

JANTO SKOWRONEK

QUALITY OF EXPERIENCE OF  
MULTIPARTY CONFERENCING  
AND TELEMEEETING SYSTEMS

METHODS AND MODELS FOR  
ASSESSMENT AND PREDICTION

TECHNISCHE UNIVERSITÄT BERLIN

# *Quality of Experience of Multiparty Conferencing and Telemeeting Systems*

## *Methods and Models for Assessment and Prediction*

vorgelegt von

Dipl.-Ing.

Janto Skowronek

geb. in Herdecke

von der Fakultät IV - Elektrotechnik und Informatik  
der Technischen Universität Berlin  
zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften

— Dr.-Ing. —

genehmigte Dissertation

Promotionsausschuss

Vorsitzender: Prof. Dr. Jean-Pierre Seifert

Gutachter: Prof. Dr.-Ing. Alexander Raake

Gutachter: Prof. Dr.-Ing. Sebastian Möller

Gutachter: Prof. Patrick Le Callet

Tag der wissenschaftlichen Aussprache: 09. Dezember 2016

Berlin 2017

# *Brief Contents*

<b>Preface</b>	<b>11</b>
<b>Zusammenfassung</b>	<b>13</b>
<b>Summary</b>	<b>15</b>
<b>I General Introduction</b>	<b>17</b>
1 General Introduction	18
<b>II Background</b>	<b>26</b>
2 Human Communication via Telemeetings	27
3 Telemeeting Technologies	51
4 Quality Assessment in Telecommunication	73
<b>III Contribution</b>	<b>88</b>
5 Conceptual Model of Telemeeting Quality	89
6 Methodology for the Perceptual Assessment of Telemeeting Quality	120
7 Perception of Telemeeting Quality with focus on Group-Communication Aspects	153
8 Perception of Telemeeting Quality with focus on Telecommunication Aspects	204
9 Perceptual Models of Telemeeting Quality	256
<b>IV Summary and Conclusions</b>	<b>290</b>
10 Summary and Conclusions	291
<b>V Appendices</b>	<b>300</b>
A Overview of Conducted Studies and Documents that Contributed to this Thesis	301
B Detailed Results: Deep Dive Analysis of Research on Group-Communication	305

<b>C Detailed Results: Study on Impact of Communication Complexity</b>	<b>329</b>
<b>D Detailed Results: Study on Impact of Involvement</b>	<b>353</b>
<b>E Detailed Results: Study on Impact of Audio-Only Telemeeting - Conversation Tests</b>	<b>363</b>
<b>F Detailed Results: Study on Impact of Audio-Only Telemeeting - Listening-Only Tests</b>	<b>371</b>
<b>G Detailed Results: Study on Impact of Audiovisual Telemeeting - Conversation Tests</b>	<b>380</b>
<b>H Detailed Results: Model Performance</b>	<b>389</b>
<b>Bibliography</b>	<b>442</b>

# Detailed Contents

<b>Preface</b>	<b>11</b>
<b>Zusammenfassung</b>	<b>13</b>
<b>Summary</b>	<b>15</b>
<b>I General Introduction</b>	<b>17</b>
<b>1 General Introduction</b>	<b>18</b>
What this chapter is about . . . . .	18
1.1 Motivation, Problem Statement, and Main Research Goal . . . . .	18
1.2 Research Scenario . . . . .	19
1.3 Research Objectives . . . . .	22
1.4 Research Method . . . . .	23
<b>II Background</b>	<b>26</b>
<b>2 Human Communication via Telemeetings</b>	<b>27</b>
What this chapter is about . . . . .	27
2.1 Introduction . . . . .	27
2.2 A Characterization of Group-Communication . . . . .	28
2.2.1 Group-Communication Purposes . . . . .	28
2.2.2 Group-Communication Processes . . . . .	29
2.2.3 Communication Modes in Group-Communication . . . . .	36
2.2.4 Shared-Work-Spaces in Group-Communication . . . . .	36
2.3 Overview of Research on Face-To-Face and Mediated Group-Communication . . . . .	38
2.3.1 Goals . . . . .	38
2.3.2 Method . . . . .	38
2.3.3 Results and Re-Interpretation . . . . .	38
2.3.4 Discussion and Conclusions . . . . .	40
2.4 Deep-Dive Analysis of Research on Face-To-Face and Mediated Group-Communication . . . . .	41
2.4.1 Motivation and Approach . . . . .	41
2.4.2 Method . . . . .	42
2.4.3 Data . . . . .	45
2.4.4 Results . . . . .	45
2.4.5 Discussion and Conclusions . . . . .	47
Summary . . . . .	49

<b>3</b>	<b>Telemeeting Technologies</b>	<b>51</b>
	What this chapter is about . . . . .	51
3.1	Introduction and Scope . . . . .	51
3.2	Main Components . . . . .	52
3.3	Central-Bridge and Peer-to-Peer Topologies . . . . .	53
3.4	End Devices . . . . .	53
	3.4.1 End Devices for Audio . . . . .	54
	3.4.2 End Devices for Video . . . . .	57
3.5	Transmission over the Network . . . . .	59
	3.5.1 Codecs . . . . .	59
	3.5.2 Packetization and Transmission . . . . .	62
	3.5.3 Buffering and Packet Error Handling . . . . .	64
	3.5.4 Audio-Video Synchronization . . . . .	64
3.6	Bridges . . . . .	65
3.7	Spatial audio . . . . .	69
	Summary . . . . .	71
<b>4</b>	<b>Quality Assessment in Telecommunication</b>	<b>73</b>
	What this chapter is about . . . . .	73
4.1	Introduction . . . . .	73
4.2	Notion of Quality and Different Perspectives . . . . .	74
	4.2.1 User-centric and Technology-Oriented Perspectives on Quality . . . . .	74
	4.2.2 A Hierarchical Perspective on Quality . . . . .	75
	4.2.3 A Process-Oriented Perspective on Quality . . . . .	75
4.3	Characterizing Quality: Quality Elements and Quality Features . . . . .	76
	4.3.1 Quality Features . . . . .	76
	4.3.2 Quality Elements . . . . .	78
4.4	Perceptual Assessment of Quality . . . . .	79
4.5	Instrumental Assessment and Prediction of Quality . . . . .	80
	Summary . . . . .	87
<b>III</b>	<b>Contribution</b>	<b>88</b>
<b>5</b>	<b>Conceptual Model of Telemeeting Quality</b>	<b>89</b>
	What this chapter is about . . . . .	89
5.1	Introduction . . . . .	89
5.2	Telemeeting Quality Formation Process . . . . .	90
	5.2.1 Basic principle . . . . .	90
	5.2.2 Reflection and Attribution . . . . .	93
	5.2.3 A more detailed process of perception . . . . .	94
	5.2.4 Quality Awareness and Quality Attention Focus . . . . .	96
	5.2.5 Summary of process model . . . . .	99
5.3	Different Aggregation Levels of Telemeeting Quality . . . . .	100
	5.3.1 Concept of Individual Connections and Individual Connection Quality . . . . .	101
	5.3.2 Mutual Influence of Individual Connection Qualities . . . . .	101
	5.3.3 Signal-based and Cue-based Quality Features . . . . .	102
	5.3.4 Interpretations of Individual Connection Quality $Q_{ij}$ . . . . .	102
	5.3.5 Single-Perspective Telemeeting Quality $Q_i$ . . . . .	104
	5.3.6 Group-Perspective Telemeeting Quality $Q_{all}$ . . . . .	105

5.3.7	Extension of the Quality Formation Process Model . . . . .	105
5.4	Telecommunication and Group-Communication Components of Telemeeting Quality . . . .	110
5.4.1	Characterization of the Telecommunication and Group-Communication Components	110
5.4.2	Influencing the Quality Attention Focus . . . . .	111
5.4.3	Quality Attention Focus in Assessment Tests . . . . .	111
5.4.4	Extension of the Quality Formation Process Model . . . . .	114
	Summary . . . . .	117
<b>6</b>	<b>Methodology for the Perceptual Assessment of Telemeeting Quality</b>	<b>120</b>
	What this chapter is about . . . . .	120
6.1	Introduction . . . . .	121
6.2	ITU-T Recommendation P.1301 . . . . .	121
6.2.1	Overview of ITU-T Recommendation P.1301 . . . . .	121
6.2.2	Selection of a Basic Test Method according to ITU-T Rec. P.1301 . . . . .	123
6.2.3	Modification and Application of a Basic Test Method according to ITU-T Rec. P.1301	124
6.2.4	Test Aspects beyond ITU-T Recommendation P.1301 . . . . .	125
6.3	Analysis Method to translate Technical Degradations into Possible Perceptual Impairments	125
6.3.1	Motivation and Basic Principle . . . . .	125
6.3.2	Step 1: Description of the Multiparty Situation . . . . .	126
6.3.3	Step 2: Identification of Degradation-Types and Degradation-Points . . . . .	127
6.3.4	Step 3: Analysis of Signal Paths and Deduction of Possible Perceptual Impairments .	133
6.3.5	Step 4: Generation of a Comprehensive Representation for all Interlocutors . . . . .	134
6.3.6	Limitations . . . . .	134
6.4	Experimental Design Method distributing Possible Perceptual Impairments instead of Tech-	
	nical Degradations . . . . .	139
6.4.1	Motivation . . . . .	139
6.4.2	Application . . . . .	139
6.5	Method to organize Perceptual Data to account for the Individual Perspectives of Interlocutors	140
6.5.1	Motivation and Basic Principle . . . . .	140
6.5.2	Application by means of an Algorithm . . . . .	143
	Summary . . . . .	151
<b>7</b>	<b>Perception of Telemeeting Quality with focus on Group-Communication Aspects</b>	<b>153</b>
	What this chapter is about . . . . .	153
7.1	Introduction . . . . .	154
7.2	Experimental Variables on a Conceptual Level . . . . .	155
7.2.1	Manipulated Variable: Communication Complexity . . . . .	155
7.2.2	Manipulated Variable: Involvement . . . . .	157
7.2.3	Manipulated Variable: Technical System Capability . . . . .	158
7.2.4	Measured Variable: Cognitive Load . . . . .	158
7.2.5	Measured Variables: Speech Communication Quality and Quality of Experience . . .	159
7.3	Quality Impact of Communication Complexity . . . . .	159
7.3.1	Goals . . . . .	160
7.3.2	Experimental Factors . . . . .	161
7.3.3	Method . . . . .	166
7.3.4	Data Acquisition . . . . .	167
7.3.5	Results Experiment CC1 . . . . .	175
7.3.6	Results Experiment CC2 . . . . .	179
7.3.7	Discussion . . . . .	183

7.4	Quality Impact of Involvement . . . . .	187
7.4.1	Goals . . . . .	187
7.4.2	Experimental Factors . . . . .	189
7.4.3	Method . . . . .	191
7.4.4	Data Acquisition . . . . .	192
7.4.5	Results Experiment INV . . . . .	196
7.4.6	Discussion . . . . .	199
	Summary . . . . .	202
<b>8</b>	<b>Perception of Telemeeting Quality with focus on Telecommunication Aspects</b>	<b>204</b>
	What this chapter is about . . . . .	204
8.1	Introduction . . . . .	205
8.2	Experimental Variables on a Conceptual Level . . . . .	206
8.2.1	Focus on Technical System Capability and Speech Communication Quality . . . . .	206
8.2.2	Quality Aggregation Levels as Measured Variables . . . . .	206
8.2.3	Involvement as Manipulated Variable across Experiments . . . . .	207
8.3	Quality Impact of an Audio-Only Telemeeting System - Conversation Tests . . . . .	207
8.3.1	Goals . . . . .	207
8.3.2	Experimental Factors . . . . .	210
8.3.3	Method . . . . .	213
8.3.4	Data Acquisition . . . . .	216
8.3.5	Results of Experiment ACT1 . . . . .	222
8.3.6	Results of Experiment ACT2 . . . . .	223
8.3.7	Results of Experiment ACT3 . . . . .	224
8.3.8	Discussion . . . . .	226
8.4	Quality Impact of an Audio-Only Telemeeting System - Listening-Only Tests . . . . .	227
8.4.1	Goals . . . . .	227
8.4.2	Experimental Factors . . . . .	228
8.4.3	Method . . . . .	230
8.4.4	Data Acquisition . . . . .	232
8.4.5	Results of Experiment LOT1 . . . . .	234
8.4.6	Results of Experiment LOT2 . . . . .	235
8.4.7	Discussion . . . . .	238
8.5	Quality Impact of an Audiovisual Telemeeting System - Conversation Tests . . . . .	240
8.5.1	Goals . . . . .	240
8.5.2	Experimental Factors . . . . .	241
8.5.3	Method . . . . .	243
8.5.4	Data Acquisition . . . . .	246
8.5.5	Results of Experiment AVCT1 . . . . .	249
8.5.6	Results of Experiment AVCT2 . . . . .	250
8.5.7	Discussion . . . . .	252
	Summary . . . . .	254
<b>9</b>	<b>Perceptual Models of Telemeeting Quality</b>	<b>256</b>
	What this chapter is about . . . . .	256
9.1	Introduction . . . . .	256
9.2	Modeling Algorithms . . . . .	257
9.2.1	Fundamental Modeling Ideas . . . . .	257
9.2.2	Modeling Functions . . . . .	262

9.2.3	Modeling Factors . . . . .	264
9.3	Model Training and Evaluation Procedure . . . . .	266
9.3.1	Motivation and Basic Principle . . . . .	266
9.3.2	Performance Measures . . . . .	267
9.3.3	Selection of the Number of Bootstrap Repetitions . . . . .	267
9.3.4	Algorithm to Split the Data into Training and Test Sets . . . . .	273
9.3.5	Selection of Proportion of Training and Test Data . . . . .	273
9.4	Model Performance . . . . .	276
9.5	Discussion . . . . .	283
9.5.1	Review of Results . . . . .	283
9.5.2	Review of the Approach and Open Questions for Future Work . . . . .	285
	Summary . . . . .	287
 <b>IV Summary and Conclusions</b>		<b>290</b>
<b>10</b>	<b>Summary and Conclusions</b>	<b>291</b>
	What this chapter is about . . . . .	291
10.1	Research Objectives Revisited . . . . .	291
10.1.1	Conceptual Model of Telemeeting Quality – Chapter 5 . . . . .	291
10.1.2	Methodology for the Perceptual Assessment of Telemeeting Quality – Chapter 6 . . . . .	292
10.1.3	Perception of Telemeeting Quality with focus on Group-Communication Aspects – Chapter 7 . . . . .	293
10.1.4	Perception of Telemeeting Quality with focus on Telecommunication Aspects – Chapter 8 . . . . .	295
10.1.5	Perceptual Models of Telemeeting Quality – Chapter 9 . . . . .	296
10.2	Final Conclusions . . . . .	298
 <b>V Appendices</b>		<b>300</b>
<b>A</b>	<b>Overview of Conducted Studies and Documents that Contributed to this Thesis</b>	<b>301</b>
	What this chapter is about . . . . .	301
A.1	Document Overview . . . . .	302
<b>B</b>	<b>Detailed Results: Deep Dive Analysis of Research on Group-Communication</b>	<b>305</b>
	What this chapter is about . . . . .	305
B.1	Analysis Results . . . . .	306
<b>C</b>	<b>Detailed Results: Study on Impact of Communication Complexity</b>	<b>329</b>
	What this chapter is about . . . . .	329
C.1	Experiment CC1 . . . . .	330
C.2	Experiment CC2 . . . . .	341
<b>D</b>	<b>Detailed Results: Study on Impact of Involvement</b>	<b>353</b>
	What this chapter is about . . . . .	353
D.1	Experiment INV . . . . .	354
<b>E</b>	<b>Detailed Results: Study on Impact of Audio-Only Telemeeting - Conversation Tests</b>	<b>363</b>
	What this chapter is about . . . . .	363
E.1	Experiment ACT1 . . . . .	364

E.2	Experiment ACT2 . . . . .	366
E.3	Experiment ACT3 . . . . .	367
<b>F</b>	<b>Detailed Results: Study on Impact of Audio-Only Telemeeting - Listening-Only Tests</b>	<b>371</b>
	What this chapter is about . . . . .	371
F.1	Experiment LOT1 . . . . .	372
F.2	Experiment LOT2 . . . . .	374
<b>G</b>	<b>Detailed Results: Study on Impact of Audiovisual Telemeeting - Conversation Tests</b>	<b>380</b>
	What this chapter is about . . . . .	380
G.1	Experiment AVCT1 . . . . .	381
G.2	Experiment AVCT2 . . . . .	384
<b>H</b>	<b>Detailed Results: Model Performance</b>	<b>389</b>
	What this chapter is about . . . . .	389
H.1	Experiment ACT1 . . . . .	390
	H.1.1 Performance across technical conditions . . . . .	390
	H.1.2 Performance per technical condition . . . . .	391
H.2	Experiment ACT2 . . . . .	395
	H.2.1 Performance across technical conditions . . . . .	395
	H.2.2 Performance per technical condition . . . . .	396
H.3	Experiment ACT3 . . . . .	398
	H.3.1 Performance across technical conditions . . . . .	398
	H.3.2 Performance per technical condition . . . . .	399
H.4	Experiment LOT1 . . . . .	408
	H.4.1 Performance across technical conditions . . . . .	408
	H.4.2 Performance per technical condition . . . . .	409
H.5	Experiment LOT2 . . . . .	414
	H.5.1 Performance across technical conditions . . . . .	414
	H.5.2 Performance per technical condition . . . . .	415
H.6	Experiment AVCT1 . . . . .	427
	H.6.1 Performance across technical conditions . . . . .	427
	H.6.2 Performance per technical condition . . . . .	428
H.7	Experiment AVCT2 . . . . .	433
	H.7.1 Performance across technical conditions . . . . .	433
	H.7.2 Performance per technical condition . . . . .	434
	<b>Bibliography</b>	<b>442</b>

# *Preface*

Many people had an influence on this PhD thesis; and for expressing my gratitude to those people, I would like to play with analogies and borrow four influencing factors from the quality research domain: the processing chain, the personal experience in the past, the context, and the personal state.

In terms of the processing chain, i.e. the thesis supervision and review, my first and foremost “thank you” goes to Alexander Raake, who – among many other things – suggested to me this fascinating topic, provided me with inspiring ideas, fought for financial support, and delved with me into deep discussions about multiparty telemeeting quality. Next, many many thanks go to Sebastian Möller and Patrick Le Callet, who were so kind to review this thesis, or in other words, who took on the laborious task of ensuring its quality.

In terms of past experience, i.e. my time at Philips Research before starting the PhD research, I want to thank Armin Kohlrausch, Steven van de Par, Martin McKinney and Jeroen Breebaart, who all taught me to do scientific research in an industrial environment. Furthermore, my decision to go for a PhD training was strongly triggered by having worked with people like Othmar Schimmel, Tom Goossens, Stefan Borchert, Michael Bruderer, Alberto Novello, Nicolas Le Goff, Tobias May, and many others at Philips Research.

In terms of context, in which this research took place at T-Labs, the list of people is long: Katrin Schoenenberg for her inspiring multidisciplinary perspective on quality, for many discussions on many details, and for being just the peer colleague that one needs as a PhD student. Mathis Schmieder and Julian Herlinghaus for their tremendous help in the technical setups of Asterisk, Netem and Co. Falk Schiffner and Maxim Spur for their support in the optimization of the different test scenarios, their help in the test conduction, and their vast endurance to run the scenarios dozens of times as a confederate subject. Hagen Wierstorf for the many discussions related with writing up my research and for showing me this amazing Tufte-style Latex template. Markus Vaalgaama and Mattias Nilsson from Skype for their support in the technical setup of the Skype test client and for the many discussions on the test design and the results. Adriana Dumitras and Bernhard Feiten for joining those discussions and for providing the financial support of Skype (Microsoft) and T-Labs (Deutsche Telekom). Marcel Wältermann and Marie-Neige Garcia for sharing their knowledge about the many tiny

details of running quality tests properly and for their inspiring PhD theses. Christine Kluge for helping in all kinds of administrative things and problems and for all the fun when we were joking around. The many present and former colleagues of the AIPA, QU, T-Labs and IMT teams as well as the ITU-T SG12 colleagues for enabling such a great working context, among them Sandra Wild, Dennis Guse, Michal Soloducha, Pierre Lebreton, Miguel Rios-Quintero, Savvas Argyropoulos, Werner Robitza, David Roegiers, Angelo De Silva, Christoph Hold, Eckhardt Schön, Stephan Werner, Judith Köppe, Anne Keller-Marz, Irene Huber-Achter, Friedemann Köster, Benjamin Weiss, Ulf Wüstenhagen, Matthias Geier, Stefan Hillmann, Mattias Schulz, Peter Hughes and Gunilla Berndtsson.

Finally, in terms of personal state, nothing of this would have been possible without the continuous, fundamental, emotional and practical support of my parents, my parents-in-law and – foremost – my wife: Barbara, grazie mille!

Janto Skowronek, March 2017

# Zusammenfassung

Die vorliegende Dissertation behandelt die Qualitätswahrnehmung von Telemeetingsystemen, d.h. audiovisuellen Telekonferenzsystemen, bei denen mehr als zwei Gesprächspartner miteinander verbunden sind. Die Dissertation legt den Schwerpunkt auf zwei Aspekte, die ein Telemeeting, d.h. Mehrpersonengespräch über ein Telemeetingsystem, von einem konventionellen (Video-)Telefonat zwischen zwei Teilnehmer unterscheiden. Der erste Aspekt ist die besondere kommunikative Situation, die sich dadurch hervorhebt, dass sie ein Gruppengespräch darstellt, welches über ein Telekommunikationssystem stattfindet. Der zweite Aspekt ist die Möglichkeit, dass die Nutzer asymmetrische Verbindungen antreffen können, d.h. dass die Gesprächspartner mit unterschiedlichen Geräten und Leitungseigenschaften verbunden sind. Das wiederum ermöglicht es, verschiedene Niveaus einer Telemeetingqualität erfahren zu können, nämlich die Qualität des gesamten Telemeetings und der Qualität der einzelnen Verbindungen der Teilnehmer.

Zunächst wird in dieser Dissertation ein konzeptionelles Model entwickelt, welches als theoretische Grundlage dient. Danach wird eine Methode zur Durchführung von Qualitätstest für Mehrpersonengespräche beschrieben. Anschließend werden drei Experimente vorgestellt und analysiert, welche den Einfluss von kommunikativen Aspekten, genauer der Komplexität der Kommunikationssituation und der Anteilnahme am Gespräch, untersuchen. Danach werden sieben auditive und audiovisuelle Experimente vorgestellt und analysiert, welche den Einfluss der technischen Systemeigenschaften auf die oben genannten verschiedenen Qualitätsniveaus untersuchen. Zuletzt werden in dieser Dissertation Modelle entwickelt, welche Urteile der Gesamtqualität anhand von Qualitätsurteilen der Einzelverbindungen schätzen. Da Eingangs- und Ausgangsgrößen der vorgestellten Modelle perzeptive Nutzerurteile und nicht instrumentelle Qualitätsschätzungen sind, liefern die Modelle eine konzeptionelle Basis für die Entwicklung technischer Lösungen.

Die wesentlichen Erkenntnisse dieser Arbeit sind:

1. Eine komplexere Kommunikationssituation kann unter denselben technischen Bedingungen zu niedrigeren Qualitätsurteilen führen als eine weniger komplexe Kommunikationssituation.
2. Der Einfluss der Anteilnahme am Gespräch in einem Qualitätstest kann größer sein als der Effekt der getesteten technischen

Bedingungen.

3. Es existiert eine gegenseitige Beeinflussung der Qualitätsurteile der Einzelverbindungen in einem Gespräch, wobei diese Beeinflussung von den technischen Bedingungen abhängt.
4. Das Qualitätsurteil des gesamten Systems kann in vielen Fällen als einfacher Mittelwert der Qualitätsurteile der Einzelverbindungen geschätzt werden. Jedoch für mehrere technische Bedingungen liefern komplexere Schätzmodelle bessere Ergebnisse.
5. Die besten Modelle erreichen Korrelationskoeffizienten von bis zu 0.95 und RMSE-Werte (Wurzel des Mittleren Quadratischen Fehlers) von 0.45 auf der 5-Punkte Absolute-Category-Rating-Skala. Das zeigt, dass es möglich ist, die Gesamtqualität anhand der Einzelverbindungsqualitäten angemessen schätzen zu können.
6. Es hat sich keine Modellfunktion herausgestellt, die als einzige alle anderen untersuchten Funktionen signifikant übertrifft. Jedoch zeigte eine der Funktionen den besten Kompromiss zwischen zwei grundsätzlichen Trends: Ist die Gesamtqualität gut, dann liegt der Mittelwert der Einzelverbindungsurteile sehr nahe an dem Gesamturteil; ist die Gesamtqualität schlecht, dann liegt das Minimum der Einzelverbindungsurteile sehr nahe an dem Gesamturteil. Demnach berechnet sich die meist versprechendste Modellierungsfunktion als eine gewichtete Kombination aus Mittelwert und Minimum der Einzelverbindungsurteile, wobei die Gewichtung selbst wiederum vom Mittelwert abhängt.
7. Einzelne Details bei Design und Durchführung der Qualitätstests zeigten unterschiedliche Messempfindlichkeiten hinsichtlich der drei Punkte Detektion der gegenseitigen Beeinflussung der Einzelverbindungen, Aufspüren von Belegen, dass eine einfache Mittelung nicht ausreicht, und Nachweisen des Zugewinns durch Verwendung komplexerer Schätzmodelle.

Als nächste Schritte für weitere Arbeiten bietet sich die Umsetzung der gewonnenen Erkenntnisse in technische Lösungen an. Dieses würde zunächst eine Nachbildung der Modellierung beinhalten, in der die perzeptiven Qualitätsurteile der Einzelverbindungen durch instrumentelle Qualitätsschätzungen ersetzt werden. Später könnten dann solche Modelle durch das Einbinden der kommunikativen Aspekte verbessert werden. Dieses wiederum erfordert die Entwicklung technischer Ansätze zur Schätzung derartiger kommunikativer Aspekte wie der hier untersuchten Komplexität der Kommunikationssituation oder der Anteilnahme am Gespräch.

# Summary

This thesis investigates quality perception of multiparty audio and video conferencing systems, in short telemeeting systems. It is built around two aspects that differentiate a multiparty telemeeting from a conventional two-party (video) telephony call. The first aspect is the special communicative situation, that is, having a group conversation over a telecommunication system. The second aspect is the possibility of encountering asymmetric connections, that is, participants are connected with different equipment or connection properties. This in turn leads to the possibility to perceive different levels of telemeeting quality, i.e. the quality of the whole telemeeting and the quality of the individual connections of participants.

First, this thesis develops a conceptual model that serves as a theoretical foundation. Then it describes a methodology to conduct multiparty quality assessment tests. Next it analyzes three experiments on the impact of the communicative aspects (i.e. communication complexity and involvement) on telemeeting quality. Then it analyzes seven audio-only and audiovisual experiments on the impact of technical conditions on the different levels of telemeeting quality. And finally it develops models that predict overall telemeeting quality scores based on the quality scores of the individual connections. Since all considered scores are perceptual ratings and not estimations obtained from instrumental methods, these models are conceptual ones, which serve as the foundation for developing technical solutions.

The main findings of the experimental chapters are:

1. An increased communication complexity can lead to lower quality ratings for the same technical conditions.
2. The effect of involvement in a quality assessment test can be larger than that due to the tested technical conditions.
3. There exists a mutual influence on the quality scores of individual connections, whereas this depends on the character of the actual technical conditions.
4. The overall telemeeting quality can be expressed as a simple mean of the individual connection quality scores. However, for some technical conditions, more complex models achieve better results.
5. With correlation coefficients up to 0.95 and root mean square errors down to 0.45 on the absolute category rating scale, it is possible

to adequately model the overall telemeeting quality based on the individual connection quality scores.

6. No single modeling function is the winning function in terms of significantly outperforming all other functions. However, one advanced modeling function appeared to provide the best compromise between two fundamental trends: if the telemeeting quality score is high, then the mean of individual connection quality scores is close to that score, if the telemeeting quality score is low, then the minimum of the individual connection quality scores is close. Accordingly, the most promising modeling function is computed as the weighted combination of the mean and the minimum value of the individual connection quality scores, whereas this weighting depends again on the mean of the individual connection quality scores.
7. Specific design details of the experimental methods distinguished the achieved sensitivity of the experiments in terms of their ability to detect the mutual influence of individual connections, to provide evidence that a simple mean is not always sufficient, and to show the added value of using advanced modeling functions.

The next steps will be to transfer the found knowledge to technical applications. Foremost, this requires to repeat the modeling work with replacing the perceptual quality ratings of the individual connections with instrumental estimations. Later on, such models could be further improved by including the impact of the communication aspects, which will require the development of technical approaches to estimate aspects such as communication complexity or involvement.

## **Part I**

# **General Introduction**

# 1

## *General Introduction*

### *What this chapter is about*

This chapter provides a general introduction to the research, it specifies the research scenario and research objectives, and it outlines the research process that the author applied.

### *1.1 Motivation, Problem Statement, and Main Research Goal*

One key factor for any successful telecommunication service is to satisfy users while optimizing the use of technological and financial resources. This requires methods to assess or predict the Quality of Experience that the service is able to provide and to link this quality with the technical parameters of the service.

Such methods exist and are continuously improved for traditional two-party (video) telephony services and they are employed by network providers and end device manufacturers. However, very little is known about the applicability of such methods for multiparty telemeeting solutions. Concerning quality assessment by means of perception tests, proposals exist to modify and apply existing perceptual quality test methods to multiparty telemeetings, but practical experience with such modified methods is still limited. Concerning quality prediction by means of algorithmic models, to the best of the author's knowledge, no work has been reported so far on the application of existing models to multiparty situations.

An alternative to applying existing methods – with or without modifications – to multiparty telemeetings is to develop new multiparty-specific methods from scratch. However, also here there are, to the best of the author's knowledge, no reports on such new methods.

One requirement common for both approaches, application of (modified) existing methods or development of new methods, is to have sufficient knowledge about the quality perception of multiparty telemeetings. Apparently there is a lot of knowledge available on Quality of Experience, but this knowledge has not been explicitly transferred to the multiparty case yet. Furthermore, there is also lot of knowledge available on group communication via telemeeting systems, but the link to Quality of Experience has not been studied in more detail either.

To conclude, it is essential to further investigate the fundamentals of multiparty telemeeting quality and to fill the apparent knowledge gaps on the corresponding assessment and prediction methods. This knowledge could then be exploited to improve the Quality of Experience of existing or future multiparty telemeeting systems.

To this aim, the present thesis will investigate two main aspects that differentiate a multiparty telemeeting from a conventional two-party (video) telephony call. One aspect is the special communicative situation, that is, having a group conversation over a telecommunication system. The other aspect is the possibility of encountering asymmetric connections, that is, participants are connected with different equipment or connection properties. These two aspects will form the basis for the present research and they will be referred to throughout the whole text.

## 1.2 Research Scenario

The research scenario is best described with an explanation of the keywords used in the title: Quality of Experience, Multiparty, Conferencing and Telemeeting Systems, Methods, Models, Assessment, and Prediction. Furthermore, there are additional aspects of telemeetings that further specify the research scenario. Those are the communication modes, the conversation scenarios, and the combination of the number of interlocutors and number of sites.

*Conferencing and Telemeeting Systems* The present work considers human-to-human communication via so-called conferencing or telemeeting systems. Such systems connect multiple participants at distant locations to join the same call.

Commonly used names of such systems, which have been traditionally used in business environments, are telephone conference systems, audio conferencing systems, video conferencing systems, or simply teleconferencing systems. Furthermore, high-end video conferencing systems that aim to enable the feeling of a virtual meeting are called telepresence systems. Nowadays, however, modern conferencing systems gain also more importance in private life and such systems are striving for higher levels of interactivity than the traditional systems did. To account for such societal and technological trends, the term telemeeting is proposed by the International Telecommunication Union (ITU). Work group “Studygroup 12 - Question 10” uses in its task description<sup>1</sup> the term telemeeting

to cover with one term all means of audio or audio-visual communication between distant locations.

Concerning the title, using both terms conferencing and telemeeting appeared to be most appropriate due to the use cases that are considered in the present work: Conferencing, because the conventional business scenarios play an important role in the present research; telemeeting because some of the applied technical scenarios may be

<sup>1</sup> ITU-T. *Question 10/12 - Conferencing and telemeeting assessment*. retrieved on August 19, 2015. International Telecommunication Union. 2015. URL: <http://www.itu.int/en/ITU-T/studygroups/2013-2016/12/pages/q10.aspx>

better described with the new term telemeeting systems. However, for the sake of brevity and due to the lack of a formally defined distinction between both terms, the present text will use the term telemeeting only.

*Multiparty* The term multiparty further specifies the research scenario to telecommunication situations, in which actually more than two interlocutors are participating in a call. This is necessary because the terms conferencing and telemeeting are not exclusively reserved for multiparty situations. For instance, the term video conferencing is very common in literature to describe two-party telecommunication using an audiovisual system.

*Quality of Experience* A precise definition for Quality of Experience is given by Qualinet<sup>2</sup>:

Quality of Experience (QoE) is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and / or enjoyment of the application or service in the light of the user's personality and current state.

<sup>2</sup> European Network on Quality of Experience in Multimedia Systems and Services. *Qualinet White Paper on Definitions of Quality of Experience*. White Paper. Version 1.1.2. Patrick Le Callet, Sebastian Möller and Andrew Perkis (Eds.) Lausanne, Switzerland: (COST Action IC 1003), Mar. 2013

According to this definition, the research presented here will investigate how users assess the quality of telemeeting systems. More specifically, all quality judgments used in this work are collected from users, who are not technical experts of telemeeting technology. Nevertheless, the perspective of a technical expert plays an important role in this work, especially when it comes to linking technical system parameters with the user quality judgments or when it comes to deriving development guidelines for improved telemeeting systems in the future.

*Assessment* The term quality assessment means in this work the act of collecting quality judgments from human raters who participated in a quality assessment test. During such a test, participants have been exposed either to a running test system in the context of a so-called conversation test or to stimuli that represent such a system in the context of a so-called non-interactive test.

*Prediction* The term quality prediction means in related literature typically the act of algorithmically computing quality judgments based on technical characteristics without conducting a quality assessment test with human participants. Nevertheless, those predicted quality judgments should represent as best as possible quality judgments that human test participants give to a system of the same technical conditions.

In this work, however, the meaning of the term prediction is slightly broader. It refers to the act of algorithmically computing quality judgments based on individual aspects of quality, which can comprise technical information about or human ratings on those individual aspects.

*Methods* In this work, the term methods refers to formalized procedures to prepare, conduct and analyze quality assessment tests. In particular, this work will put a strong emphasis on the details of such experimental methods. This is motivated by the knowledge in the field that quality perception is context-dependent, emphasizing the need for systematic and well-defined experimental methods.

*Models* In this work, the term models refers to descriptions of the relation between a set of input and output variables. More specifically, the input variables here describe different aspects of a telemeeting, and the output variables describe different aspects of quality.

*Communication Modes* Many telemeeting systems provide different means to communicate, ranging from audio-only to audiovisual communication, potentially augmented with additional features such as text chat, screen sharing, or joint editing of documents. In this work, the focus here lies on audio-only and audiovisual communication without any additional communication means.

*Conversation Scenarios* Nowadays the use cases for telemeetings can vary widely, ranging from the classical pre-scheduled business teleconference scenario following a predefined agenda to spontaneous group calls between friends. That means, the communication in telemeetings can differ in numerous aspects such as topic and purpose, roles of and relations between the interlocutors, communication behavior of interlocutors, structure of the conversation and so on. In this work, well-defined and structured conversation scenarios are used for triggering conversations, which follow a rather fixed structure addressing similar topics between interlocutors who take upfront-determined roles. This allows a sufficient comparability for proper analysis of the collected results, as it controls – at least to a certain extent – for those different influencing aspects.

*Number of Interlocutors and Number of Sites* The combination of number of interlocutors and number of sites can have an impact on the quality perception of a multiparty telemeeting. The reason is that the particular setting influences the communicative situation in terms of possible mixtures of face-to-face conversations and conversations via the telecommunication system.

To illustrate this consider two examples: In one telemeeting, two remote sites are connected. At one site two persons are located in the same room, while at the other site one person is alone in his or her room. Such a telemeeting is typically characterized by a mix of a face-to-face conversation between the two interlocutors at the one site and a conversation via a telecommunication system to the interlocutor at the other site.

In another telemeeting using the same system, three remote sites are connected while one person is located at every site. Such a telemeeting is actually a “pure” multiparty conference via a telecommu-

nication system. As a result, the quality perception of the telemeeting system in both examples can differ due to the different communicative situations (face-to-face vs. telecommunication).

In this work, the focus lies on the latter type of situations that avoid any face-to-face communication in the telemeeting.

### 1.3 *Research Objectives*

The overall research objective is to investigate the fundamentals of multiparty telemeeting quality and to fill a number of knowledge gaps on the corresponding assessment and prediction methods. In particular, the goal is to provide a set of prototypical methods and models for the quality assessment and prediction of multiparty telemeeting systems. This set of methods and models shall serve as a foundation for further research on the Quality of Experience (QoE) of multiparty telemeetings and they shall be directly applicable for practitioners in the field.

For that purpose, this thesis shall develop conceptual approaches to address the fundamentals of multiparty telemeeting QoE and practical approaches to prove the feasibility of assessing and predicting multiparty telemeeting QoE. More precisely, the present text will first provide appropriate background knowledge on the communication aspects (Chapter 2) and technical aspects (Chapter 3) of telemeetings as well as quality assessment (Chapter 4). Then the text will report in individual chapters on the research results for the following detailed objectives.

*Conceptual Model of Telemeeting Quality – Chapter 5* The objective is to derive from existing taxonomies of quality perception a general overview of aspects and processes that may play a role when test participants form a quality judgment of telemeeting systems. Since this work will be based on a literature review and theoretical reasoning, the result is of a descriptive conceptual nature. The intention of such a conceptual model is two-fold. On the one hand it summarizes the known fundamentals of telemeeting quality perception and can thus help to identify open issues and questions for future research. On the other hand it serves as a tool to put individual research activities on telemeeting quality assessment into a broader context, which would help to interpret results and properly draw conclusions.

*Methodology for the Perceptual Assessment of Telemeeting Quality – Chapter 6* When starting the research for this thesis, there were no standardized multiparty-specific methods available for conducting quality tests with test participants. Hence, the objective was to develop a set of such testing methods and synthesize them into a set of recommendations that can be used by practitioners in the field. The chapter provides an overview of these methods and recommendations.

*Perception of Telemeeting Quality with focus on Group-Communication Aspects – Chapter 7* The objective is to investigate the impact of a number of communicative aspects on quality ratings in the context of multiparty telemeetings. This refers to the first of the two main differentiators between two-party (video) telephony and multiparty telemeetings, the special communicative situation. More specifically, a set of experimental studies shall provide empirical results on the relation between relevant communication aspects and the perceived quality.

*Perception of Telemeeting Quality with focus on Telecommunication Aspects – Chapter 8* The objective is to investigate the impact of various technical system characteristics on quality ratings in the context of multiparty telemeetings. This refers to the second of the two main differentiators between two-party (video) telephony and multiparty telemeetings, the possibility of perceiving asymmetric conditions. More specifically, a set of experimental studies shall provide empirical results on the relation between the individual connections between the interlocutors and the perceived quality of the whole telemeeting.

*Perceptual Model of Telemeeting Quality – Chapter 9* While the previous objectives addressed conceptual and methodological aspects of telemeeting quality, tangible results in form of computational models are also required as a proof-of-concept in two aspects. First, such computational models, if successful, can serve as algorithmic implementation of the gained knowledge on certain aspects of quality formation, providing thus evidence for the conceptual work done beforehand. Second, they could serve as basis for developing technical solutions for quality monitoring or planning, providing thus evidence for their usefulness. For that reason, the objective is to develop a set of perceptual models, which predict how test participants form a quality judgment on individual aspects that they perceive.

#### 1.4 Research Method

*Motivation and Basic Approach* The two aspects that differentiate multiparty telemeeting from two-party (video) telephony, i.e. special communicative situation and possible asymmetry of connections, suggest that telemeeting quality is a complex construct of communicative and technical aspects. At the time of starting the present research, there were different approaches to decompose the complexity of quality perception, e.g. the taxonomy of Möller<sup>3</sup>, but it has, to the best of the author's knowledge, not been attempted for the specific situation of multiparty telemeetings. Furthermore, it is known in the field that quality perception can be influenced by the context, which means that the quality assessment of telemeetings needs to follow well-defined test protocols in order to provide reproducible results. Again, at the time of starting the present research, individual proposals for the specific situation of multiparty telemeetings have been made, see

<sup>3</sup> Sebastian Möller. *Assessment and Prediction of Speech Quality in Telecommunications*. Kluwer Academic publishers, 2000

e.g. Baldis<sup>4</sup>, Raake<sup>5</sup> or Berndtsson<sup>6</sup>, but a standardized method was not developed yet. For that reason, the research activities conducted throughout this thesis have a highly iterative character, specifically addressing the two mentioned aspects complexity and methodology.

Concerning the methodology, the experimental work comprises (in most cases) two similar experiments for a given research task, e.g. two listening-only tests for investigating the impact of number of interlocutors. The first experiment serves as a pilot experiment for a second full-scale experiment, with the idea that the first test proves the general feasibility of the particular method and identifies improvement points, which are then applied in the second test. Here, the pilot experiments should not be misunderstood as some informal trying-out; they are conducted as formal tests like the corresponding full-scale experiments.

Concerning the complexity, many aspects are expected to influence the results, meaning that the experimental work should address a number of those aspects individually in an iterative manner. That means the individual aspects are not only investigated separately by keeping the other aspects as constant as possible – which is obviously the usual scientific approach – but when proceeding to the next experimental phase, the knowledge gained from the first experimental phase should be incorporated to better specify the research context of the second experimental phase. More specifically, a first set of experiments is intended to investigate the communicative aspects of telemeeting quality, while not being too stringent concerning the choice of the technical aspects. Then a second set of experiments is intended to investigate technical aspects of telemeeting quality, in which the gained knowledge on the communicative aspects is used to better specify the considered use cases and target variables.

*Visualization of Research Process* Figure 1.1 summarizes the iterative research process. First, it shows a number of aspects that constitute an experimental study: concept, method, experiment, data, and results. Then, it shows a number of stages (black blocks) that a researcher goes through when conducting an experimental study: develop, use, prepare, do, get, analyze, and interpret. Furthermore, it shows also a number of stages (red blocks) that were gone through when extracting knowledge from an experimental study: review, report, and feedback. Finally, it also shows the iterative character (blue paths) in which a new study is conducted based on the lessons learned from the previous study.

Concerning the feedback paths, the insights can stem from the own review, the external feedback or both. Then, those insights can be used to improve the concept (e.g. when conducting a full study based on a pilot study) or develop a new concept (e.g. when addressing a new research question, i.e. a new aspect of telemeeting quality with the next study). However, the insights can also be used to directly update the experimental method and the practical preparation and conduction of the experiment.

<sup>4</sup> Jessica J. Baldis. “Effects of Spatial Audio on Memory, Comprehension, and Preference during Desktop Conferences”. In: *Proceedings of the ACM CHI 2001 Human Factors in Computing Systems Conference*. Ed. by Michel Beaudouin-Lafon and Robert J. K. Jacob. Vol. 3. 1. 2001, pp. 166–173

<sup>5</sup> Alexander Raake et al. “Listening and conversational quality of spatial audio conferencing”. In: *Proceedings of the AES 40th International Conference*. Tokyo, Oct. 2010

<sup>6</sup> Gunilla Berndtsson, Mats Folkesson, and Valentin Kulyk. “Subjective quality assessment of video conferences and telemeetings”. In: *19th International Packet Video Workshop*. 2012

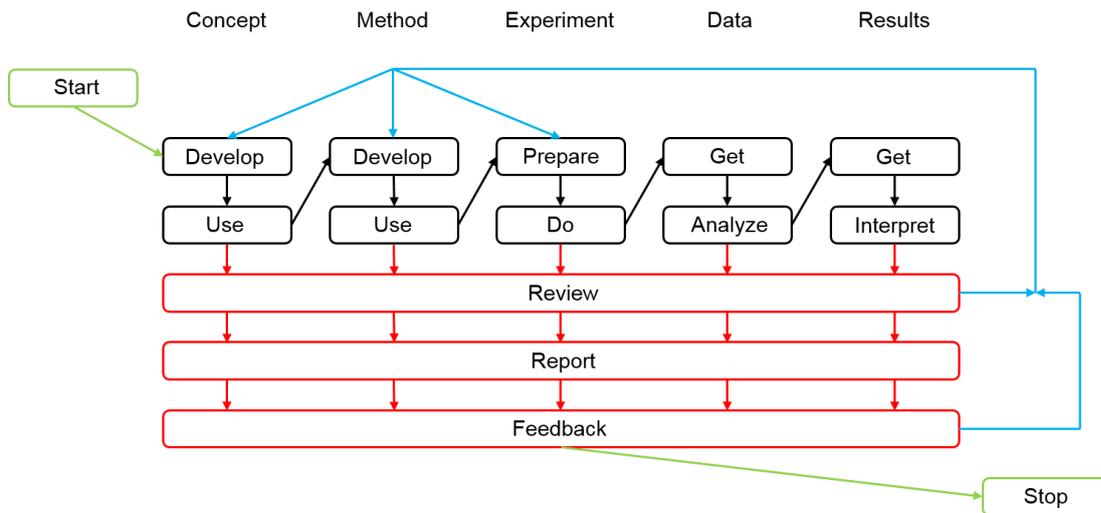


Figure 1.1: Visualization of the research process, defining the progress of activities through a number of stages per experimental study.

The approach in the present research project is to consider – whenever possible – all five aspects during the review and feedback stages and to integrate the obtained insights – whenever applicable – into all three aspects (concept, method, and experiment) when continuing from one experiment to the next.

*Synthesis of Results for the Present Text* The iterative character of the individual studies led to a number of individual documents in form of publications, technical reports and one patent application (see Appendix A for the full list). The present text, however, is actually based on an additional iteration, in which the author revisited and – whenever it deemed necessary – re-analyzed and re-interpreted the results of the individual documents.

## **Part II**

# **Background**

## Human Communication via Telemeetings

### *What this chapter is about*

The present text is built around two aspects that differentiate a multiparty telemeeting from a conventional two-party (video) telephony call. One aspect is the special communicative situation, that is, having a group conversation over a telecommunication system. Bearing this in mind, the present chapter provides background knowledge on the communicative aspects that characterize two-party and multiparty communication. Both face-to-face communication and mediated communication, that is communication via a telecommunication system, are considered.

### 2.1 Introduction

There is a vast amount of literature on human communication and in particular on mediated communication, that is communication via a telecommunication system. This amount of publications is written from different perspectives and research disciplines and inherently showing the complexity of this process.

On the one hand, this chapter aims to provide a concise overview of the existing literature, covering two-party and multiparty situations. On the other hand, this chapter also aims to cover a number of different perspectives that are taken in the literature. For that purpose, the present text approaches in three sections the literature from three different perspectives and following two different reviewing methods:

1. Section 2.2 will provide an overview about aspects described in literature that can be used to characterize human communication in general, independently from the question whether it is a face-to-face or mediated communication. In terms of Light & Pillemer<sup>1</sup>, this overview is a narrative literature review.
2. Section 2.3 will discuss the results of a series of literature reviews by Fjermestad & Hiltz<sup>2</sup>, who investigated the differences between face-to-face and mediated communication. In terms of Light & Pillemer, the reviews of Fjermestad & Hiltz are quantitative, i.e. data-driven, reviews. Such reviews are based on a quantitative analysis of data extracted from the literature.

<sup>1</sup> Richard J. Light and David B. Pillemer. *Summing Up*. Cambridge, MA, USA: Harvard University Press, 1984

<sup>2</sup> Jerry Fjermestad and Starr Roxanne Hiltz. "Experimental studies of group decision support systems: an assessment of variables studied and methodology". In: *Proceedings of the 30th Annual Hawaii International Conference on System Sciences*. Vol. 2. IEEE. 1997, pp. 45-65

Jerry Fjermestad and Starr Roxanne Hiltz. "An Assessment of Group Support Systems Experimental Research: Methodology and Results". In: *Journal of Management Information Systems* 15.3 (1999), pp. 7-149

Jerry Fjermestad and Starr Roxanne Hiltz. "Case and field studies of group support systems: an empirical assessment". In: *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*. IEEE. 2000, pp. 10-19

Jerry Fjermestad. "An analysis of communication mode in group support systems research". In: *Decision Support Systems* 37 (2004), pp. 239-263

3. Section 2.4 will present an own literature analysis on similar literature with the focus on identifying potential measurable aspects that can characterize mediated communication. In terms of Light & Pillemer, this review belongs also to the data-driven reviews.

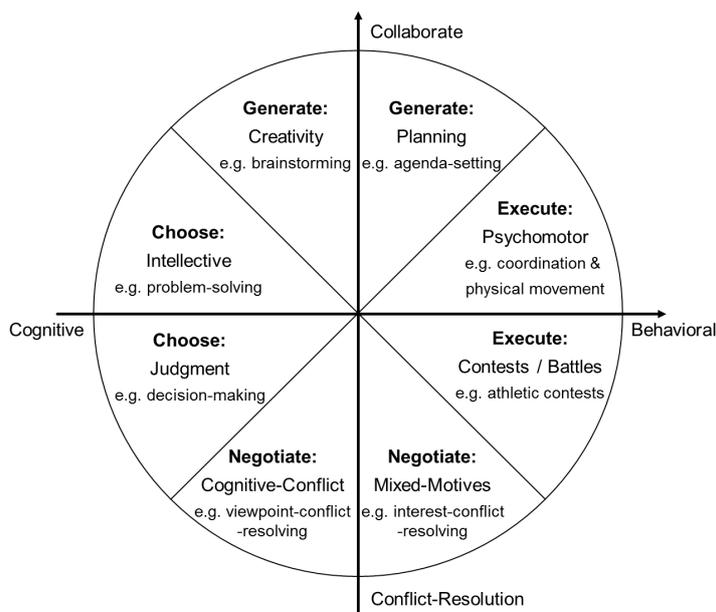
## 2.2 A Characterization of Group-Communication

This section provides some fundamental knowledge on group-communication (including a group size of two) that can be found in literature. The aim is to decompose the complexity of group-communication by providing a systematic characterization. The chosen approach is a separation of communication into purpose, process (with six prominent aspects), modes (also referred to as media), and the aspect of a shared work space.

### 2.2.1 Group-Communication Purposes

Group-communication is considered here as a means to serve a certain set of purposes, i.e. achieve certain goals or fulfill certain needs. Different communication purposes are possible and can be roughly categorized in accomplishing tasks, socializing, and exchange of information.

Concerning the accomplishment of tasks, an often cited categorization of group tasks is McGrath's task circumplex<sup>3</sup>, which was in parts empirically evaluated by Straus<sup>4</sup>. This circumplex model categorizes group tasks in four categories and eight subcategories and organizes them along two dimensions, see Figure 2.1.



<sup>3</sup> Joseph Edward McGrath. *Groups: Interaction and performance*. Prentice-Hall Englewood Cliffs, NJ, 1984

<sup>4</sup> Susan G. Straus. "Testing a Typology of Tasks: An Empirical Validation of McGrath's (1984) Group Task Circumplex". In: *Small Group Research* 30 (1999), pp. 166–187. DOI: 0.1177/104649649903000202

Figure 2.1: Two-dimensional circumplex model of tasks – adapted from McGrath. Joseph Edward McGrath. *Groups: Interaction and performance*. Prentice-Hall Englewood Cliffs, NJ, 1984

With the term socializing the present text refers to those group-communications that serve the social needs of group members to feel

connected, to belong to the same group, etc. Examples for interlocutors who know each other are telling about the state of mood, health, and recent daily-life events. An example for interlocutors who do not know each other is getting acquainted with each other, which essentially means telling – to a certain degree – about personality, life experience and opinions.

With the term exchange of information the present text refers to those group-communications that solely serve the purpose of bringing each group member up-to-date on a specific topic. Examples are announcements, news, or useful information for group members in case of something specific is happening in the foreseeable future, e.g. who would act as sick leave cover.

With such a categorization of communication purposes, it is possible to characterize group communication in a formal systematic way. In turn, when it comes to the interpretation of experimental results, this enables to consider the group-communication purpose as a possible mediating factor that might explain certain results.

### 2.2.2 Group-Communication Processes

In the literature, a number of approaches are described to characterize group-communication processes from different perspectives. Six prominent approaches are *Turn-Taking*, *Listener Feedback Signals*, *Conversational Surface Structure*, *Grounding*, *Conversational Games* and *Conversational Moves*. As a way to describe the perspectives and thus to relate the approaches to each other, the author chose to use five aspects: (1) the temporal granularity, i.e. a microscopic moment-by-moment perspective vs. a macroscopic in-retrospect perspective; (2) the level of organization, i.e. focus on organizing speaker changes vs. focus on ensuring that content is properly conveyed; (3) the characterization power, i.e. intrinsic property of communication vs. measure to characterize communication; (4) the reliance on the semantic content, i.e. requiring semantic analysis of the utterances vs. semantics-independent characterization; and (5) the focus on the interlocutors' contributions to the communication, i.e. focus on speaker vs. focus on listener vs. considering both. Table 2.1 categorizes the six approaches accordingly to the five aspects, while the following paragraphs describe each aspect in more detail and motivate the categorization.

Aspect	Temporal Granularity	Level of Organization	Characterization Power	Reliance on Semantics	Focus on Interlocutor
Turn-Taking	microscopic	speaker changes	intrinsic property	yes	both
Listener Feedback Signals	microscopic	speaker changes & conveying content	intrinsic property	yes	listener
Conversational Surface Structure	microscopic & macroscopic	speaker changes	measure to characterize	no	both
Grounding	macroscopic	conveying content	intrinsic property	yes	both
Conversational Games	macroscopic	conveying content	intrinsic property	yes	both
Conversational Moves	microscopic & macroscopic	conveying content	intrinsic property	yes	speaker

Table 2.1: Overview of six approaches (Column 1) characterizing communication processes from different perspectives, each perspective described along five dimensions (Columns 2 to 6).

*Turn-Taking* Turn-taking refers to the temporal organization of and transitions between speech utterances of each interlocutor; in other words it describes who is speaking when and how a change of speakers is accomplished. Probably the most cited publication on turn-taking is a model proposed by Sacks et al.<sup>5</sup>. This model describes turns to be composed of turn construction units (e.g. sentential, clausal, phrasal and lexical constructions). At the end of each such unit, there is a so-called transition-relevance place, which may lead to a continuation of the current turn or to a speaker change. Furthermore, the model provides a set of four if-then rules that determine the outcome of each transition-relevance place. Thus, the model essentially describes turn-taking to take place on a moment-by-moment basis: at the end of each turn construction unit a decision is made for either turn-keeping or turn-taking.

The model of Sacks et al. focuses on turn-taking behavior based on verbal information provided by the speaker. However, speakers also use non-verbal signals to manage turn-taking. For instance, speakers (and listeners) use eye movements to indicate when it is time to change a turn<sup>6</sup>. In addition, speakers can use gestures to deter listeners to make a turn-taking attempt, or they can send turn surrendering signals by ceasing to gesticulate<sup>7</sup>.

Regarding the categorization in Table 2.1, turn-taking concerns the organization of speaker changes, whereas the process, i.e. decision to keep or change a turn, relies on the semantical meaning of the shared information, both verbally and non-verbally. Since turn-taking refers to a process that interlocutors actually use, it can be considered as an intrinsic property of communication. Furthermore, turn-taking, and especially the model of Sacks et al., looks at communication on a moment-by-moment and thus microscopic time scale. And finally, turn-taking focuses on both the speaker and the listener, since turns can be allocated by speaker selection or self-selection, and the decision to keep or change a turn can be taken by both, speaker and listener.

*Listener feedback signals* Listener feedback signals – also referred to as backchannels<sup>8</sup> or listener responses<sup>9</sup> – are short signals to the speaker indicating the listener's degree of attention, understanding, and acceptance of the communicated message. Those signals may be produced vocally, e.g. "mm", "uh uh", "right", "okay", and "yes", or by means of facial expressions, gestures and posture changes, e.g. straightening upper body part, head nods, and establishing mutual eye gaze.

While listener feedback signals mainly refer to the listener's understanding of the speakers' messages, it can also serve the organization of speaker changes as part of the turn-taking processes. For instance, the listener may use such signals to approve or disapprove that the current speaker took the turn, to show satisfaction with the current turn so far, i.e. to encourage the current speaker to continue, or to indicate the need for repairing or re-initiating a turn-taking process<sup>10</sup>.

Regarding the categorization in Table 2.1, listener feedback signals concern both the organization of speaker changes and the conveyance

<sup>5</sup> Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. "A Simplest Systematics for the Organisation of Turn-Taking for Conversation". In: *Language* 50 (4 Dec. 1974), pp. 696–735

<sup>6</sup> Mark L. Knapp and Judith A. Hall. *Nonverbal communication in human interaction*. 7th edition. Boston, USA: Wadsworth, Cengage Learning, 2010

<sup>7</sup> Owen Daly-Jones, Andrew Monk, and Leon Watts. "Some advantages of video conferencing over high-quality audio conferencing: fluency and awareness of attentional focus". In: *International Journal of Human-Computer Studies* 49 (1998), pp. 21–58

<sup>8</sup> Brid O'Conaill, Steve Whittaker, and Sylvia Wilbur. "Conversations Over Video Conferences: An Evaluation of the Spoken Aspects of Video-Mediated Communication". In: *Human-Computer Interaction* 8 (1993), pp. 389–428

Owen Daly-Jones, Andrew Monk, and Leon Watts. "Some advantages of video conferencing over high-quality audio conferencing: fluency and awareness of attentional focus". In: *International Journal of Human-Computer Studies* 49 (1998), pp. 21–58

<sup>9</sup> Mark L. Knapp and Judith A. Hall. *Nonverbal communication in human interaction*. 7th edition. Boston, USA: Wadsworth, Cengage Learning, 2010

<sup>10</sup> Owen Daly-Jones, Andrew Monk, and Leon Watts. "Some advantages of video conferencing over high-quality audio conferencing: fluency and awareness of attentional focus". In: *International Journal of Human-Computer Studies* 49 (1998), pp. 21–58

of content. The process, i.e. the listeners' choices of the feedback signal, relies on the semantic meaning of the shared information. Since interlocutors actually use feedback signals, they can be considered as an intrinsic property of communication. Furthermore, listener feedback signals are produced on a moment-by-moment and thus microscopic time scale and they focus on the listener, obviously.

*Conversational surface structure* Conversational surface structure describes the organization of speaker changes from a semantics-independent perspective by means of employing a probabilistic state model, an idea introduced by Brady<sup>11</sup>. The basic concept encompasses three steps. The first step is to define all combinations of talking and non-talking interlocutors as so called *states*: no-one is talking (one silence state), each speaker is talking alone ( $N$  single-talk states), two speakers are talking simultaneously ( $\sum_{i=1}^{N-1} (N-i)$  double-talk states), up to all  $N$  speakers are talking simultaneously (one  $N$ -talk state). This is done for the perspective or – as Schoenberg<sup>12</sup> writes – *perceptual reality* of each interlocutor, as those states can be different for each interlocutors when there is transmission delay on the connection. The second step is to employ a Markov-Chain model by heuristically determining for each state the probability to remain in that state (sojourn probability), and the probabilities to move to one of the other states (transition probabilities). Figure 2.2 gives visual examples for a two-party and a three-party conversation. The third step is to characterize a conversation by directly using those state probabilities, or by extracting from those state probabilities a number of secondary parameters. Such secondary parameters can transform the state probabilities to temporal variables by including the timing information about when the conversation was in which state. Or they reflect simple statistics on specific state walks during a conversation, again related to timing information. An example for the former parameter type is the State Sojourn Time per state<sup>13</sup> (mean time of staying in that state); examples for the latter parameter type are the Speaker Alternation Rate<sup>14</sup> (mean rate at which speaker changes occurred) or Utterance Rhythm<sup>15</sup> (mean time from one utterance of a speaker to his or her next utterance).

Regarding the categorization in Table 2.1, conversational surface structure concerns the organization of speaker changes. It represents a way of characterizing (measuring) communication that does not rely on the semantic meaning of the shared information; however, it is not an intrinsic property of communication as such. Furthermore, the underlying state model means that conversational surface structure works on a moment-by-moment and thus microscopic time scale. However, it can be argued that some of the parameters represent statistics of a whole conversation, meaning that conversational surface structure can apply also an in-retrospect and thus macroscopic time scale. And finally, conversational surface structure focuses on one hand on the speaker, since it analyzes speaker state probabilities, on the other hand also on the listener as it refers to the perceptual reality

<sup>11</sup> Paul T. Brady. "A technique for investigating on-off patterns of speech". In: *Bell Syst. Tech. J.* 44.1 (1965), pp. 1–22

Paul T. Brady. "A statistical analysis of on-off patterns in 16 conversations". In: *Bell Syst. Tech. J.* 47.1 (1968), pp. 73–99

Paul T. Brady. "Effects of transmission delay on conversational behavior on echo-free telephone circuits". In: *Bell Syst. Tech. J.* 50.1 (1971), pp. 115–134

<sup>12</sup> Katrin Schoenberg. "The Quality of Mediated-Conversations under Transmission Delay". PhD Thesis. Technische Universität Berlin, Germany, 2016

<sup>13</sup> Paul T. Brady. "A technique for investigating on-off patterns of speech". In: *Bell Syst. Tech. J.* 44.1 (1965), pp. 1–22

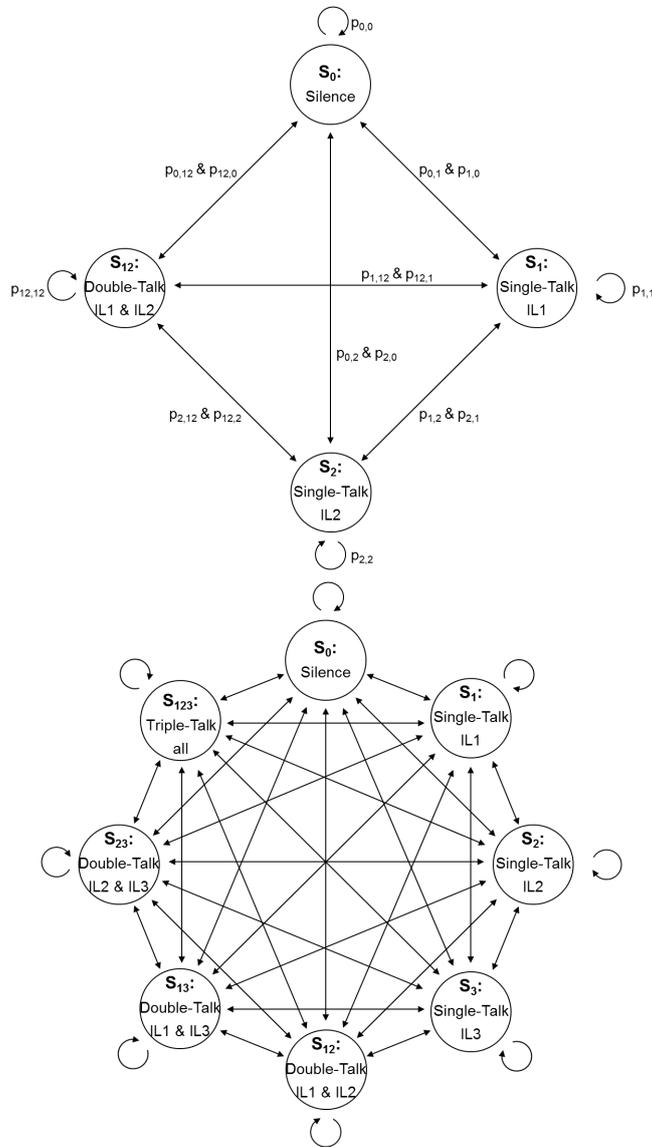
Paul T. Brady. "A statistical analysis of on-off patterns in 16 conversations". In: *Bell Syst. Tech. J.* 47.1 (1968), pp. 73–99

Paul T. Brady. "Effects of transmission delay on conversational behavior on echo-free telephone circuits". In: *Bell Syst. Tech. J.* 50.1 (1971), pp. 115–134

<sup>14</sup> Florian Hammer, Peter Reichl, and Alexander Raake. "Elements of interactivity in telephone conversations". In: *Proceedings of the 8th International Conference on Spoken Language Processing (Interspeech 2004)*. Jeju Island, Korea, 2004, pp. 1741–1744

Sebastian Egger, Raimund Schatz, and Stefan Scherer. "It takes two to tango – assessing the impact of delay on conversational interactivity on perceived speech quality". In: *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech 2010)*. Makuhari, Japan, 2010, pp. 1321–1324

<sup>15</sup> Katrin Schoenberg et al. "On interaction behaviour in telephone conversations under transmission delay". In: *Speech Communication* 63 (2014), pp. 1–14



Remarks:

Each state  $S_x$  (with indices  $x \in [0,1,2,12]$  for two-party and  $x \in [0,1,2,3,12,13,23,123]$  for three-party conversation) is visualized by a circle. The sojourn probabilities  $p_{x,x}$  and transition probabilities  $p_{x,y}$  (read:  $p$  from state  $x$  to state  $y$ ) are indicated by arrows. To simplify the visualization, the two possible transitions between each state pair – from  $x$  to  $y$  and from  $y$  to  $x$  – are merged into one double-headed arrow, and the nomenclature  $p_{x,y}$  is only shown for the two-party case.

Figure 2.2: Visualization of Markov Chain models for a two-party (top panel) and three-party conversation (bottom panel).

of each listener.

*Grounding* Grounding, introduced as a term by Clark & Brennan<sup>16</sup>, refers to the process of establishing a mutual believe between speaker and listeners that the message or – in the terms used by Clark & Brennan – a contribution has been correctly understood. This means, turn taking behavior as described by Sacks et al.<sup>17</sup> is driven by the attempt of interlocutors to reach this mutual understanding, this common ground. Clark & Brennan describe that the understanding of a contribution consists of a presentation phase (speaker presents an utterance) and an acceptance phase (listener give evidence that they believe they understood the utterance). Thus, the grounding process can cover one to several turns: one turn – the oririnal utterance as such – if the utterance was positively acknowledged by listener feedback signals (which do not lead to a turn change); more than one turn if listeners request clarification.

Regarding the categorization in Table 2.1, grounding concerns the conveyance of content and is, obviously, relying on a semantic analysis of the information shared. Furthermore, grounding is a process that establishes mutual understanding of the previous episodes of a conversation; hence it can be seen as an in-retrospect process and as an intrinsic property of communication. Thus it rather refers to a macroscopic time scale, not in the sense of a whole conversation but of individual parts. And finally, since grounding is about mutual understanding, it requires consideration of both, speaker and listener.

*Conversational games and moves* Conversational games and moves refer – as Doherty-Sneddon et al.<sup>18</sup> describe – to the “pragmatic functions of utterances with respect to achieving speakers’ goals”. That means, a conversational game is a unit of a conversation that encompasses a set of utterances that serves the accomplishment or, alternatively, the abandonment of a certain goal<sup>19</sup>. In that respect, grounding, as discussed above, can be seen as one such type of conversational games, while an example for other conversational games is to ask someone to perform a certain action and making sure that this instruction is correctly understood.

Each conversational game can consist of one or more conversational moves, which are utterances that can be classified according to their purpose, such as initiating a conversational game or acknowledging understanding. Different coding schemes for conversational moves have been used in a number of studies using different conversation tasks. Carletta et al. proposed and evaluated such a coding scheme that was especially tailored to the map task<sup>20</sup>. While that scheme consists of 12 categories of conversational moves, alternative coding schemes have been used in the literature as well. For instance, Veinott et al.<sup>21</sup> used – again for the map task – a slightly modified version of that scheme, essentially merging some categories and introducing two additional categories for utterances not directly related to the task at hand. Furthermore, Kraut et al.<sup>22</sup> tailored their

<sup>16</sup> Herbert H. Clark and Susan E. Brennan. “Grounding in Communication”. In: *Perspectives on socially shared cognition*. Ed. by Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley. American Psychological Association, 1991, pp. 127–149

<sup>17</sup> Harvey Sacks, Emanuel A. Schegloff, and Gain Jefferson. “A Simplest Systematics for the Organisation of Turn-Taking for Conversation”. In: *Language* 50 (4 Dec. 1974), pp. 696–735

<sup>18</sup> Gwyneth Doherty-Sneddon et al. “Face-to-Face and Video-Mediated Communication: A Comparison of Dialogue Structure and Task Performance”. In: *Journal of Experimental Psychology: Applied* 3.2 (1997), pp. 105–125

<sup>19</sup> Jean Carletta et al. *HCRC dialogue structure coding manual*. Report. University of Edinburgh, UK, 1996. URL: <http://www.lancaster.ac.uk/fass/projects/eagles/maptask.htm>

Jean Carletta et al. “The reliability of a dialogue structure coding scheme”. In: *Computational Linguistics* 23.1 (1997), pp. 13–31

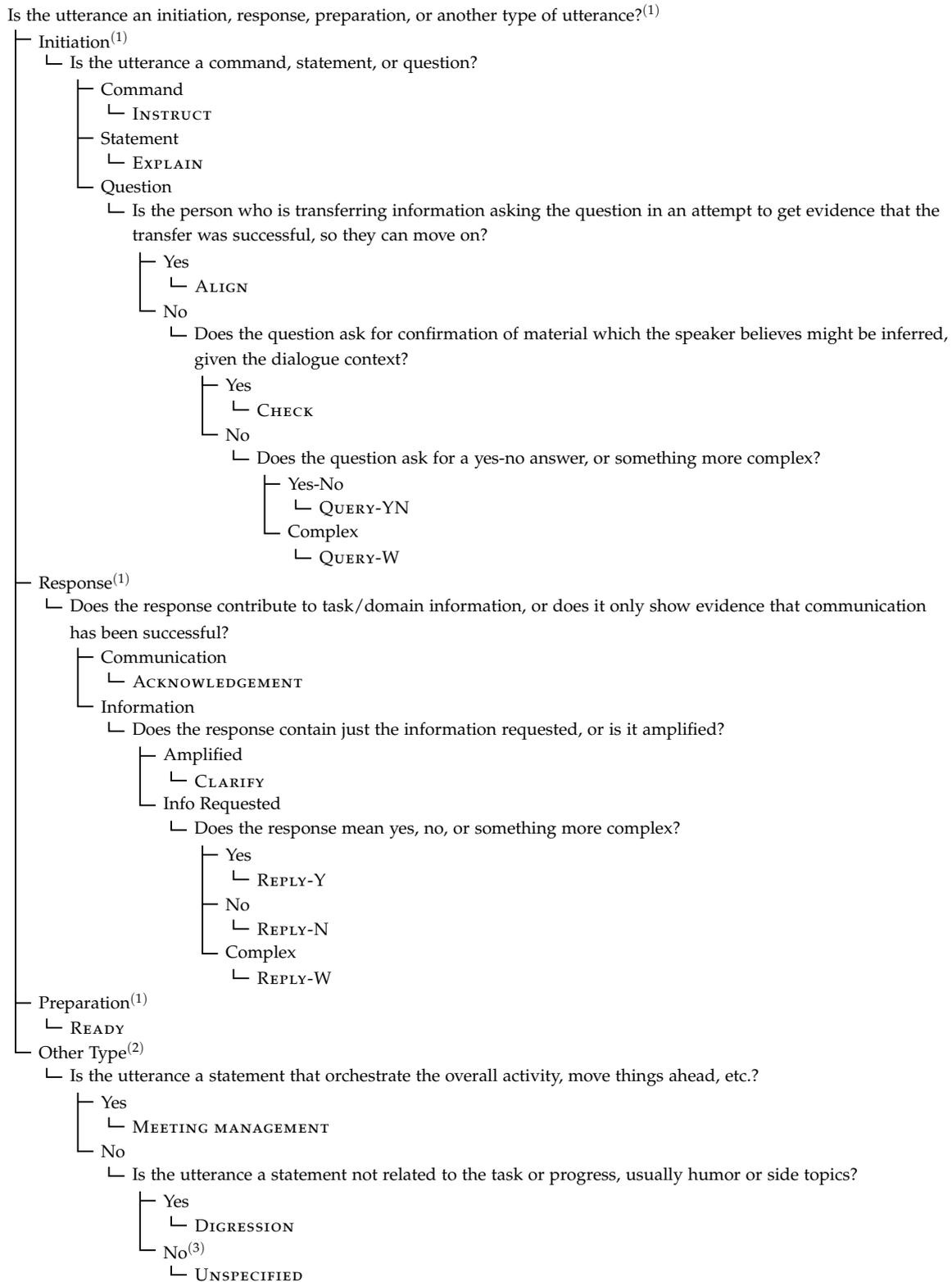
<sup>20</sup> Anne H. Anderson et al. “The HCRC Map Task Corpus”. In: *Language and Speech* 34.4 (1991), pp. 351–366

<sup>21</sup> Elizabeth S. Veinott et al. “Video Helps Remote Work: Speakers Who Need to Negotiate Common Ground Benefit from Seeing Each Other”. In: *Proceedings of CHI 1999*. 1999

<sup>22</sup> Robert E. Kraut, Susan R. Fussell, and Jane Siegel. “Visual Information as a Conversational Resource in Collaborative Physical Tasks”. In: *Human-Computer Interaction* 18 (2003), pp. 13–49

coding scheme to the helper-worker-task (bicycle repair task) that they used. However, Kraut's categories can be reinterpreted as merged categories from the coding scheme of Carletta et al. This means, the combined coding schemes of Carletta et al. and Veinott et al. appear to form a representative and hierarchical set of definitions for conversational moves. Figure 2.3 provides the definition of the 14 resulting conversational moves.

Regarding the categorization in Table 2.1, conversational games and moves concern the conveyance of content and are, obviously, relying on a semantic analysis of the information shared. Furthermore, as conversational moves and games build on grounding to a certain extent, they represent intrinsic properties of communication. In terms of the time scale, conversational games refer to the function of episodes of a conversation and therefore look at conversation on a rather macroscopic time scale. The time scale for conversational moves, however, is ambiguous. On the one hand, conversational moves are categorized by the function of utterances in light of the conversation goals, which suggests a macroscopic time scale. On the other hand, conversational moves are determined per utterances, which suggests the same microscopic time scale as defined for turn-taking. And finally, conversational games refer to the mutual achievement or abandonment of goals, which means they concern both speaker and listener, while conversational moves is a categorization of spoken turns, which means they concern the speaker.



Notes:

- (1): Coding branches according to Carletta et al., with literal wording of the definitions.
- (2): Coding branches according to Veinott et al.; the original definitions are rephrased as questions with minimum modifications, to achieve consistency in the combined scheme.
- (3): Additional branch to account for utterances that do not fit any of the categories.

Figure 2.3: Coding scheme for conversational moves, combination of the schemes from Carletta et al.(1996) and Veinott et al.(1999)

### 2.2.3 *Communication Modes in Group-Communication*

Communication Mode – also referred to as Communication Modality or Communication Media – describes the medium over which interlocutors are conversing<sup>23</sup>. On a high level, communication mode refers to the communication modalities in terms of face-to-face, audiovisual, audio-only, and text-only communication. On a finer level, there are many different instantiations possible, depending on the actual technical implementation. For instance, audio-only communication can be realized as full-duplex (both sides can simultaneously talk) or half-duplex (only one side can talk at a time), audiovisual communication can use different video frame rates, and text-only communication can be realized as real-time synchronous chat or message-based asynchronous email. Furthermore, another very important aspect of group-communication is the availability of a shared-workspace as part of the communication mode, which will be discussed next.

### 2.2.4 *Shared-Work-Spaces in Group-Communication*

The term shared work space is usually used in the context of task-driven communication. Shared work space refers to a physical or virtual space that allows the interlocutors to share information and objects, which supports achieving the task goals at hand. In terms of Buxton<sup>24</sup>, a shared work space – or task space – refers to a “copresence in the domain of the task being undertaken”. A typical example for a physical, also often referred to as co-located, shared work space is a whiteboard in a meeting room, which serves the interlocutors to visualize (draw and write) all relevant information currently discussed about. An example for a virtual, also often referred to as distributed, shared work space is given by Buxton: a data spreadsheet presented on networked computers at different physical sites, whereas each interlocutor can assess, mark and edit the spreadsheet, visibly to the others. Another example for a virtual shared work space that is common in today’s computer-based conferencing solutions is screen sharing, in which one interlocutor sends a visual representation of his or her computer desktop to the remote participants.

Since shared work spaces can be realized in quite different ways, a systematic approach to characterize the essential features of shared work spaces is beneficial. A first such feature obviously is the distinction between physical/co-located and virtual/distributed shared work spaces. Two more such features are introduced by Park<sup>25</sup>, who provides a categorization of co-located and distributed variations of shared work spaces by examining two dimensions: visibility and controlability. Visibility means the extent to which an owner of information is sharing the view with interlocutors, and controlability means the extent to which an owner of information is sharing the control with interlocutors. Both dimensions range from private (never share) via mixed (partially share) to public (always share).

Relating the concept of shared work space to group communication, it is intuitive that the absence or presence of a shared work space

<sup>23</sup> ITU-T. *Recommendation P.1301 - Subjective quality evaluation of audio and audiovisual telemeetings*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2012

<sup>24</sup> William A. S. Buxton. “Telepresence: integrating shared task and person spaces”. In: *Proceedings of Graphics Interface '92*. Amsterdam, The Netherlands, Oct. 1992, pp. 123–129

<sup>25</sup> Kyoung Shin Park. “Enhancing Cooperative Work in Amplified Collaboration Environments”. PhD Thesis. Graduate College of the University of Illinois at Chicago, 2003

can influence the communication. This assumption is supported by Gergle et al.<sup>26</sup>, Kraut et al.<sup>27</sup> and Whittaker et al.<sup>28</sup>, but it should be further validated in the light of the vast amount of existing related literature (for example,  $\geq 200$  relevant studies were identified in two reviews by Fjermestad & Hiltz<sup>29</sup>).

Regardless of the empirical evidence, from a theoretical and methodological perspective, the present text advocates a distinction between communication with and without shared work space. Buxton<sup>30</sup>, for instance, introduced a formal distinction between the shared work space and a person space, which refers to the perceived co-presence of the interlocutors. Along that line of thought, Masoodian<sup>31</sup> proposed to categorize the experimental studies of face-to-face and mediated communication into four categories:

- Studies that investigated the person space, by changing the communication medium but keeping the shared workspace unchanged.
- Studies that investigated the shared work space, by changing the shared work space but keeping the communication medium unchanged.
- Studies that investigated person and shared work space.
- Studies that investigated other aspects.

Buxton discussed that depending on the task at hand, the separation of person and shared work space can differ from none at all to distinct. His examples for the former case are negotiation or counseling, his example for the latter case is working with a shared data spreadsheet. Given this continuum between shared work space and person space, this concept of the two spaces suggests to consider the two spaces not as two distinct categories, but as two dimensions along which communication scenarios can be characterized: to what extent is the communication concerned with or focusing on the person space, to what extent with shared work. Such characterization can, for instance, be done by developing a corresponding conversational coding scheme. A starting point for such a scheme could be the coding scheme from Kraut et al.<sup>32</sup> defining six types of messages, which can be grouped into shared work space related messages (procedural, task status, and referential utterances) and person space related messages (internal state, acknowledgments, and other utterances). A complementary coding scheme from Gergle et al.<sup>33</sup> examines the use of deictic expressions (e.g., “that left one”) when referring to the task object.

<sup>26</sup> Darren Gergle, Robert E. Kraut, and Susan R. Fussell. “Language Efficiency and Visual Technology – Minimizing Collaborative Effort with Visual Information”. In: *Journal of Language and Social Psychology* 23.4 (Dec. 2004), pp. 1–27

<sup>27</sup> Robert E. Kraut, Susan R. Fussell, and Jane Siegel. “Visual Information as a Conversational Resource in Collaborative Physical Tasks”. In: *Human-Computer Interaction* 18 (2003), pp. 13–49

<sup>28</sup> Steve Whittaker and Brid O’Conaill. “An Evaluation of Video Mediated Communication”. In: *INTERACT ’93 and CHI ’93 Conference Companion on Human Factors in Computing Systems*. ACM. 1993, pp. 73–74

<sup>29</sup> Jerry Fjermestad and Starr Roxanne Hiltz. “An Assessment of Group Support Systems Experimental Research: Methodology and Results”. In: *Journal of Management Information Systems* 15.3 (1999), pp. 7–149

Jerry Fjermestad. “An analysis of communication mode in group support systems research”. In: *Decision Support Systems* 37 (2004), pp. 239–263

<sup>30</sup> William A. S. Buxton. “Telepresence: integrating shared task and person spaces”. In: *Proceedings of Graphics Interface ’92*. Amsterdam, The Netherlands, Oct. 1992, pp. 123–129

<sup>31</sup> Masood Masoodian. “Human-to-Human Communication Support for Computer-Based Shared Workspace Collaboration”. PhD Thesis. Department of Computer Science, The University of Waikato, Hamilton, New Zealand, 1996

<sup>32</sup> Robert E. Kraut, Susan R. Fussell, and Jane Siegel. “Visual Information as a Conversational Resource in Collaborative Physical Tasks”. In: *Human-Computer Interaction* 18 (2003), pp. 13–49

<sup>33</sup> Darren Gergle, Robert E. Kraut, and Susan R. Fussell. “Language Efficiency and Visual Technology – Minimizing Collaborative Effort with Visual Information”. In: *Journal of Language and Social Psychology* 23.4 (Dec. 2004), pp. 1–27

### 2.3 Overview of Research on Face-To-Face and Mediated Group-Communication

There is a vast amount of literature on Face-To-Face and Mediated Group-Communication, mainly in the context of collaborative working. In a series of reviews, Fjermestad & Hiltz<sup>34</sup>, have examined more than 200 relevant studies published in journals and peer-reviewed conferences. Considering that they excluded other publication types such as PhD theses or book chapters, that they focused on studies with at least three interlocutors, and that their last review is already ten years old, today's real number of relevant studies will be much higher. Nevertheless, since these reviews provide – to the author's knowledge – the most comprehensive overview of studies, the present section will be based on these review articles.

#### 2.3.1 Goals

The main goal of Fjermestad & Hiltz was to provide a systematic and graspable overview of the vast amount of literature in the context of collaborative working. More specifically, they examined existing research from the perspective of a potential added value of collaborative working tools, which they call *Group Support Systems*, compared to face-to-face collaboration.

In terms of the present text, on the one hand, they analyzed collaborative working for different communication modes, thus including both face-to-face and mediated communication. On the other hand they analyzed collaborative working with and without providing a shared work space, which they refer to as *Task Support Tools*.

#### 2.3.2 Method

The analysis is based on a framework that allows a systematic characterization of the vast amount of experimental input and output variables. The framework defines four variable categories: contextual factors (i.e. independent variables); intervening factors; adaptation factors; and outcome factors. Table 2.2 provides a short overview of those four types, for the full description the present text refers to the original papers<sup>35</sup>.

In order to extract numerical results from the individual studies, Fjermestad & Hiltz first coded all located studies according to their framework and put them into a database. Then they computed simple statistics in terms of counts, i.e. how many experiments studied certain variables or variable combinations ("What has been studied") and how often an effect of the system was observed ("Face-to-face is better/equal/worse than system-based").

#### 2.3.3 Results and Re-Interpretation

Here the author summarizes and re-interprets the main results reported in the two main publications of the four found reviews from

<sup>34</sup> Jerry Fjermestad and Starr Roxanne Hiltz. "Experimental studies of group decision support systems: an assessment of variables studied and methodology". In: *Proceedings of the 30th Annual Hawaii International Conference on System Sciences*. Vol. 2. IEEE. 1997, pp. 45–65

Jerry Fjermestad and Starr Roxanne Hiltz. "An Assessment of Group Support Systems Experimental Research: Methodology and Results". In: *Journal of Management Information Systems* 15.3 (1999), pp. 7–149

Jerry Fjermestad and Starr Roxanne Hiltz. "Case and field studies of group support systems: an empirical assessment". In: *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*. IEEE. 2000, pp. 10–19

Jerry Fjermestad. "An analysis of communication mode in group support systems research". In: *Decision Support Systems* 37 (2004), pp. 239–263

<sup>35</sup> Jerry Fjermestad and Starr Roxanne Hiltz. "An Assessment of Group Support Systems Experimental Research: Methodology and Results". In: *Journal of Management Information Systems* 15.3 (1999), pp. 7–149

Jerry Fjermestad. "An analysis of communication mode in group support systems research". In: *Decision Support Systems* 37 (2004), pp. 239–263

Contextual Factors	Intervening Factors
<p><b>Technology:</b> e.g. Communication Mode, Task Support Tools, Process Structure</p> <p><b>Group:</b> e.g. Group Characteristics, Leadership, Member Characteristics, Meeting Structure</p> <p><b>Task:</b> e.g. Task Type, Structure, Equivocality, Complexity</p> <p><b>Context:</b> e.g. Environment, Organizational, Cultural</p>	<p><b>Methods:</b> e.g. Experimental Design, Task Implementation, Training</p> <p><b>Communication Dimensions:</b> e.g. bandwidth, media richness, social presence</p> <p><b>Group Member Perception &amp; Problem Solving:</b> Task Performance Strategies, Member Knowledge and Skill, Commitment, Level of Interest, Biases</p> <p><b>Organizing Concepts:</b> e.g. Information Processing System, Consensus Generating System, Behavior Motivation &amp; Regulation</p> <p><b>Operating Conditions:</b> e.g. Available Modalities, Changes in Task, Divisions of Labor</p>
<p>Adaptation Factors</p> <p><b>Group Adaptation Factors:</b> Structural Features, Process Variables, Process Issues</p> <p><b>Process Gains/Losses:</b> e.g. Synergy, Learning, Attenuation Blocking, Information Overload</p> <p><b>Intermediate Role Outcomes:</b> e.g. Role Assumption by Technology, Actual Roles of Participants, Values</p>	<p>Outcome Factors</p> <p><b>Consensus:</b> e.g. Decision Agreement, Commitment</p> <p><b>Efficiency Measures:</b> e.g. Decision Time, Number of Decision Cycles</p> <p><b>Effectiveness Measures:</b> Decision Quality, Process Quality, Level of Understanding, Task Focus</p> <p><b>Satisfaction Measures:</b> e.g. participation, Confidence, Attitude, General Satisfaction</p> <p><b>Usability Measures:</b> e.g. Learning Time, Willingness to work together again, Design Preference</p>

Fjermestad & Hiltz. One review<sup>36</sup> was conducted in 1998 and concerns the full set of studies located at that time. The other review<sup>37</sup> was conducted in 2000 and provides a more detailed analysis for those studies that examined the impact of communication mode, thus excluding all studies that manipulated only the other contextual factors (see Table 2.2).

Fjermestad & Hiltz found that in the majority of studies no significant effect between mediated and face-to-face communication was found, and that for the remaining studies the ratio of positive and negative effects, i.e. mediated communication is better/worse than face-to-face, is almost equal to one. The counts in terms of tested hypotheses across all studies in the 1998-review were 628 for no effect, 158 for mediated being better than face-to-face, and 164 for mediated being worse than face-to-face. The corresponding counts in the 2000-review were 239 with no effect, 177 for mediated being better, and 191 for mediated being worse.

Looking at the results for the individual variables, again no significant effect could be found for most variables. While Fjermestad & Hiltz present a comprehensive analysis of the many variables considered, the author decided to focus here only on three dependent variables that appear to be most relevant for the present chapter: *Efficiency*, *Effectiveness* and *Satisfaction*. The results for these outcome variables are shown in Table 2.3. Here, interesting observations can be made with respect to the typical assumptions about the added

Table 2.2: Short overview of a framework to characterize studies on collaborative working, according to Fjermestad & Hiltz.

<sup>36</sup> Jerry Fjermestad and Starr Roxanne Hiltz. "An Assessment of Group Support Systems Experimental Research: Methodology and Results". In: *Journal of Management Information Systems* 15.3 (1999), pp. 7-149

<sup>37</sup> Jerry Fjermestad. "An analysis of communication mode in group support systems research". In: *Decision Support Systems* 37 (2004), pp. 239-263

value of mediated communication, considering that two contradicting assumptions are reasonable: face-to-face communication is better, given the richness and naturalness of the communication mode with its potential for fast interaction and turn-taking behavior; or mediated communication is better, as the restrictions on the communication mode force people to focus on the task and to minimize non-task related conversations.

Outcome Factor	Total Count	No Effect	Pos. Effect	Neg. Effect	Other Effect
1998-Review					
<i>Efficiency</i>	97	22	10	26	39
<i>Effectiveness</i>	617	230	73	44	270
<i>Satisfaction</i>	280	133	18	29	100
2000-Review					
<i>Efficiency</i>	49	4	11	29	5
<i>Effectiveness</i>	259	89	72	57	41
<i>Satisfaction</i>	126	54	24	25	24

Table 2.3: Results from Fjermestad & Hiltz for a selection of considered outcome variables. The numbers show the counts in terms of observed effects over all considered experiments. *Pos. Effect* means mediated communication better than face-to-face communication, *Neg. Effect* the opposite, and *Other* refers to cases without measured main effects or with significant interaction effects only.

Concerning *Efficiency*, the data from Fjermestad & Hiltz shows a clear tendency for the naturalness argument preferring face-to-face communication. Concerning *Effectiveness*, the data suggests that in most cases no significant effect can be found, but if there is an effect, then the mediated communication reaches more often higher *Effectiveness*. This would support the argument that participants focus more on mediated communication. Concerning *Satisfaction*, face-to-face is more preferred than mediated communication in those cases in which there is an effect, again supporting the naturalness argument. However, the vast majority of studies showed no effect at all, which suggest that other aspects than naturalness and media richness contribute to *Satisfaction*, such as the *Effectiveness*. In the author's view, an alternative explanation could be that participants are used to the necessity of assuming a different communication behavior in mediated communication, which is therefore not decreasing satisfaction in most cases.

#### 2.3.4 Discussion and Conclusions

Although the reviews of Fjermestad & Hiltz were very comprehensive, i.e. covering more than 200 experiments, they were conducted from a very specific perspective: evaluate the added value of what they call *Group Support Systems* compared to face-to-face communication.

However, looking at Fjermestad & Hiltz' descriptions of *Group Support Systems*, a substantial number of those systems provided mixed mediated and face-to-face communication, the so-called *Decision Room* systems (related to 347 out of 705 variables in the 2000-review). Furthermore, a second group of systems (related to 268 out

of 705 variables in the 2000-review) were called *Computer-Mediated-Communication* systems, which are primarily providing text-based communication, i.e. chat and email, even though they may include audio or audiovisual communication channels as well. This means that the body of tested systems was beyond the scope of the present work, which considers audio and audiovisual communication only, and only between remote interlocutors, i.e. excluding combinations of mediated and face-to-face communication. As a consequence, the results of Fjermestad & Hiltz may not be perfectly evaluating the effect of “purely” mediated group-communication vs. face-to-face communication.

Nevertheless, given the comprehensiveness of the reviews, it can be concluded that the effect of mediated group-communication vs. face-to-face communication is overall highly ambiguous, since roughly one-third of the studies showed no effect, one-third of the studies an effect favoring mediated communication and one-third of the studies favoring face-to-face communication. This result, however, does not mean that the communication behavior of the interlocutors is not affected, it only means that the measurable aspects of the communication processes and their outcomes did not differ. In other words, it is likely that people assume a different communication behavior when conversing over a telecommunication system in order to reach the communication goals despite the limitations of the particular communication channel, whereas those changes in communication behavior could not be consistently measured across studies.

## 2.4 *Deep-Dive Analysis of Research on Face-To-Face and Mediated Group-Communication*

### 2.4.1 *Motivation and Approach*

Despite the comprehensiveness of the reviews by Fjermestad & Hiltz (see previous section), the author decided to conduct an own deep-dive analysis of research on face-to-face and mediated group-communication. The motivation was to obtain a complementary picture of existing literature from a different perspective than the typical research question “Is face-to-face communication better or worse than mediated communication?”. Instead the goal was to obtain insights which variables might explain differences in the communication media for the purpose of characterizing human communication via telemeeetings.

Emphasizing the aspect of obtaining complementary results to Fjermestad & Hiltz, the present deep-dive literature analysis includes complementary literature, i.e. including publications later than 2000, including two-party studies, including studies investigating only mediated communication, and including non-article style publications such as PhD theses. While increasing the broadness of included studies, the present analysis focused on much less studies, i.e. 44 experiments described in 42 publications, noting that some publications

covered multiple experiments, some experiments were covered by multiple publications. The motivation for this trade-off was to be able to conduct a detailed analysis of individual variables in light of the available resources.

#### 2.4.2 *Method*

The present analysis is a cross-study analysis and looks at the effects of communication media on different dependent variables. Since those studies did not use standardized protocols, measures, or other anchors allowing cross-study comparisons, it is not possible to express the results of individual experiments on one common scale in terms of quantitative values. For that reason, the author opted for a qualitative analysis, using a visualization method that essentially produces a ranking of communication media. The idea is to visualize per study which communication media were found to be statistically significantly different, and then to combine those visualizations across studies.

Concerning the visualization per study, the number of ranks shown in each plot represents the number of communication media that statistically differed. In addition, the visualization method also illustrates whether the media significantly differed in a study or not by means of hypothetical errorbars: overlapping errorbars represent non-significant differences, adjacent errorbars represent significant differences. This means, even though the ranking and sizes of those errorbars are based on the quantitative results of a study, those errorbars are a qualitative visualization and do not represent the real study data.

Furthermore, the ranking, expressed by the middle points of the errorbars, and the width of the errorbars needed to be constructed such that a proper across-study comparison is possible. This is not trivial since different studies tested different combinations of communication media. The chosen approach was to normalize per individual study the position and widths of the errorbars to the number of ranks that were found.

Figure 2.4 shows five representative hypothetical examples for studies that investigated three communication media. Furthermore, it explains per example how the visualization is constructed.

The plots in Figure 2.4 are examples for individual studies. The next step is to combine such plots across studies in one figure such that the impact of individual communication media across studies is visible. For that reason the errorbars are grouped per communication medium, Figure 2.5 gives an example combining three hypothetical studies, and provides an explanation how to interpret such a plot. Furthermore, the analysis was conducted separately for studies that investigated two-party communication and studies that investigated multiparty communication.

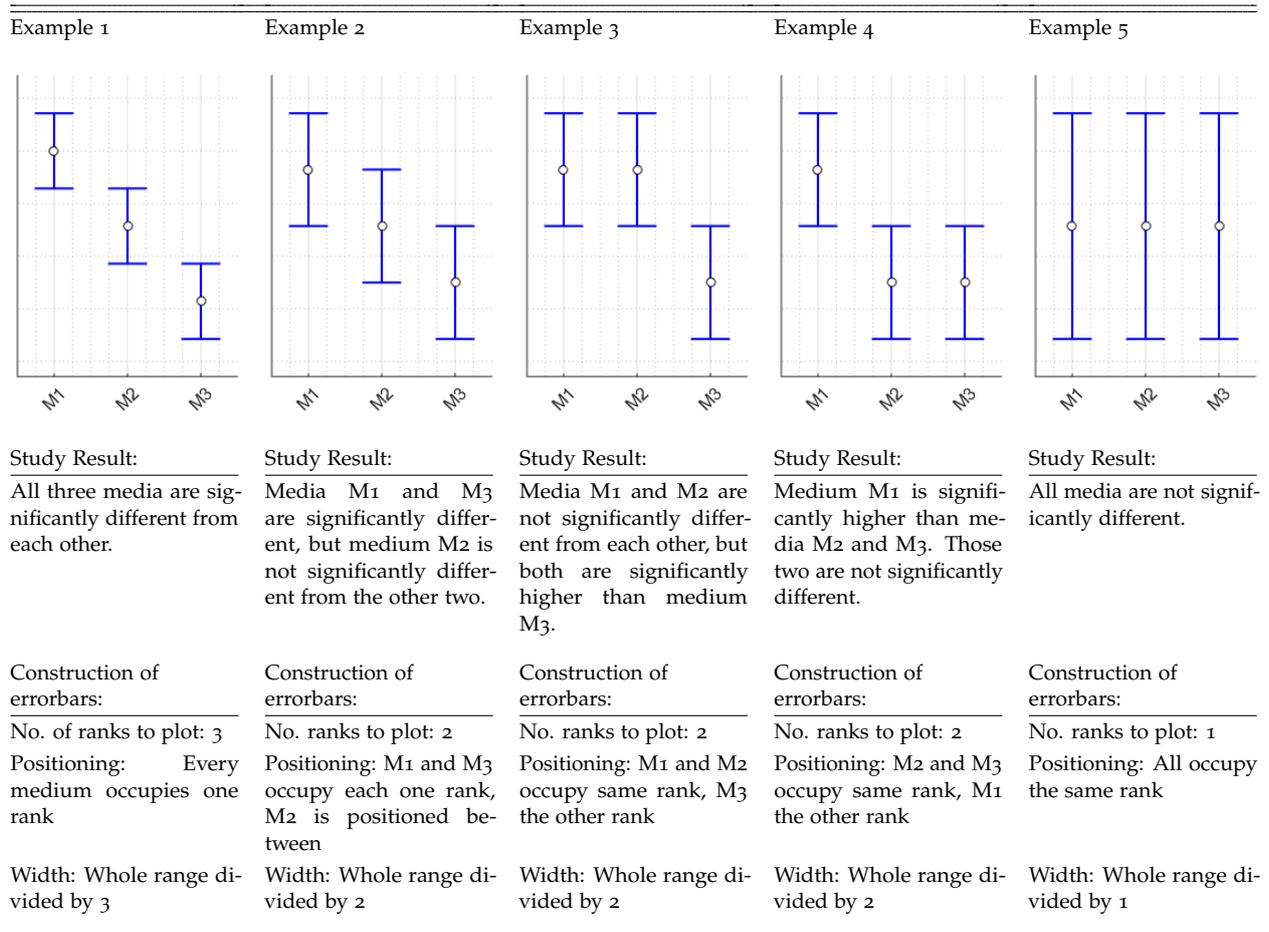
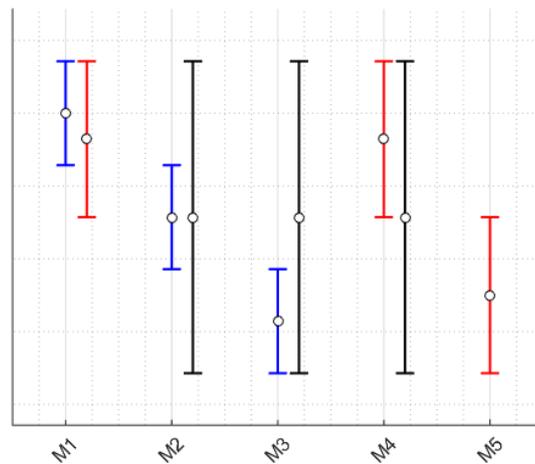


Figure 2.4: Hypothetical examples to visualize the impact of communication media on experimental results.



Combination of the results of studies “Blue”, “Red”, and “Black”:

Study “Blue” investigated media M<sub>1</sub>, M<sub>2</sub>, and M<sub>3</sub>, and it showed a pattern corresponding to Example 1 in Figure 2.4.

Study “Red” investigated media M<sub>1</sub>, M<sub>4</sub>, and M<sub>5</sub>, and it showed a pattern corresponding to Example 3 in Figure 2.4.

Study “Black” investigated media M<sub>2</sub>, M<sub>3</sub>, and M<sub>4</sub>, and it showed a pattern corresponding to Example 5 in Figure 2.4.

Interpretation Method:

1. Look at the positions and sizes of the errorbars per medium  $M_x$ .
2. The more the errorbars for the medium  $M_x$  are clustered towards the upper end, the more likely  $M_x$  leads to an increased value of the current measure (e.g. collaboration performance) compared to other media. Example: M<sub>1</sub>.
3. The more the errorbars are clustered toward the lower end, the more likely  $M_x$  leads to a decreased value of the current measure (e.g. collaboration performance). Example: If there were more studies with a pattern as study “Red”, then M<sub>5</sub>.
4. The more centrally positioned and at the same time the smaller the errorbars are for the medium  $M_x$ , the more likely  $M_x$  has an average value of the current measure (e.g. collaboration performance) compared to other media. Example: If there were more studies with a pattern as study “Blue”, then M<sub>2</sub>.
5. The more centrally positioned and at the same time the wider the errorbars are for the medium  $M_x$ , the more likely  $M_x$  does not differ in the current measure (e.g. collaboration performance) compared to other media. Example: If there were more studies with a pattern as study “Black”, then M<sub>2</sub>.
6. The more the errorbars of individual studies are spread for medium  $M_x$ , the more likely the impact of  $M_x$  on the current measure (e.g. collaboration performance) depends on the specific characteristics of the corresponding studies. Example: If M<sub>2</sub>, M<sub>3</sub>, and M<sub>4</sub> would represent three technical implementations of the same medium, then this combination would be an example.

Figure 2.5: Hypothetical example to combine results of different studies in one plot.

### 2.4.3 Data

The analysis comprised 44 experiments from 42 publications; see Table 2.4 for the bibliographic details. In those experiments 45 different instantiations of communication media were tested, covering face-to-face, audiovisual, audio-only, text-only communication, with and without shared workspace, covering different technical versions per communication medium, and including special conditions such as video only, virtual avatar and even no communication<sup>38</sup>. In total, 253 measures (dependent variables) were collected across the studies, whereas just two measures were used in more than 10 experiments: discussion or completion time in 17 studies, and number of turns or utterances in 13 studies. Fourteen further measures were used in two to seven experiments, and all remaining measures were used in one study only. For that reason, the present analysis grouped the measures into 24 categories of related measures, which cover four principle measurement paradigms:

- A: task performance in terms of outcome, quality and efficiency,
- B: conversational statistics not relying on analysis of content and function,
- C: conversational coding relying on analysis of content and function, and
- D: questionnaire answers, i.e. perceived aspects.

In order to avoid skewed visualizations, this categorization required some care not only in terms of assigning which individual measures to which group but also in terms of correctly interpreting the direction of such related measures, i.e. when one measure reflects the inverse or opposite of another measure. For that reason, the visualizations aim for maximum transparency by including a list of the current individual measures and by marking those measures that are inverted for the plot.

### 2.4.4 Results

*Overview* The full analysis resulted in 46 plots that are presented in 23 figures in Appendix B, noting that

- (a) there were in total 24 groups of measures,
- (b) the results for the first category *Task Outcome* was split into four plots, each comprising related tasks that were used,
- (c) not all categories were tested in both, two-party and multiparty studies,
- (d) each figure has comprehensive legends providing identifiers of the experiments ([Author\_Year\_ExperimentInPublication]), the shown measures, and the individual communication media, and

<sup>38</sup> In such conditions, participants saw each other shortly before the task, and then did the task individually.

- Anne H. Anderson et al. "Virtual team meetings: An analysis of communication and context". In: *Computers in Human Behavior* 23 (2007), pp. 2558–2580
- Mark Apperley and Masood Masoodian. "An Experimental Evaluation of Video Support for Shared Work-Space Interaction". In: *Proceedings of CHI 1995*. 1995
- Jennifer L. Blaskovich. "Exploring the Effect of Distance: An Experimental Investigation of Virtual Collaboration, Social Loafing, and Group Decisions". In: *Journal of Information Systems* 22.1 (2008), pp. 27–46
- Erin Bradner and Gloria Mark. "Why Distance Matters: Effects on Cooperation, Persuasion and Deception". In: *Proceedings of CSCW 2002*. 2002
- Jeannette Brosig, Joachim Weimann, and Axel Ockenfels. "The Effect of Communication Media on Cooperation". In: *German Economic Review* 4.2 (2003), pp. 217–241
- Judee K. Burgoon et al. "Trust and Deception in Mediated Communication". In: *Proceedings of the 36th Hawaii International Conference on System Sciences*. 2003
- Milton Chen. "Conveying Conversational Cues Through Video". PhD Thesis. Stanford University, 2003
- Joanie B. Connell et al. "Effects of communication medium on interpersonal perceptions". In: *Proceedings of the 2001 International ACM SIG-GROUP Conference on Supporting Group Work*. ACM. 2001, pp. 117–124
- Marcus Crede and Janet A. Sniezek. "Group judgment processes and outcomes in video-conferencing versus face-to-face groups". In: *International Journal of Human-Computer Studies* 59 (2003), pp. 875–897
- Owen Daly-Jones, Andrew Monk, and Leon Watts. "Some advantages of video conferencing over high-quality audio conferencing: fluency and awareness of attentional focus". In: *International Journal of Human-Computer Studies* 49 (1998), pp. 21–58
- Gwyneth Doherty-Sneddon et al. "Face-to-Face and Video-Mediated Communication: A Comparison of Dialogue Structure and Task Performance". In: *Journal of Experimental Psychology: Applied* 3.2 (1997), pp. 105–125
- Eckehard Doerry. "An Empirical Comparison of Copresent and Technologically-Mediated Interaction based on Communicative Breakdown". PhD Thesis. Department of Computer and Information Science, Graduate School of the University of Oregon, 1995
- Thomas Erickson et al. "Telepresence in Virtual Conferences: An Empirical Comparison of Distance Collaboration Technologies". In: *Proceedings of CSCW 2010*. 2010
- Susan R. Fussell, Robert E. Kraut, and Jane Siegel. "Coordination of Communication: Effects of Shared Visual Context on Collaborative Work". In: *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. 2000
- Susan R. Fussell, Leslie D. Setlock, and Robert E. Kraut. "Effects of Head-Mounted and Scene-Oriented Video Systems on Remote Collaboration on Physical Tasks". In: *Proceedings of the Conference on human factors in computing systems CHI 03*. 2003
- Darren Gergle, Robert E. Kraut, and Susan R. Fussell. "Language Efficiency and Visual Technology – Minimizing Collaborative Effort with Visual Information". In: *Journal of Language and Social Psychology* 23.4 (Dec. 2004), pp. 1–27
- Michel J. J. Handgraaf et al. "Web-conferencing as a viable method for group decision research". In: *Judgment and Decision Making* 7.5 (2012), pp. 659–668
- Matthew Jackson et al. "Impact of video frame rate on communicative behaviour in two and four party groups". In: *Proceedings of the 2000 ACM conference on Computer Supported Cooperative Work*. ACM. 2000, pp. 11–20
- Robert E. Kraut, Susan R. Fussell, and Jane Siegel. "Visual Information as a Conversational Resource in Collaborative Physical Tasks". In: *Human-Computer Interaction* 18 (2003), pp. 13–49
- Andreas Löber, Sibylle Grimm, and Gerhard Schwabe. "Audio vs. chat: Can media speed explain the differences in productivity?" In: *Proceedings of the European Conference on Information Systems*. 2006, pp. 2172–2183
- Andreas Löber, Gerhard Schwabe, and Sibylle Grimm. "Audio vs. chat: The effects of group size on media choice". In: *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*. IEEE. 2007, pp. 41–41
- Masood Masoodian, Mark Apperley, and Lesley Frederickson. "Video support for shared work-space interaction: an empirical study". In: *Interacting with Computers* 7.3 (1995), pp. 237–253
- Masood Masoodian and Mark Apperley. "User perceptions of human-to-human communication modes in CSCW environments". In: *Proceedings of ED-MEDIA'95*. 1995, pp. 17–21
- Masood Masoodian. "Human-to-Human Communication Support for Computer-Based Shared Workspace Collaboration". PhD Thesis. Department of Computer Science, The University of Waikato, Hamilton, New Zealand, 1996
- Giacinto Matarazzo. "The effects of video quality and task on remote collaboration: a laboratory experiment". In: *17th International Symposium on Human Factors in Telecommunication*. Copenhagen, Denmark, May 1999
- Poppy Laretta McLeod et al. "The Eyes Have It: Minority Influence in Face-To-Face and Computer-Mediated Group Discussion". In: *Journal of Applied Psychology* 82.5 (1997), pp. 706–718
- David Nguyen and John Canny. "MultiView: Spatially Faithful Group Video Conferencing". In: *Proceedings of CHI 2005*. 2005
- Brid O'Conaill, Steve Whittaker, and Sylvia Wilbur. "Conversations Over Video Conferences: An Evaluation of the Spoken Aspects of Video-Mediated Communication". In: *Human-Computer Interaction* 8 (1993), pp. 389–428
- Claire O'Malley et al. "Comparison of face-to-face and video-mediated interaction". In: *Interacting with Computers* 8.2 (1996), pp. 177–192
- Alison Sanford, Anne H. Anderson, and Jim Mullin. "Audio channel constraints in video-mediated communication". In: *Interacting with Computers* 16 (2004), pp. 1069–1094
- Abigail J. Sellen. "Speech Patterns in Video-Mediated Conversations". In: *Proceedings of CHI 1992*. 1992
- Diane H. Sonnenwald, Mary C. Whitton, and Kelly L. Maglaughlin. "Evaluating a scientific collaboratory: Results of a controlled experiment". In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 10.2 (2003), pp. 150–176
- Diane H. Sonnenwald, Kelly L. Maglaughlin, and Mary C. Whitton. "Designing to support situation awareness across distances: an example from a scientific collaboratory". In: *Information Processing and Management* 40 (2004), pp. 989–1011
- Susan G. Straus and Joseph E. McGrath. "Does the Medium Matter? The Interaction of Task Type and Technology on Group Performance and Member Reactions". In: *Journal of Applied Psychology* 79.1 (1994), pp. 87–97
- Kil Soo Suh. "Impact of communication medium on task performance and satisfaction: an examination of media-richness theory". In: *Information & Management* 35 (1999), pp. 295–312
- Federico Tajariol, Jean-Michel Adam, and Michel Dubois. "Seeing the Face and Observing the Actions: the Effects of Nonverbal Cues on Mediated Tutoring Dialogue". In: *Intelligent tutoring systems*. Vol. 5091. Lecture Notes in Computer Science. Springer, 2008, pp. 480–489
- Lori Foster Thompson and Michael D. Covert. "Teamwork Online: The Effects of Computer Conferencing on Perceived Confusion, Satisfaction, and Postdiscussion Accuracy". In: *Group Dynamics: Theory, Research, and Practice* 7.2 (2003), pp. 135–151
- Lori Foster Thompson and Michael D. Covert. "Stepping Up to the Challenge: A Critical Examination of Face-to-Face and Computer-Mediated Team Decision Making". In: *Group Dynamics: Theory, Research, and Practice* 6.1 (2002), pp. 52–64
- Elizabeth S. Veinott et al. "Video Helps Remote Work: Speakers Who Need to Negotiate Common Ground Benefit from Seeing Each Other". In: *Proceedings of CHI 1999*. 1999
- Peter J. Werkhoven, Jan Maarten Schraagen, and Patrick A.J. Punte. "Seeing is believing: communication performance under isotropic teleconferencing conditions". In: *Displays* 22 (2001), pp. 137–149
- Steve Whittaker and Brid O'Conaill. "An Evaluation of Video Mediated Communication". In: *INTERACT '93 and CHI '93 Conference Companion on Human Factors in Computing Systems*. ACM. 1993, pp. 73–74
- Wanmin Wu et al. "Quality of Experience in Distributed Interactive Multimedia Environments: Toward a Theoretical Framework". In: *Proceedings of the 17th ACM international conference on Multimedia*. 2009

Table 2.4: Bibliographic details of the 42 publications included in the literature analysis.

- (e) each figure includes the main finding based on the interpretation method of Figure 2.5.

Table 2.5 summarizes those detailed results.

*Measures that characterize communication* If a category of measures shows strong and consistent trends in terms of significantly different results for different communication media, than it can be concluded that this type of measures is characterizing communication. However, none of the categories shows such trends for both two-party and multiparty situations, which means that the literature sample does not allow to identify a generally valid type of measures. Instead, some types of measures do characterize either two-party or multiparty situations. This means that the literature sample is not sufficient to show consistent trends for both situations, which is likely since the amount of data did not allow for proper conclusions in many cases. It can also mean that two-party and multiparty communication are such different that different aspects of communication (i.e. different types of measures) are influenced, which is also likely since some measures, e.g. Category B.2, show an effect for one situation but not for the other.

A further aspect that complicates the picture is that in those cases in which an effect of medium was found, not all tested media showed that effect, e.g. Category B.3 for multiparty situations. That means there is likely some interaction between actual measures and tested communication medium.

Finally, Table 2.5 reveals a further aspect that is that results are – unfortunately – ambiguous to a certain extent. In several cases some evidence for an effect could be found, while in detail those effects are accompanied with inconsistent results, e.g. Category A.2 for two-party situations. Similarly, in several cases the evidence against an effect is also compromised by inconsistent results, e.g. Category B.1 for two-party situations.

#### 2.4.5 *Discussion and Conclusions*

The results of the literature analysis are discouraging as they did not show a clear result of the form: “measures x, y and z are consistently influenced by the communication medium, hence these measures do characterize group communication”. On the one hand, this can be in parts explained by the size of the literature sample, which resulted in small numbers of studies per such group of measures that did not allow to draw proper conclusions. On the other hand, this can be also explained by strong inconsistencies of results across studies, which could be observed for many cases that have sufficient data for proper analysis.

Looking at this from an engineering perspective, the two aspects sample size and amount of inconsistencies suggest that there is a very large amount of measurement noise. This means a vast amount of data is necessary to identify a small set of generally characteristic

Measures	Two-Party Studies		Multiparty Studies	
	Result	Comments	Result	Comments
A.1.1) Positive task outcome for collaboration and survival tasks	?	Insufficient data for proper interpretation	✓	Trend: f2f differs from audiovisual
A.1.2) Positive task outcome for design, decision, minority opinion, mystery, and estimation task	?	Insufficient data for proper interpretation	(✓)	Trend: f2f differs from text-only but not consistent: often no differences, or differences not in all studies
A.1.3) Positive task outcome for map task	(×)	Not consistent: often no differences, or differences not in all studies, or contrary differences between studies	?	No data
A.1.4) Positive task outcome for helper-worker tasks	(×)	Not consistent: often no differences, or differences not in all studies, or contrary differences between studies	?	No data
A.2) Task completion time	(✓)	Trend: f2f < audiovisual < text-only, no trend for audio-only but not consistent: often no differences, or differences not in all studies	×	No effect
B.1) Turn-taking rate and related measures	(×)	Not consistent: often no differences, or differences not in all studies, or contrary differences between studies	(×)	Not consistent: often no differences, or differences not in all studies, or contrary differences between studies
B.2) Word rate and related measures	✓	Trend: f2f < audiovisual, no trend for audio-only	×	No effect
B.3) Amount of overlap and speaker-state measures	×	No effect	✓	Trend: f2f > audio-only, no trend for audiovisual
B.4) Amount of speaker changes	×	No effect	(×)	Not consistent: contrary differences between studies
C.1) Amount of back channels	?	Insufficient data for proper interpretation	?	Insufficient data for proper interpretation
C.2) Amount of instruction	✓	Trend: interaction of medium and availability of shared-workspace: audiovisual with shared workspace > audio-only with shared workspace, but audiovisual without shared workspace ≈ audio-only with shared workspace	?	No data
C.3) Amount of aligns and checks	(✓)	Trend: f2f requires relative consistently a smaller amount of aligns and checks while other media require sometimes higher, sometimes smaller amount of aligns and checks but not consistent: often no differences, or contrary differences between studies	?	No data
C.4) Amount of explanations	×	No effect	?	Insufficient data for proper interpretation
C.5) Amount of questions	(✓)	Trend: f2f requires relative consistently a smaller amount of questions while other media require sometimes higher, sometimes smaller amount of questions but not consistent: often no differences, or differences not in all studies, or contrary differences between studies	?	Insufficient data for proper interpretation
C.6) Amount of acknowledgments	(×)	Not consistent: often no differences, or differences not in all studies	?	No data
C.7) Amount of object, process and status utterances	(×)	Not consistent: often differences not in all studies, or contrary differences between studies	?	No data
C.8) Amount of attempts to manage the conversation or amount of conversation failures	?	Insufficient data for proper interpretation	?	Insufficient data for proper interpretation
D.1) Satisfaction	(×)	Not consistent: often no differences, or differences not in all studies, or contrary differences between studies	✓	Trend: f2f and audiovisual higher satisfaction than audio-only and text-only
D.2) Perceived media richness	?	Insufficient data for proper interpretation	?	No data
D.3) Amount of modifications and exchange of opinions	?	No data	×	No effect
D.4) Perception of positive work process and collaboration	?	Insufficient data for proper interpretation	(×)	Not consistent: often differences not in all studies, or contrary differences between studies
D.5) Perception of positive grounding processes	?	Insufficient data for proper interpretation	?	Insufficient data for proper interpretation
D.6) Perception of interlocutors	?	Insufficient data for proper interpretation	✓	Trend: interlocutors are more positively perceived for f2f communication vs. audio-only and other media, with audiovisual in-between
D.7) Perception of engagement	?	Insufficient data for proper interpretation	?	Insufficient data for proper interpretation
D.8) Perception of trust	?	Insufficient data for proper interpretation	?	Insufficient data for proper interpretation
D.9) Perception of positive turn-taking processes	?	Insufficient data for proper interpretation	(×)	Not consistent: often differences not in all studies, or contrary differences between studies
D.10) Positive perception of visual cues	?	Insufficient data for proper interpretation	?	Insufficient data for proper interpretation

Legend: ✓: communication media show a trend for an impact on a type of measures; (✓): weak trend; (×): non-consistent behavior; ×: no effect; ?: no results/conclusions possible

Table 2.5: Summary of literature analysis, based on detailed results shown in Appendix B.

measures, if such a set actually exists.

With regard to the literature reviews of Fjermestad & Hiltz (see previous section), two observations can be made here: First, even though the goal of the present literature review – identify measures that characterize communication – differed from the goals of Fjermestad & Hiltz – investigate added value of mediated communication versus face-to-face communication – the fundamental result is quite similar: no consistent effect of communication medium for most parts of the considered data. Second, the reviews of Fjermestad & Hiltz covered about 200 studies, which can be considered as quite exhaustive, and they still found no strong effects favoring either mediated or face-to-face communication.

This suggests that simply increasing the sample of the current literature review may also not be enough to reduce the measurement noise for the present goal as well. Instead, it may be necessary to develop a more comprehensive framework that provides a list of the desired characteristic measures together with a systematic description of the context (e.g. communication medium, two-party vs. multiparty, tasks) for which those measures work. The overview in Table 2.5 as well as the detailed results in Appendix B may be used as a starting point for such a project, which is out of scope of the present thesis.

### *Summary*

One way to characterize group-communication (including a group size of two) is to address its purpose, the processes involved, and the medium used. The purpose can be focussed on accomplishing tasks, socializing, or exchange of information, and a widely used approach for finer distinction of task-oriented communication is McGrath's task circumplex. Communication processes can be characterized by five prominent aspects: turn-taking, listener feedback signals, conversational surface structure, grounding, and conversational games and moves. These aspects of analysis address communication from different perspectives, which the present chapter categorized along five dimensions: temporal granularity (moment-by-moment, in retrospect), organizational focus (temporal order of contributions, organization of content), reliance on the content (yes, no), characterization depth (describing intrinsic property, providing measurement tool), and focus on the persons (speaker, listener, both). Communication media are determined by the modality in terms of face-to-face, audiovisual, audio-only, and text-only, as well as on the technical realization. Furthermore, the availability of a shared workspace has a crucial impact on communication and therefore constitutes an important aspect of the communication medium.

Comparing face-to-face and mediated communication, an extensive literature review by Fjermestad & Hiltz revealed that reviewed studies can be roughly equally divided into studies favoring face-to-face, studies favoring mediated communication, and studies showing no significant effects. This means that the measurable aspects of the com-

munication processes and their outcomes did not differ consistently enough to give a clear picture. However, it is still likely that people assume a different communication behavior when conversing over a telecommunication system in order to reach the communication goals despite the limitations of the particular communication channel, whereas those changes in communication behavior could not be consistently measured across studies.

An own literature analysis did not lead to a result of the form: "measures x, y and z are consistently influenced by the communication medium, hence these measures do characterize group communication". Instead, the analysis revealed that the sample size, i.e. number of reviewed studies, must be sufficiently large and that there is a large amount of inconsistencies in those cases for which the sample was large enough. This suggests that there is a large amount of measurement noise that needs to be properly addressed in order to identify such a set of generally characteristic measures. In future work, which is outside the scope of this thesis, it may be necessary to develop a more comprehensive framework that provides a list of the desired characteristic measures together with a systematic description of their mutual interaction and the context for which those measures work.

Concerning the present thesis, one of its objectives is to investigate the relation between communication aspects and quality perception in the context of multiparty telemeetings. To this aim, this thesis will not develop new approaches or measures to characterize mediated group-communication. Instead, two experimental studies will be conducted, in which participants judge the quality of telemeeting systems for few specific communication scenarios. For that purpose, the background knowledge presented in this chapter will be used to define those communication scenarios and the experimental variables.

# 3

## Telemeeting Technologies

### *What this chapter is about*

The present text is built around two aspects that differentiate a multiparty telemeeting from a conventional two-party (video) telephony call. One aspect is the possibility of encountering asymmetric connections, that is, participants are connected with different equipment or connection properties. Bearing this in mind, the present chapter provides background knowledge on the technical aspects and components that constitute a telemeeting system.

### 3.1 *Introduction and Scope*

Today's technology allows to establish multiparty communication between distant locations by very different means. Those are ranging from the traditional telephone conference bridges over mobile or PC-based audiovisual solutions to high-end telepresence rooms.

To organize this variety of possible systems, the present text starts with defining some coarse categories along two characteristics: *communication modality* and *system size*. Concerning the first characteristic, ITU-T Recommendation P.1301<sup>1</sup> defines five types of *communication modalities*: audio-only, video-only (i.e. for hearing-impaired people), audiovisual, text (i.e. chat), graphical (i.e. slide/screen sharing, referring to the concept of shared work space described in Section 2.2). Concerning the second characteristic, the term *system size* refers to the physical size and setup complexity of the end device(s) that the user needs for connecting to the telemeeting. Thus, it does not refer to any central or network components of the telemeeting system. In that line of thought, the present text distinguishes between single-device, small-size and large-size setups.

Single-device setups are those for which the user is just using one physical device: telephone, mobile phone, tablet, laptop etc. Small-size setups require more than one physical device, but which occupy only a rather small amount of space, i.e. work desk, part of small office or living room. The typical example is a computer-based setup with connections to external camera, microphone(s), display, and loudspeakers. Large-size setups require more than one physical device and occupy a large amount of space, i.e. a dedicated room. The

<sup>1</sup> ITU-T. *Recommendation P.1301 - Subjective quality evaluation of audio and audiovisual telemeetings*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2012

typical example is a high-end telepresence room with multiple large-size displays, microphone and loudspeaker arrays, and dedicated furniture arrangement.

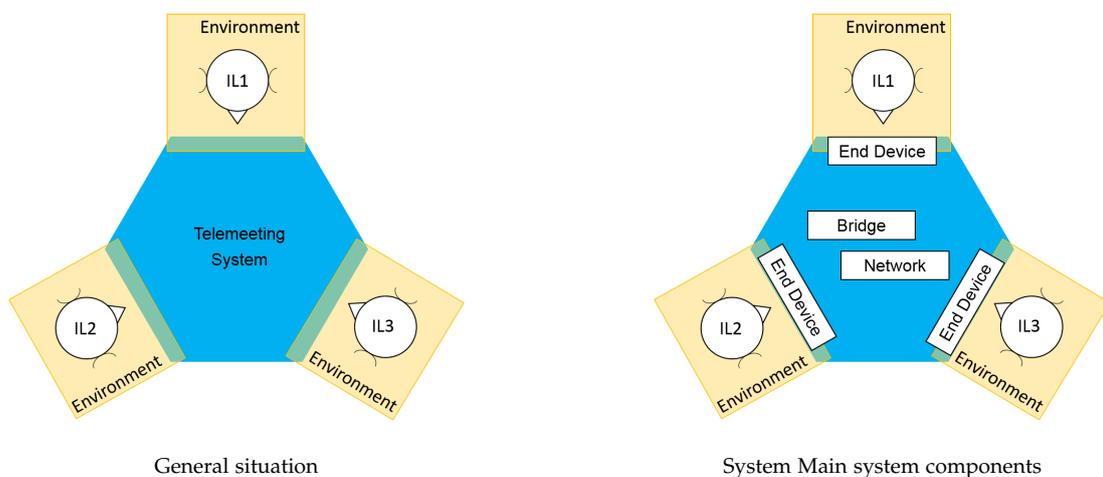
Obviously, those categories are not exhaustive, since there are further types of telemeeting systems that do not fit exactly into those categories. Two examples are mobile telepresence systems using robot avatars (see for instance Paepcke et al.<sup>2</sup>), and telemedicine systems (see for instance Kyriacou et al.<sup>3</sup>).

Nevertheless, with these three categories it is possible to define the scope of the present text: The following sections will provide an overview of the technical components of audio-only and audiovisual, single-device and small-size telemeeting systems.

### 3.2 Main Components

A multiparty telemeeting system connects multiple interlocutors, who are situated in at least two distant locations. The left panel of Figure 3.1 visualizes a situation for three interlocutors (IL1, IL2, IL3), in which every interlocutor is located in a different physical environment and all are connected via the same telemeeting system.

Going into more details, the right panel of Figure 3.1 adds the three main system components: the end device(s), the bridge (often called conferencing bridge), and the network. The end devices transduce the sound and light waves into electrical signals and prepare them for digital transmission over the network and vice versa. The bridge provides the multiparty functionality, i.e. it realizes for every interlocutor the mixing of the audio and video signals coming from the other interlocutors. The following sections describe those three components in more detail, starting with a description of different system topologies.



<sup>2</sup> Andreas Paepcke et al. "Yelling in the hall: using sidetone to address a problem with mobile remote presence systems". In: *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM. 2011, pp. 107–116

<sup>3</sup> E.C. Kyriacou, C.S. Pattichis, and M.S. Pattichis. "An overview of recent health care support systems for eEmergency and mHealth applications". In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2009)*. IEEE. 2009, pp. 1246–1249

Figure 3.1: Example setting of a three-party telemeeting and main components of the telemeeting system.

### 3.3 Central-Bridge and Peer-to-Peer Topologies

Telemeeting systems can be categorized into two main topologies: central-bridge and peer-to-peer. The two topologies are characterized by the location of the bridge, as visualized in Figure 3.2.

In a central-bridge topology, the telemeeting system uses one bridge which is located somewhere in the network. The end devices at the user side – henceforth referred to as client side – are sending the audiovisual data streams to that central bridge. The bridge generates the mixed signals for every interlocutor and sends them to the corresponding end devices.

In a peer-to-peer topology, every client side has an own bridge which is directly connected or even integrated into the end device. Each bridge ensures that the own audiovisual data is sent to all interlocutors and it processes (mixes) all incoming data streams for presentation via the end device.

The disadvantage of a peer-to-peer topology is the extra computational load for client side devices due to the bridge and the extra network traffic load due to the multiple data streams. Here, a central-bridge topology with dedicated servers has a clear advantage. However, operating central-bridge servers requires resources, which eventually increase the service subscription costs for the users.

While these two topologies describe the two main types of telemeeting architectures, hybrid solutions are also possible. Examples are peer-to-peer systems in which one client bridge serves also as a sub-bridge for further end devices that are not peer-to-peer-capable, or mixed systems in which the computational load is distributed between central and client-side bridges.

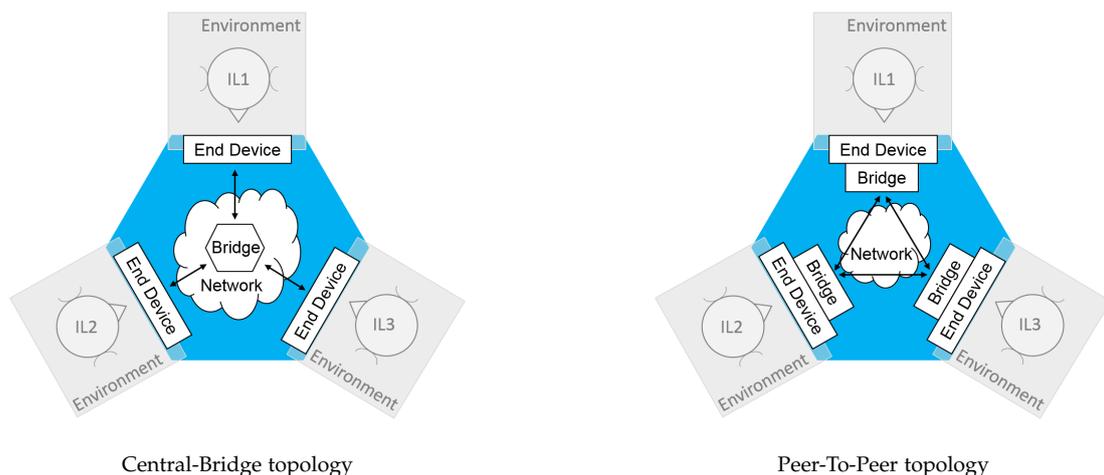


Figure 3.2: Visualization of the two main topologies for telemeeting systems.

### 3.4 End Devices

The variety of end devices used for audio and video capture at the send-side and reproduction at the receive-side is vast. Devices can

differ in terms of their form factors, the sound and video transduction principles, and the signal processing implemented in those devices for the purpose of signal enhancement.

### 3.4.1 *End Devices for Audio*

*Form factor* There are two general types of end devices: those which combine both sound capture and reproduction in one physical device, and those which require separate microphones and loudspeakers. Examples for the first type are telephones, mobile phones, tablets or laptops with integrated microphones and loudspeakers. Examples for the second type are computer-based setups that do not have integrated loudspeakers or microphones at all or only with an unsatisfactory quality.

End devices of the first type can be further categorized into three groups: handsets, headsets, hands-free terminals. Each of those types has advantages and disadvantages in terms of comfort and the possible signal quality. Typical comfort aspects comprise wearing comfort (of handsets and headsets) and setup complexity (e.g. cable connections, placement of devices).

Concerning the possible signal quality, the form factor influences the impact of the distance between mouth and microphone: the smaller the distance, the lower the acoustic problems such as environmental noise, room reverberation and echo. If the form factor does not allow for small mouth-to-microphone distances, as it is the case for modern mobile phones or hands-free terminals, then heavy signal enhancement (see below) is usually necessary.

It can be expected that different users have different preferences for either type, which may also depend on the actual situation. To account for this, modern telephone devices often combine all three form factors: they can be conventionally used as handsets, they are equipped with additional microphones and loudspeakers to enable a handsfree mode, and they provide connectors for a suitable headset.

Concerning the second type of devices, those that require separate microphones and loudspeakers, the complexity of the technical setup in terms of cables and software installation may become a comfort issue. In addition, the available degree of freedom in placing the microphones and loudspeakers can also impact signal quality.

In the context of the present research, the form factor plays a role when it comes to the assessment of the Quality of Experience beyond the assessment of pure signal quality. Referring to the book chapter by Reiter et al.<sup>4</sup>, there are many influencing factors, and it is straightforward to argue that the form factor belongs to such factors, given the discussed implications for comfort and signal quality. Furthermore, in case of using headsets, the form factor also plays a practical role when it comes to the technical setup of a laboratory experiment. If no echo cancellation is desired or available, an acoustic feedback from the headphones to the microphones needs to be avoided. Then headsets with close-talking microphones and closed headsets are the preferred

<sup>4</sup> Ulrich Reiter et al. "Factors Influencing Quality of Experience". In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 55-72

choice.

*Transducer Types* There are different principles to transduce sound waves to electric signals (microphones) and vice versa (loudspeakers). In the context of the present research, the transduction principle plays a role, if the transducer needs electric power supply. This can become a non-trivial issue when it comes to the technical setup of a laboratory experiment, in which different end devices shall be tested. As an example, high-end studio sound cards, which are the preferred choice to avoid any influence of the sound card, and consumer headsets are often not compatible: Most electret microphones integrated in the consumer market computer headsets require a small power supply of 1 - 10 Volt<sup>5</sup>, not for polarizing their membrane, but for an internal preamplifier. While on-board laptop sound cards usually provide this voltage, high-end studio sound cards do not provide it as they are compliant to other standards (+48 V phantom power for condenser microphones).

*Signal processing* In most modern audio end devices different signal processing algorithms are implemented for signal enhancement purposes. Table 3.1 lists some of the most common processing algorithms including a short description of their main characteristics. To summarize this table, the algorithms presented here are mainly concerned with the enhancement of the microphone signals, and most algorithms apply highly non-linear, non-standardized or even undisclosed processing stages. In the context of the present research, such signal processing plays a role when it comes to the technical setup of laboratory experiments in which the lack of control can become an issue, or when it comes to instrumental assessment of quality, in which the lack of standardization or available information on the processing can become an issue.

<sup>5</sup> Tomi Engdahl. *Powering Microphones*. retrieved on: August 13, 2015. 2012. URL: [http://www.epanorama.net/circuits/microphone\\_powering.html](http://www.epanorama.net/circuits/microphone_powering.html)

Algorithm	Basic Idea	Further Reading
<p><b>Acoustic Echo Cancellation</b></p> <p>Cancel echo caused by acoustic feedback from the loudspeaker to the microphone.</p>		<p>Peter Vary and Rainer Martin. <i>Digital Speech Transmission – Enhancement, coding and Error Concealment</i>. Chichester, West Sussex, UK: Wiley, 2006</p>
<p><b>Noise Reduction</b></p> <p>Reduce unwanted environmental noise which interferes with the desired speech signal.</p>		<p>Peter Vary and Rainer Martin. <i>Digital Speech Transmission – Enhancement, coding and Error Concealment</i>. Chichester, West Sussex, UK: Wiley, 2006</p>
<p><b>De-reverberation</b></p> <p>Reduce reverberation in the microphone signal caused by acoustic room reflections.</p>		<p>Patrick A. Naylor and Nikolay D. Gaubitch, eds. <i>Speech Dereverberation</i>. Berlin, Germany: Springer, 2010</p>
<p><b>Automatic Gain Control (Dynamic Range Control)</b></p> <p>Adjust too low, too high or varying speech levels to a comfortable and understandable level.</p>		<p>Udo Zölzer. <i>Digital Audio Signal Processing</i>. 2nd ed. Chichester, West Sussex, UK: Wiley, 2006</p>
<p><b>Side Tone Generation</b></p> <p>Simulate talker side tone, i.e. the normal acoustic feedback path from own mouth to own ear in case the acoustic interface is shielding one or both ears, such as for handsets or closed headphones. In addition, simulate listener side tone, i.e. the path for environmental sound into that ear that is shielded by the device.</p>		<p>Numerous Patents</p>
<p><b>Voice Activity Detection</b></p> <p>Avoid transmission of unnecessary data in time intervals in which the speaker does not talk by sending data only in case of active speech.</p>		<p>Antti Vahätalo and Ingemar Johansson. "Voice activity detection for GSM adaptive multi-rate codec". In: <i>Proceedings of IEEE Workshop on Speech Coding</i>. 1999, pp. 55–57</p> <p>Javier Ramirez, Juan Manuel Górriz, and José Carlos Segura. "Voice Activity Detection - Fundamentals and Speech Recognition System Robustness". In: <i>Robust Speech Recognition and Understanding</i>. Ed. by Michael Grimm and Kristian Kroschel. In-Tech Education and Publishing, 2007</p>
<p><b>Artificial Bandwidth Extension</b></p> <p>Increase sound experience for speech transmitted at low bandwidth.</p>		<p>Peter Vary and Rainer Martin. <i>Digital Speech Transmission – Enhancement, coding and Error Concealment</i>. Chichester, West Sussex, UK: Wiley, 2006</p>
<p><b>Beam-forming</b></p> <p>Reduce unwanted environmental noise which interferes with the desired speech signal.</p>		<p>Peter Vary and Rainer Martin. <i>Digital Speech Transmission – Enhancement, coding and Error Concealment</i>. Chichester, West Sussex, UK: Wiley, 2006</p>

Table 3.1: Overview of most common audio signal enhancement algorithms used in end devices.

### 3.4.2 End Devices for Video

*Form factor* Similar to audio devices, end devices for video have both video capture and reproduction in one physical device (e.g. mobile phones, tablets, laptops) or they require separate cameras and displays (e.g. desktop-computer-based setups). Accordingly, similar comfort and placement aspects apply as in the audio case.

However, the most important form factor aspect for both types of video devices is the issue of unnatural eye gaze, which is determined by the distance between camera and display. Since users usually look at the display and not directly into the camera, the other interlocutors get the impression that the user is looking somewhere else. While there is a lot of literature on the issue of eye gaze, Grayson and Monk<sup>6</sup> argue that users are able to learn to interpret if the interlocutor is looking at him or her. However, the data of Grayson and Monk also show that this has some limitations, especially in case of horizontal disparity of camera and display. Concerning the form factor discussed here, this means that if the distance between camera and display is too strong, then eye gaze becomes an issue despite this learning ability of users, and thus Quality of Experience can be affected.

*Signal processing* In most modern video end devices signal processing algorithms are implemented for signal enhancement purposes. Table 3.2 lists some of the most common processing algorithms including a short description of their main characteristics.

To summarize this table, the algorithms presented here are mainly concerned with the enhancement of the camera signals and optimizing them for coding, either by directly processing the signals or by controlling the camera. In the context of the present research, such signal processing plays a role when it comes to the technical setup of laboratory experiments in cases in which the signal enhancement can not be controlled as desired.

<sup>6</sup> David M. Grayson and Adrew F. Monk. "Are You Looking at Me? Eye Contact and Desktop Video Conferencing". In: *ACM Transactions on Computer-Human Interaction* 10.3 (Sept. 2003), pp. 221–243

Algorithm Purpose	Basic Idea	Further Reading
<b>Auto Focus</b> Automatically adjust the camera focus to obtain sharp pictures.	Iteratively change the camera focus and estimate the picture's sharpness until that sharpness is maximized. Typical methods for estimating sharpness: contrast detection or phase detection.	Junichi Nakamura, ed. <i>Image Sensors and Signal Processing for Digital Still Cameras</i> . Boca Raton, Florida, USA: Taylor & Francis Group, 2006
<b>Auto Exposure</b> Automatically adjust the camera exposure (aperture and shutter) time to let a proper amount of incident light falling onto the sensor.	Measure the amount of light falling onto the light sensor, done usually just before taking the real picture, and calculate the desired exposure time to allow an optimal usage of the camera's dynamic range.	Junichi Nakamura, ed. <i>Image Sensors and Signal Processing for Digital Still Cameras</i> . Boca Raton, Florida, USA: Taylor & Francis Group, 2006
<b>Auto White Balance</b> Provide the camera with a reference point for white to account for the fact that the human vision system adapts this white point depending on the context/scene.	Estimate from the picture the white point and then transform the RGB color data accordingly. Typical method for estimation: Combination of average of the scene (assumed to be middle gray), brightest color (assumed to be white), and color gamut of the scene (based on distribution of colors in the scene). Typical methods for adaptation: chromatic adaptation, color constancy.	Junichi Nakamura, ed. <i>Image Sensors and Signal Processing for Digital Still Cameras</i> . Boca Raton, Florida, USA: Taylor & Francis Group, 2006
<b>Automatic Gain Control</b> Improve the picture's contrast by modifying the pixel intensities.	Compute the pixel intensity histogram of the image and apply a transformation to achieve a desired histogram. Typical methods: full-scale histogram stretch / contrast stretch	AI Bovik, ed. <i>The Essential Guide to Image Processing</i> . London, UK: Elsevier Inc., 2009
<b>Noise Reduction</b> Reduce unwanted noise in the picture caused by the electronic components, often triggered by low light conditions.	Apply linear or non-linear processing to smooth the effect of noisy pixels. Typical methods: linear filters (e.g. moving average, lowpass, gaussian, multiscale approaches), non-linear filters (e.g. weighted median smoothers), processing combined with dedicated edge detection algorithms.	AI Bovik, ed. <i>The Essential Guide to Image Processing</i> . London, UK: Elsevier Inc., 2009
<b>Image Stabilizer</b> Remove motion blur caused by unwanted camera movements.	Use mechanical or optical stabilization (e.g. using floating lens controlled by gyroscopic sensors), or apply digital signal processing comprising the two main steps Motion Estimation and Motion Compensation.	Mohammad Javad Tanakian, Mehdi Rezaei, and Farahnaz Mohanna. "Digital Video Stabilization System by Adaptive Fuzzy Kalman Filtering". In: <i>Information Systems and Telecommunication 1.4</i> (Oct. 2013), pp. 223-232 Paresh Rawat and Jyoti Singhai. "Review of Motion Estimation and Video Stabilization techniques For hand held mobile video". In: <i>Signal and Image Processing : An International Journal (SIPIJ) 2.2</i> (June 2011), pp. 159-168
<b>Color Space Transformation and Chroma Subsampling</b> Optimize color representation for efficient coding by exploiting that the human visual system is less sensitive to color than to luminance.	Transform the incoming RGB data into another color space separating the luminance component from the chroma (i.e. color difference) components, and then store the chroma components with a lower number of bits. Typical method: Computation according to ITU-R Recommendation BT.601.	Iain E. G. Richardson. <i>H.264 and MPEG-4 Video Compression – Video Coding for Next-generation Multimedia</i> . Chichester, West Sussex, UK: Wiley, 2003 ITU-R. <i>Recommendation BT.601-7 - Studio encoding parameters of digital television for standard 4:3 and wide screen 16:9 aspect ratios</i> . International Standard. Geneva, Switzerland: International Telecommunication Union, 2011
<b>De-blocking and De-ringing filter</b> Reduce the impact of some typical coding artefacts, i.e. pixel blocks and edge distortions (ringing).	Apply linear or non-linear filters to smooth the effect of blocking and ringing, which can be put as post-processor after the decoder or inside the encoding and decoding scheme. Example methods: Computation according to MPEG4-2 Visual (formerly ISO/IEC14496-2) or ITU-R Recommendation H.264	Iain E. G. Richardson. <i>H.264 and MPEG-4 Video Compression – Video Coding for Next-generation Multimedia</i> . Chichester, West Sussex, UK: Wiley, 2003 ISO/IEC. <i>ISO/IEC 14496-2:2004 - Information technology – Coding of audiovisual objects – Part 2: Visual</i> . International Standard. International Organization for Standardization & International Electrotechnical Commission, 2004 ITU-T. <i>Recommendation H.264 - Advanced video coding for generic audiovisual services</i> . International Standard. Geneva, Switzerland: International Telecommunication Union, 2014

Table 3.2: Overview of most common video signal enhancement algorithms used in end devices.

### 3.5 Transmission over the Network

The principle process of transmission over an IP-based network is to encode and packetize the captured and enhanced audio and video signals at the send side (usually done in the end device), to send the data packets over the network, and to depacketize and decode the data streams at the receive side in order to obtain the audio and video signals for reproduction.

#### 3.5.1 Codecs

The purpose of the audio and video codecs is to reduce the amount of data, i.e. number of bits, that are needed for transmission. There is a vast amount of codecs for speech, general audio and video available; descriptions of the most relevant codecs for telemeetings can be found in the G- and H-Series of the ITU-T <sup>7</sup> and MPEG1 and MPEG4 of the Moving Pictures Experts Group <sup>8</sup>.

This section text will focus on the most relevant codecs for the present research context of telemeeting quality, and refers to Raake <sup>9</sup>, Wältermann <sup>10</sup>, Belmudez <sup>11</sup> and Garcia <sup>12</sup> for further reading on existing codecs in the context of quality assessment for IP-based solutions, and to Chu <sup>13</sup>, Vary and Martin <sup>14</sup>, Richardson <sup>15</sup>, and Hanzo et al. <sup>16</sup> for further reading on the technologies.

*Speech Codecs* The basic principle of speech codecs is to exploit characteristics of the speech signal or human speech production process. Conceptually, this aspect separates speech codecs from the codecs for music and general audio, such as MPEG-1 Layer 3 (mp3) <sup>17</sup> or Advanced Audio Coding (AAC) <sup>18</sup>, which in turn exploit characteristics of the human auditory system. Since music transmission is out of scope, the present text focuses on a brief introduction to the principles used in speech codecs.

The simplest approach used for speech coding is to optimize the quantization steps according to the non-uniform amplitude histogram of the speech signal, which is done for example by the ITU-T G.711 codec <sup>19</sup>.

More advanced approaches apply the technique of linear prediction, which is used in three categories of speech codecs: waveform codecs, parametric codecs, and hybrid codecs. Common to all three types is to represent at encoder side each input speech sample as a linear combination of its past samples plus a residual part. Technically this leads to the computation of a set of filter coefficients representing the linear weighting, and the construction of a residual signal. Then data reduction is possible by exploiting the reduced dynamic range of the residual signal, as this is essentially a difference signal which has much lower amplitudes than the original speech signal.

For the first category, the waveform codecs, the processing steps comprise at encoder side the computation of the residual signal based on linear prediction and its quantization, on the channel the trans-

<sup>7</sup> ITU-T. *ITU-T Recommendations by series*. retrieved on August 19, 2015. International Telecommunication Union. 2015. URL: <http://www.itu.int/ITU-T/recommendations/index.aspx>

<sup>8</sup> Moving Pictures Expert Group. *MPEG Standards*. retrieved on August 19, 2015. International Organization for Standardization & International Electrotechnical Commission (ISO/IEC). 2013. URL: <http://mpeg.chiariglione.org/standards>

<sup>9</sup> Alexander Raake. *Speech Quality of VoIP – Assessment and Prediction*. Chichester, West Sussex, UK: Wiley, 2006

<sup>10</sup> Marcel Wältermann. *Dimension-based Quality Modelling of Transmitted Speech*. Springer, 2013

<sup>11</sup> Benjamin Belmudez. *Assessment and Prediction of Audiovisual Quality for Videotelephony*. Springer, 2009

<sup>12</sup> Marie-Neige Garcia. *Parametric Packet-based Audiovisual Quality model for IPTV services*. Springer, 2014

<sup>13</sup> Wai C. Chu. *Speech Coding Algorithms - Foundation and Evolution of Standardized Coders*. Hoboken, New Jersey, USA: Wiley, 2014

<sup>14</sup> Peter Vary and Rainer Martin. *Digital Speech Transmission – Enhancement, coding and Error Concealment*. Chichester, West Sussex, UK: Wiley, 2006

<sup>15</sup> Iain E. G. Richardson. *H.264 and MPEG-4 Video Compression – Video Coding for Next-generation Multimedia*. Chichester, West Sussex, UK: Wiley, 2003

<sup>16</sup> Lajos Hanzo, Peter J. Cherriman, and Jürgen Streit. *Video Compression and Communications – From Basics to H.261, H.263, H.264, MPEG4 for DVB and HSDPA-Style Adaptive Turbo-Transceivers*. Chichester, West Sussex, UK: Wiley, 2007

<sup>17</sup> ISO/IEC. *ISO/IEC 11172-3:1993 - Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s – Part 3: Audio*. International Standard. International Organization for Standardization & International Electrotechnical Commission, 1993

ISO/IEC. *ISO/IEC 13818-7:1995 - Information technology – Generic coding of moving pictures and associated audio information – Part 3: Audio*. International Standard. International Organization for Standardization & International Electrotechnical Commission, 1995

<sup>18</sup> ISO/IEC. *ISO/IEC 13818-7:2006 - Information technology – Generic coding of moving pictures and associated audio information – Part 7: Advanced Audio Coding (AAC)*. International Standard. International Organization for Standardization & International Electrotechnical Commission, 2006

<sup>19</sup> ITU-T. *Recommendation G.711 - Pulse Code Modulation (PCM) of voice frequencies*. International Standard. Geneva, Switzerland: International Telecommunication Union, 1988

mission of the quantized residual signal, and at decoder side the reconstruction of the output signal from the residual signal again by linear prediction. The predominant waveform codec is adaptive differential pulse code modulation (ADPCM) which computes as the residual signal the simple difference between two consecutive samples and applies adaptive, i.e. signal-dependent, quantization of that signal. Common codecs using ADPCM are the ITU-T G.726 codec<sup>20</sup>, which is typically applied in DECT (Digital Enhance Cordless Telecommunications) telephones, and the ITU-T G.722<sup>21</sup>, which is typically applied in Voice-over-IP (VoIP) telephony.

For the second category, the parametric codecs, the input signal is completely replaced by a parametric representation. For this purpose, the source-filter-model of human speech production is exploited: a source/excitation signal is produced by the human glottis and then filtered by the human vocal tract. There are different technical implementations possible<sup>22</sup>, such as channel vocoders and the LCP vocoder. Focussing on the LPC vocoder, the encoder first applies linear prediction to model the vocal tract, which is parameterized by the set of the computed filter coefficients. Then the encoder extracts a parametric description of the residual signal, which represents the excitation signal, by detecting whether the current input signal is a voiced or unvoiced signal, and estimates gain factors as well as for the voiced signal the pitch period. With this approach, only the parameters (LP filter coefficients, excitation parameters) are transmitted. At the decoder side, the excitation signal is first constructed using noise for unvoiced or pulse trains for voiced signal frames, which is then fed into a linear prediction stage to construct the output signal.

For the third category, the hybrid codecs, combine both principles by transmitting linear prediction coefficients as side information as well as the residual signal. Of the many possible variants, one important approach is to replace the residual signal by a sequence of preset signals taken from a code book (referred to as CELP: code-excited linear prediction), whereas this replacement is based on a search procedure finding those code book entries that minimize the perceptual error of the synthesized signal (therefore also referred to as Analysis-by-Synthesis coding). A common codec using CELP is the ITU-T G.729 codec<sup>23</sup>, which is applied in VoIP telephony.

Another very relevant codec for VoIP telephony is SILK<sup>24</sup>, which is also based on the CELP approach, but extended with a number of optimizations such as, for example, a noise shaping algorithm to minimize the audibility of the quantization noise. One of the key features of SILK is to provide speech transmission at different audio signal bandwidth, ranging from narrowband (NB: 300 Hz - 3400 Hz) to superwideband (SWB: 20 Hz - 12 kHz). A successor of SILK is OPUS<sup>25</sup>, which is extended with an additional coding scheme<sup>26</sup> that is more suited for music and general audio signals.

Another variation of the CELP approach is ACELP, algebraic code-excited linear prediction, which extends CELP with a patented method<sup>27</sup> to optimize the code book structure. Common codecs

<sup>20</sup> ITU-T. *Recommendation G.726 - 40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM)*. International Standard. Geneva, Switzerland: International Telecommunication Union, 1990

<sup>21</sup> ITU-T. *Recommendation G.722 - 7 kHz audio-coding within 64 kbit/s*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2012

<sup>22</sup> Peter Vary and Rainer Martin. *Digital Speech Transmission – Enhancement, coding and Error Concealment*. Chichester, West Sussex, UK: Wiley, 2006

<sup>23</sup> ITU-T. *Recommendation G.729 - Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2012

<sup>24</sup> Koen Vos, Søren Skak Jensen, and Karsten Vandborg Sørensen. *SILK speech codec*. International Standard. Internet Engineering Task Force (IETF), 2010. URL: <https://tools.ietf.org/id/draft-vos-silk-02.txt>

<sup>25</sup> Koen Vos et al. "Voice Coding with OPUS". in: *Proceedings of the 135th AES Convention*. New York, USA, Oct. 2013

<sup>26</sup> Jean-Marc Vallin et al. "High-Quality, Low-Delay Music Coding in the OPUS Codec". In: *Proceedings of the 135th AES Convention*. New York, USA, Oct. 2013

<sup>27</sup> Claude Lamblin. "Algebraic code-excited linear prediction speech coding method". Pat. US 5717825 A (US). France Telecom. Feb. 10, 1995

using ACELP are 3GPP's AMR<sup>28</sup> (Adaptive Multi Rate) and ITU G.722.2<sup>29</sup> (AMR - Wideband), both the defacto standard codecs for mobile telephony.

*Video Codecs* The basic principle of video codecs is to exploit the spatial similarity within frames and temporal similarity between frames. Spatial similarity refers to the observation that in a single video frame, a still picture, neighboring pixels are rather similar as long as there is no sudden transition between two areas in the picture, such as a sharp edge of an object. Temporal similarity refers to the observation that in a sequence of video frames, the pictures as such are rather similar as long as there is no sudden change of the picture, such as a sharp scene cut. Bearing this in mind, the most common video codecs, MPEG2<sup>30</sup>, ITU-T H.263<sup>31</sup>, ITU-T H.264<sup>32</sup> and ITU-T H.265<sup>33</sup>, apply both strategies by assigning a sequence of video frames (also called Group-Of-Pictures) to four categories: I-frames for intra-frame coding, and P-, reference B-, and non-reference B-frames for inter-frame coding.

In intra-frame coding, the I-frame picture is first split into so-called macroblocks of pixels, typically 16x16, 8x8 or 4x4 pixels. Then two processes are applied, spatial prediction and transform coding, whereas the order of those two may depend on the actual codec<sup>34</sup>. For H.264, first the spatial prediction is applied, which means that each macroblock is represented as a combination of neighboring macroblocks and a residual, i.e. the difference of the real and the predicted macroblock. Since the overall codec applies lossy coding, i.e. quantization of the residual signal, error propagation problems may occur if the prediction is based on the original input signal and not on the quantized signal. For this reason, the encoder incorporates also a decoder, whose output is then used for the prediction. The next stage in H.264 is then transform coding by means of a separable integer transform. With similar properties as the Discrete Cosine Transform, this transform computes per residual macroblock a block of coefficients that represent spatial frequencies. For typical video signals, this leads to a better energy compaction towards lower spatial frequencies, i.e. less energy in high spatial frequencies. This enables a higher quantization efficiency by treating different spatial frequencies with different quantization steps.

In inter-frame coding, the P-frames are predicted from the previous I-frame, and the reference B-frames are predicted from the previous and next I- or P-frames, and the non-reference B-frames are predicted from the previous and next I-, P- or reference B-frame. The principle is essentially the same for both unidirectional (P-frames) and bidirectional (B-frames) prediction: The codec computes a prediction of each macroblock from other frames, thus it essentially applies a temporal prediction and not a spatial prediction as in intra-frame coding. Then the prediction residual is encoded using transform coding. This temporal prediction is realized in two steps. First, the codec searches for the most similar macroblocks between the current frame (P-, B-

<sup>28</sup> 3GPP. *TS 26.071 version 12.0.0 Release 12 - Mandatory speech CODEC speech processing functions; AMR speech CODEC, General description*. Standard. Sophia Antipolis Cedex, France: 3rd Generation Partnership Project, 2014

<sup>29</sup> ITU-T. *Recommendation G.722.2 - Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2003

<sup>30</sup> ISO/IEC. *ISO/IEC 13818-7:1995 - Information technology - Generic coding of moving pictures and associated audio information - Part 2: Video*. International Standard. International Organization for Standardization & International Electrotechnical Commission, 2013

<sup>31</sup> ITU-T. *Recommendation H.263 - Video coding for low bit rate communication*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2005

<sup>32</sup> ITU-T. *Recommendation H.264 - Advanced video coding for generic audiovisual services*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2014

<sup>33</sup> ITU-T. *Recommendation H.265 - High efficiency video coding*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2015

<sup>34</sup> Thomas Wiegand et al. "Overview of the H.264/AVC Video Coding Standard". In: *IEEE Transactions On Circuits And Systems For Video Technology* 7 (July 2003), pp. 560-576

frame) and the one serving as reference (I-, P- or reference B-frame). Second, the codec computes so-called motion vectors, which describe the displacement between the identified pairs of macroblocks. At the decoder side, the frames are then restored by reconstructing the prediction and correcting that prediction by decoding the residual macroblocks.

In ITU-T H.264, this principle of different frame types and macroblocks is further extended by introducing slices: instead of applying the intra- and intercoding strategies on whole frames, i.e. whole video pictures, H.264 allows to apply these two strategies on parts of the picture, the slices.

### 3.5.2 Packetization and Transmission

In recent years, the traditional Public Switched Telephony Network (PSTN) is more and more replaced by packet-based transmission techniques, which are in turn predominantly built upon the Internet Protocol (IP). For that reason, the present paragraph limits to a brief description of the IP-based mechanisms that are used for audio-only and audiovisual communication.

The basic approach is to use the mechanisms of the Internet Model<sup>35</sup>, also referred to the TCP/IP model, which is similar to the Open Systems Interconnection (OSI) reference model<sup>36</sup>. The Internet model defines a set of four different layers; the OSI model defines seven layers. Each layer in either model takes care of one specific aspect of the data transmission by building on the services provided by the next lower layer and by hiding the details of the current functionality from the next higher layer. This encapsulation of hierarchical layers allows an essentially unlimited flexibility with respect to the actual transmission technologies, given that the communication between layers is well defined by protocols.

According to this model, the transmission of data from sender to receiver is a process of going down the layers at the send side, using the lowest layer to access the actual transmission medium, and going up the layers at receive side. For IP-based audiovisual transmission using the Internet model, this process is described in Figure 3.3.

<sup>35</sup> R. Braden. *Requirements for Internet Hosts – Communication Layers*. International Standard. Internet Engineering Task Force (IETF), 1989. URL: <https://tools.ietf.org/rfc/rfc1122.txt>

<sup>36</sup> ISO/IEC. *ISO/IEC 7498-1:1994 - Information technology – Open Systems Interconnection – Basic Reference Model: The Basic Model*. International Standard. International Organization for Standardization & International Electrotechnical Commission, 1994

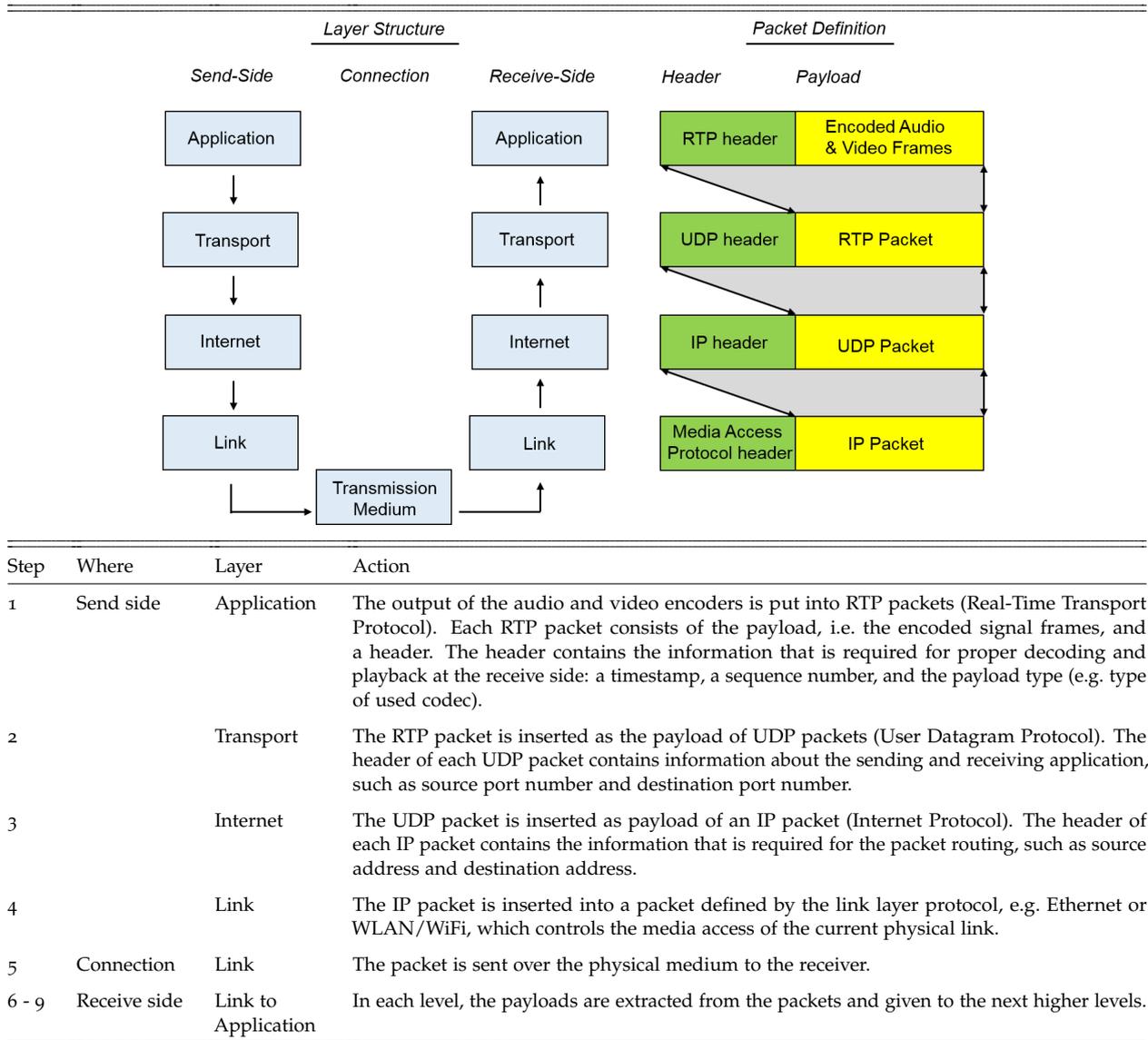


Figure 3.3: Outline of the IP-based packetization and transport process for an audio-only or audiovisual communication link.

### 3.5.3 Buffering and Packet Error Handling

One key characteristic of IP-based transmission over the network is that packets are transported independently from each other. This means that individual packets from the same sender to the same receiver can actually take different routes through the network. Therefore, packets can reach the receiver with different transmission delays, which is referred to as jitter, and they can even arrive out of order. For that reason, a jitter buffer is installed in the receiver, which is holding packets for some constant or adaptive time in order to sort them back in the right order and to play them out in equidistant time intervals. This leads to additional transmission delay. Typical network transmission delays are, according to Thomsen and Jani<sup>37</sup>, in the order of 50 - 200 ms.

There are three different problems possible that eventually lead to the same error: packet loss. First, when network congestion occurs, i.e. too many data packets shall be sent over a network link, the corresponding network router needs to drop packets. Second, if packets arrive the receiver at a later moment in time than the jitter buffer length allows for, they are dropped by the receiver. Third, if packets are corrupted, then they are also dropped, whereas corrected packets are usually detected by verifying whether checksums stored in the media access protocol header (e.g. Ethernet), the IP header and the UDP header correspond to the actual header or payload sizes.

In order to deal with lost packets, there are two different options: sender-based and receiver-based loss recovery. Sender-based techniques such as Forward Error Correction (FEC) or Low-Bitrate Redundancy (LBR) send redundant duplicates of packets (FEC) or redundant versions which were encoded at a lower bitrate (LBR) over the network. Those techniques allow to restore the lost data to a certain amount, depending on the amount of redundant data sent, but they also lead to a further increase of network traffic and delay.

Receiver-based techniques, referred to as Packet Loss Concealment, attempt to restore the missing signal parts at receiver side. Various approaches are possible, ranging from zero or noise insertion, over repetition of the previous frame, to more advanced methods using time stretching or interpolation.

A typical strategy is to use a combination of both options, that is sender-based techniques are applied in any case, and if they fail, the receiver-based techniques will further minimize the effects. For more details the present text refers to Perkins et al.<sup>38</sup>, Korhonen<sup>39</sup> and Raake<sup>40</sup>, focusing on audio transmission, and Garcia<sup>41</sup> addressing also video transmission.

### 3.5.4 Audio-Video Synchronization

Since the audio and video data is encoded and decoded in separate codecs, measures need to be taken to synchronize both signal streams. There are two options for this and both can be found in real-life applications.

<sup>37</sup> G. Thomsen and Y. Jani. "Internet telephony: Going like crazy". In: *IEEE Spectrum* 37.5 (2000), pp. 52-58

<sup>38</sup> C. Perkins, O. Hodson, and V. Hardman. "A survey of packet loss recovery techniques for streaming audio". In: *IEEE Network Magazine* 11.5 (1998), pp. 40-48

<sup>39</sup> Jari Korhonen. "New methods for robust audio streaming in a wireless environment". PhD Thesis. Tampere University of Technology, Finland, 2006

<sup>40</sup> Alexander Raake. *Speech Quality of VoIP - Assessment and Prediction*. Chichester, West Sussex, UK: Wiley, 2006

<sup>41</sup> Marie-Neige Garcia. *Parametric Packet-based Audiovisual Quality model for IPTV services*. Springer, 2014

The first option is to insert a multiplexer between the encoder outputs and the packetizer, which puts encoded audio and video frames that belong to the same period of time into one RTP packet. At receiver side, a de-multiplexer separates again the audio and video data and sends it to the corresponding decoder. The advantage is that both streams are automatically synchronized, as both arrive with the same data packet; the disadvantage is that if this packet is lost, both audio and video channels are degraded.

The second option is to put the audio and video data into separate RTP packets, and then use in the receive-side jitter buffer the timestamp and sequence number information of the RTP header to synchronize the arriving audio and video packets. The advantage is that if one packet is lost, only one of the two modalities, audio or video, is affected; the disadvantage is that more processing is required in the receiver.

### 3.6 Bridges

The purpose of a conference bridge is to establish the connections between the multiple interlocutors, which means, to mix and transfer the audio and video signals in an appropriate manner. There are various possibilities that allow to optimally implement the bridge for the specific scenario, ranging from central-bridges on dedicated servers to client-side bridges on the end devices. This section gives a brief overview of typical approaches that can be found in existing systems.

*Audio Signal Mixing on Central Bridges* The most straight-forward approach is to decode all transmitted audio signals and to generate for every interlocutor an individual mix excluding the own input signal, which is necessary to avoid audible echo. A convenient way to implement this is a matrix computation as shown in Table 3.3.

	$IN_1$	$IN_2$	$\dots$	$IN_N$
$OUT_1$	0	1	$\dots$	1
$OUT_2$	1	0	$\dots$	1
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$OUT_N$	1	1	$\dots$	0

Table 3.3: Mixing matrix for audio conferencing bridge: basic approach.

This implementation, however, does not scale well with an increasing number of interlocutors  $N$ . The number of required operations, i.e. the number of matrix elements, is  $N^2$ , or  $N^2 - N$  when excluding the zero-elements on the diagonal.

For that reason, the audio mixing is usually implemented differently; two typical improvements are presented here. The first improvement, see for example Singh and Schulzrinne<sup>42</sup>, is to first generate a submix of all input streams, and then to subtract for each

<sup>42</sup> Kundan Singh, Gautam Nair, and Henning Schulzrinne. "Centralized conferencing using SIP". in: *Internet Telephony Workshop*. Vol. 7. 2001, pp. 57–63

output from this mix the corresponding input. This leads to a matrix as shown in Table 3.4.

	$IN_1$	$IN_2$	$\dots$	$IN_N$	$IN_1 + \dots + IN_N$
$OUT_1$	-1	0	$\dots$	0	1
$OUT_2$	0	-1	$\dots$	0	1
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	1
$OUT_N$	0	0	$\dots$	-1	1

Table 3.4: Mixing matrix for audio conferencing bridge: advanced approach using a submix and subtraction.

With the expense of computing a submix containing all inputs ( $N$  operations) and inserting this as an additional input to the mixer, and with the number of non-zero matrix entries being now  $2 \cdot N$ , the overall number of operations is now  $3 \cdot N$  instead of  $N^2 - N$ , which leads to reduced computation already for  $N = 5$ .

Another optimization, see for example Firestone et al.<sup>43</sup>, is to use voice activity detection (VAD) on each input channel, and allow only the active speech signals in the mix. This can be further restricted to allow only the  $x$  loudest speech signals. This leads to a matrix shown in Table 3.5, assuming that the first  $x$  inputs are the ones chosen by the VAD mechanism.

	$IN_1$	$IN_2$	$\dots$	$IN_x$	$IN_{x+1}$	$\dots$	$IN_N$
$OUT_1$	0	1	$\dots$	1	0	$\dots$	0
$OUT_2$	1	0	$\dots$	1	0	$\dots$	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\dots$	0
$OUT_x$	1	1	$\dots$	0	0	$\dots$	0
$OUT_{x+1}$	1	1	$\dots$	1	0	$\dots$	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	0	$\dots$	0
$OUT_N$	1	1	$\dots$	1	0	$\dots$	0

<sup>43</sup> Scott Firestone, Thiya Ramalingam, and Steve Fry. *Voice and Video Conferencing Fundamentals*. Cisco Press, 2007

Table 3.5: Mixing matrix for audio conferencing bridge: advanced approach using voice activity detection.

With the expense of having VAD in the signal paths, now the number of operations is down to  $x^2 - x + (N - x) \cdot x$ , or even  $x^2 - x + 1 \cdot x$  if one considers that the channels  $OUT_{x+1}$  to  $OUT_N$  have exactly the same signal. Furthermore, if VAD is implemented in the end devices, the extra computation for VAD on the bridge input can be omitted for those end devices, which further reduces the computational requirements for the bridge. Another advantage of this approach is to improve the signal-to-noise ratio as this avoids an unnecessary accumulation of any background noise from all non-relevant channels.

Finally, both improvements can be combined, leading to a matrix shown in Table 3.6, with  $3 \cdot x + 1$  operations for the matrix and the submix, including the optimization that channels  $OUT_{x+1}$  to  $OUT_N$  contain the same signal.

	$IN_1$	$IN_2$	$\dots$	$IN_x$	$IN_{x+1}$	$\dots$	$IN_N$	$IN_1 + \dots + IN_x$
$OUT_1$	-1	0	$\dots$	0	0	$\dots$	0	1
$OUT_2$	0	-1	$\dots$	0	0	$\dots$	0	1
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\dots$	0	1
$OUT_x$	0	0	$\dots$	-1	0	$\dots$	0	1
$OUT_{x+1}$	0	0	$\dots$	0	0	$\dots$	0	1
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	0	$\dots$	0	1
$OUT_N$	0	0	$\dots$	0	0	$\dots$	0	1

Table 3.6: Mixing matrix for audio conferencing bridge: advanced approach using both submix with subtraction and voice activity detection.

*Audio Signal Mixing for Peer-to-Peer Topologies* In case of a straight-forward peer-to-peer system, each client side receives the incoming streams from the other interlocutors, decodes those streams and generates an own mix of the audio signals. That means, the computational load for a central-bridge is distributed among the clients, i.e. the matrices in Tables 3.3 and 3.5 are reduced to single rows, while the approaches using submix and subtraction (Tables 3.4 and 3.6) are not meaningful.

*Audio Packet-Stream Mixing* With the packet-based transmission considered in this text, the above described audio signal mixing approaches require to decode all incoming audio packets and, in case of central-bridges, also the re-encoding of the mixed signals. There are a number of approaches that aim at avoiding unnecessary decoding and re-encoding, which are subsumed here under the term *audio packet-stream mixing*.

The first approach is applicable for the ITU-T G.711 codec, whose encoded data essentially reflects a compressed version of the signal. The idea is to avoid the decoding, i.e. the decompression, by integrating corresponding level-dependent gain factors into the bridge and do operations in the bitstream domain, see for instance Singh and Schulzrinne<sup>44</sup>.

The next approach is to exploit VAD-based transmission from the end device, i.e. only packets with active streams are transmitted, which is for instance implemented in the AMR<sup>45</sup> and ITU-T G.729 Annex B<sup>46</sup> codecs. Here the number of channels that need to be decoded is automatically limited to the number of actually transmitting devices.

If VAD-based transmission is not supported by the end devices, then all input channels require decoding. However, proposals exist to estimate the signal energy in incoming packets without fully decoding the data<sup>47</sup>, which, however, is only possible if the coding principle allows such “partial decoding”<sup>48</sup>.

Another related approach using VAD is to apply “speaker selection and forward” schemes, see for instance Smith et al.<sup>49</sup>. Such methods use a mixture of directly forwarding individual streams in single-talk states (only one speaker is talking) and generating mixes only in multi-

<sup>44</sup> Kundan Singh, Gautam Nair, and Henning Schulzrinne. “Centralized conferencing using SIP”. in: *Internet Telephony Workshop*. Vol. 7. 2001, pp. 57–63

<sup>45</sup> 3GPP. *TS 26.071 version 12.0.0 Release 12 - Mandatory speech CODEC speech processing functions; AMR speech CODEC, General description*. Standard. Sophia Antipolis Cedex, France: 3rd Generation Partnership Project, 2014

<sup>46</sup> ITU-T. *Recommendation G.729 - Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2012

<sup>47</sup> Dror Nahumi. “Conferencing arrangement for compressed information signals”. Pat. US 5390177 A (US). At&T Corp. Feb. 14, 1993

<sup>48</sup> Paxton J. Smith, Peter Kabal, and Rafi Rabipour. “Speaker Selection for Tandem-Free Operation VoIP Conference Bridges”. In: *Proceedings of IEEE Workshop on Speech Coding*. IEEE. Tsukuba, Japan, Oct. 2002, pp. 120–122

<sup>49</sup> Paxton J. Smith et al. “Tandem-free VoIP conferencing: A bridge to next-generation networks”. In: *IEEE Communications Magazine* 41.5 (2003), pp. 136–145

talk states (multiple speakers are talking). In case the end devices can perform the decoding and mixing of multiple incoming streams, the central-bridge only takes care of selecting the required streams and leaves decoding and mixing to the end device. Obviously, such approaches reflect hybrid versions of central-bridge and peer-to-peer technology.

*Video Mixing* The most straight-forward approach is to decode all incoming video streams and to generate a new video stream in which the pictures from all interlocutors are visible, also called Continuous Presence Conferences<sup>50</sup>. There are various layouts possible to arrange the individual video pictures, such as all are shown in equally sized frames, or the active speaker is shown in a larger frame, or the own picture is separated and presented smaller, and so forth.

In contrast to the audio domain, in which echo needs to be avoided, it is not necessary per se to remove the own video stream from this new compiled video. This simplifies matters at first glance, since no individual video mixes per interlocutor are necessary. However, actually a mirrored version of the own picture should be presented, in order to show any own movements in those directions that one is used to from a real mirror. This, in turn, requires the generation of individual video mixes per interlocutor. Furthermore, more advanced video conferencing systems allow the users to select an own layout, which again requires individual video mixes.

There are a number of possibilities to reduce the computational load that video mixing requires. A few are presented here, acknowledging that this list is non-exhaustive. The first approach concerns the presentation of the own picture. Instead of having a central-bridge inserting the own picture into the video mix, the (mirrored version of the) own picture may be directly played back on the end device without any encoding or transmission. While this does not reduce the number of required video mixes, it reduces the number of pictures that need to be mixed per interlocutor on the central-bridge.

The next approach is to exchange the Continuous Presence paradigm (all interlocutors are visible) with a Voice-Activated Switched video mode<sup>51</sup>. The idea is to use VAD that determines the current speaker, and the system is only showing the video stream of that active speaker. While this greatly reduces the computational load down to one video stream in total, this has some strong impact on the audiovisual communication experience: it does not show the whole group, so one can not see for instance the non-verbal feedback channels of the other interlocutors (see Section 2.2), and it does not support multi-talk states as only one speaker is visible at a time. For that reason, such a Voice-Activated Switched paradigm can be likely found only in older systems, the services on today's market do provide the Continuous Presence paradigm.

A final approach in this list concerns peer-to-peer topologies and addresses the problem of limited processing power on the end devices, such as mobile phones. Wu et al.<sup>52</sup> propose a peer-to-peer

<sup>50</sup> Scott Firestone, Thiya Ramalingam, and Steve Fry. *Voice and Video Conferencing Fundamentals*. Cisco Press, 2007

<sup>51</sup> Scott Firestone, Thiya Ramalingam, and Steve Fry. *Voice and Video Conferencing Fundamentals*. Cisco Press, 2007

<sup>52</sup> Yu Wu et al. "vSkyConf: Cloud-assisted multi-party mobile video conferencing". In: *Proceedings of the second ACM SIGCOMM workshop on Mobile cloud computing*. ACM, 2013, pp. 33–38

architecture in which the heavy video processing is not done on the end device but on a number of virtual machines running on internet servers (in the “cloud”), which act as surrogates for the end devices. Obviously, such an approach reflects a hybrid version of central-bridge and peer-to-peer technology.

### 3.7 Spatial audio

Spatial audio reproduction is of particular interest for telemeeting systems, since a number of studies have shown a benefit of spatial sound reproduction<sup>53</sup> in terms of reducing the mental effort for separating all speakers in a telemeeting. For that reason, the key concepts and their application in telemeeting systems are presented in this section.

*Background* The term spatial audio refers to sound reproduction techniques that exploit characteristics of the human auditory system to render a sound field in which sound sources are perceived from certain spatial locations. The idea is to generate sound such that the auditory system is able to extract a number of cues that determine the perceived location.

The most important cues for the location in the horizontal plane are Interaural Time Difference (ITD) and Interaural Level Difference (ILD), that are the time and level differences occurring between the two ear signals, as well as interaural coherence, which describes the similarity between the two ear signals. Concerning sound localization in closed rooms another important cue is the ability of the auditory system to dissolve the direction of the source from the room reflections which are coming from all directions. Here, the precedence effect – or law of the first wavefront – is mainly responsible to determine the direction of the source.

The cues for the location in the median plane are a direction-dependent perceived frequency weighting with specific resonances in the ear canal (Blauert’s directional bands) combined with direction-dependent reflections from ear, head and torso, expressed as Head-Related Transfer Functions (HRTF). In addition, tiny unconscious head movements with the resulting changes of ITD and ILD contribute to the localization in the median plane. The most important cues for distance perception are level, frequency weighting and the ratio of direct to reverberated sound waves. For more details on the vast field of spatial hearing, the author refers to Blauert<sup>54</sup>, Moore<sup>55</sup> and Pulkki and Karjalainen<sup>56</sup>.

There are a number of techniques to realize spatial sound, ranging from traditional stereo reproduction over commercial multichannel reproduction (e.g. “5.1 surround sound”) to very advanced approaches using large numbers of loudspeakers (e.g. Wave-Field Synthesis, Higher-Order Ambisonics). For headphones reproduction, another technique is available, commonly known as Binaural Sound Reproduction. Since this technique is applied in the present research,

<sup>53</sup> Jessica J. Baldis. “Effects of Spatial Audio on Memory, Comprehension, and Preference during Desktop Conferences”. In: *Proceedings of the ACM CHI 2001 Human Factors in Computing Systems Conference*. Ed. by Michel Beaudouin-Lafon and Robert J. K. Jacob. Vol. 3. 1. 2001, pp. 166–173

Ryan Kilgore, Mark Chignell, and Paul Smith. “Spatialized audioconferencing: what are the benefits?” In: *Proceedings of the 2003 conference of the Centre for Advanced Studies on Collaborative research CASCON ’03*. IBM Press, 2003, pp. 135–144

Alexander Raake et al. “Listening and conversational quality of spatial audio conferencing”. In: *Proceedings of the AES 40th International Conference*. Tokyo, Oct. 2010

<sup>54</sup> Jens Blauert. *Spatial Hearing – The Psychophysics of Human Sound Localization – Revised edition*. Harvard, MA, USA: The MIT Press, 1997

Jens Blauert, ed. *The Technology of Binaural Listening*. Berlin, Germany: Springer, 2013

<sup>55</sup> Brian C. J. Moore. *Introduction to the Psychology of Hearing*. 6th ed. Leiden, The Netherlands: Brill, 2013

<sup>56</sup> Ville Pulkki and Matti Karjalainen. *Communication Acoustics: An Introduction to Speech, Audio and Psychoacoustics*. Chichester, West Sussex, UK: The MIT Press, 2015

a few more details are provided here.

The principle of binaural sound reproduction is to convolve the monotic and dry source signals, i.e. single-channel signals without any spatial cues, with Head-Related Impulse Responses HRIR (the inverse Fourier transform of above mentioned HRTFs), which then introduce the spatial cues. HRIRs have a very complex structure, which changes for every direction. For that reason, binaural sound reproduction means that for every direction another set of HRIRs (for left and right ear) must be used during rendering, requiring even real-time head tracking and on-the-fly switching of HRIRs if maximum possible naturalness is desired. Furthermore, HRIRs are not synthesized from scratch, but acoustically measured for a certain number of directions, and – if necessary – interpolated during playback for directions which have not been measured.

HRIRs describe in the time domain the direction-dependent modification of sound that reaches the ear in the free-field scenario, i.e. without any room reflections. Usually HRIRs are measured in anechoic chambers producing nearly no reflections from floor, walls and ceiling, since true free-field measurements would contain ambient background noise. Binaural Room Impulse Responses (BRIR) are an alternative in which the impulse response incorporates also the reflections of a room, thus BRIRs are measurements of an acoustic impulse played back in a real and enclosed space. Thus, recordings with HRIRs sound “dry” while recordings with BRIRs contain some reverberation that gives the listener some impression of being in a room.

Concerning the practical realization, a common approach is to measure HRIRs with a dummy head for as many positions of the impulse source as desired. A drawback of this approach is that every human has a different physiology/anatomy of the ear and thus a different set of HRIRs, which means that dummy head recordings are just an approximation. Especially, one common problem with binaural sound reproduction using dummy head recordings is the perception of an unwanted elevation (the sound source is perceived to come from a higher or lower position than intended), which can even lead to the so-called front-back confusions. Alternatively, individual HRIRs can be measured per individual listener by placing small microphones into the ear canal, a procedure that is usually not applied given the required effort.

*Application in telmeetings* The technical principle is rather straightforward: the microphone signal of every interlocutor is rendered with a set of HRIRs belonging to a certain position in the virtual acoustical space and every interlocutor is assigned to another position in that space. The practical realization in real life, however, is not trivial: First, a system needs to decide, where to place all interlocutors in the virtual acoustic space. Some guidance can be for instance found in Hyder et al.<sup>57</sup>, who suggest that sound sources should be presented at the same height, but those are applicable for cases in which the

<sup>57</sup> Mansoor Hyder, Michael Haun, and Christian Hoene. “Placing the participants of a spatial audio conference call”. In: *Proceedings of the 7th IEEE Consumer Communications and Networking Conference (CCNC)*. IEEE, 2010, pp. 1–7

number of interlocutors is fixed and known upfront. Roegiers<sup>58</sup> on the contrary explored, how a system could adapt the positioning of interlocutors, when new interlocutors join the telemeeting, and he proposed four algorithms that were positively perceived by subjects.

Second, little is known on the possible interaction of typical transmission impairments, pre-dominantly packet loss, and spatial audio; two example studies were conducted by Bachl<sup>59</sup> and Spur<sup>60</sup>. The former study suggested that audio frames of both channels should be put into the same packets, as packet loss on one side yields strong negative effects. The latter study suggested that packet loss in a spatial audio conferencing system can be perceived by test subjects similar to packet loss in non-spatial context, i.e. perceived quality degrades with increasing amount of packet loss.

Third, the added value of spatial audio has been mainly shown for cases in which interlocutors used close-talk microphones, hands-free cases were out of scope. However, such cases are highly relevant and technically challenging: On the one hand, the system is faced with additional acoustical distortions (reverb, noise, echo), whereas the effects on spatial audio have – to the author’s knowledge – not been investigated. On the other hand, the system is faced with multiple speakers in the same room, which would require for a spatial separation per speaker an extra sound source separation and channel assignment module<sup>61</sup>.

### Summary

One way to address the various possibilities for implementing telemeeting systems is to characterize the technology in terms of the three main components: *End Device(s)*, *Bridge(s)*, and *Network*. Those components are connected in either a central-bridge topology, a peer-to-peer topology, or hybrid topologies mixing the former two, depending on the physical location of the bridges (central servers in the network or on each client side).

Concerning end devices, various technical properties, such as hardware design (i.e. form factor) and built-in signal enhancement algorithms can influence the quality of the overall system. Regarding the form factor, two quality-relevant examples are the distance between mouth and microphone, which influences the speech signal quality, and the disparity of camera and display, which influences the perception of eye gaze. Regarding the signal enhancement algorithms, the complexity and non-linearity of many of those algorithms makes it often difficult to analytically characterize or even to control for their impact on the signal quality.

Concerning the network, packet-based transmission over the internet is increasingly used. For this purpose, first the audio and video signals are encoded using dedicated speech and video codecs, then the encoded audio and video frames are packetized and sent over the network, using the Internet Protocol stack model. Quality-relevant aspects are the data compression techniques of the different speech and

<sup>58</sup> David Roegiers. “Dynamical Aspects of Spatial-Audio in Multi-participant Teleconferences”. Master Thesis. Universiteit Gent, Belgium, 2013

<sup>59</sup> Maximilian Bachl. “Impact of packet loss on localization and subjective quality in 3D-telephone calls”. Bachelor Thesis. Quality and Usability Lab, Technische Universität Berlin, Germany, 2014

<sup>60</sup> Maxim Spur. “Implementation of a Spatial VoIP Conferencing Demonstrator”. Master Thesis. Quality and Usability Lab, Technische Universität Berlin, Germany, 2015

<sup>61</sup> Martin Rothbucher et al. “3D Audio Conference System with Backward Compatible Conference Server using HRTF Synthesis”. In: *Journal of Multimedia Processing and Technologies* 2 (4 Dec. 2011), pp. 159–175

video codecs, which can lead to different types of signal distortions, the packet-based transmission paradigm, which can lead to different transmission delays for individual packets and even packet loss, and the design of jitter buffers and packet error handling mechanisms, which mitigate such transmission errors.

The key functionality of multiparty telemeetings is provided by the bridges, which provide for each interlocutor appropriate mixes of the audio and video signals. The quality of the audio mixes may be influenced by different techniques which minimize scalability issues in terms of the required number of individual mixes, ranging from subtraction of the own signal to limiting the number of input signals by means of voice activity detection. Furthermore, conventional central-bridges usually require extra decoding and encoding of the signals, which can additionally impact the signal quality. In case of packet-based transmission, however, alternative techniques apply “speaker selection and forward” schemes, which use voice activity detection or partial decoding, to pass only packets with active speech to the clients. Concerning the mixing of video, even with today’s technology the required computational load needs to be properly addressed. Typical solutions range from dedicated powerful central-bridge servers to peer-to-peer solutions that use virtual machines in the network cloud as surrogates.

Next to those components, spatial audio reproduction is of particular interest for modern telemeeting systems. This is motivated by the benefit of spatial sound reproduction in terms of reducing the mental effort for separating all speakers in a telemeeting. The idea is to render the microphone signals of each interlocutor with different sets of Head-Related Impulse Responses such that every interlocutor is assigned to another position in the virtual acoustic space.

# 4

## *Quality Assessment in Telecommunication*

### *What this chapter is about*

The present text concerns the quality assessment of multiparty tele-meeting systems. This chapter provides a concise recapitulation of the most important conceptual aspects of quality and it provides an overview of the most common perceptual and instrumental assessment methods for telecommunication systems. In that respect, the present chapter limits to existing literature on quality assessment that predominantly concerns conventional two-party (video-) telephony systems and related services. More in-depth discussions on concepts and assessment methods for the multiparty scenario are addressed in the following chapters.

### *4.1 Introduction*

There are various aspects that influence the success of a certain telecommunication service or product. Certainly, economic-related aspects, ranging from the manufacturer's brand image to the required costs for the end user, play an important role. However, another aspect that should not be neglected is the quality of the service or product. The argumentation here is twofold: first, if the quality is not sufficient, then the end user will not accept the system; second, if two competing systems share the same economical characteristics, it is most likely that user favors the system with the better quality.

At this point it needs to be emphasized that quality refers to the quality as it is perceived by the end user. Obviously, this is linked to the technical performance of the system, but it is not the same. In this line of thought, there exist a number of definitions of quality in the literature, each formulated from different perspectives and discussed in Section 4.2. Continuing from such definitions, it is further possible to characterize quality more specifically; one approach that can be found in the literature is presented in Section 4.3.

Once the target "quality" is properly defined, researchers and practitioners can decide on an adequate method to assess the quality of the system. Here a comprehensive set of perceptual and instrumental methods are available. The former type of methods requires the conduction of perceptual tests in which human participants rate

the system under test, either by directly communicating over the system or by rating the system's output signals. The latter type of methods requires an algorithm that computes for the system under test estimations of quality, which should represent as best as possible quality judgments that human test participants would have given for that system. Section 4.4 provides an overview of the most common perceptual assessment methods, Section 4.5 of the most common instrumental methods.

## 4.2 Notion of Quality and Different Perspectives

There are various definitions for and interpretations of the concept of quality in the context of Information and Communication Technology. Garcia<sup>1</sup> extensively discusses those different definitions, and Möller and Raake<sup>2</sup> give a nice and compact overview on how these definitions evolved over the last decades, starting from the technical perspective of telephone service providers towards a user-centric perspective. One result of these developments is the definition of the term *Quality of Experience*, provided by Qualinet<sup>3</sup>:

"Quality of Experience (QoE) is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and / or enjoyment of the application or service in the light of the user's personality and current state."

This term serves as the fundamental concept for the present work.

### 4.2.1 User-centric and Technology-Oriented Perspectives on Quality

The definition of QoE emphasizes the user's perspective: it refers to what a user perceives from a system, it does not necessarily refer to the technical performance of a system. For the second perspective, the term *Quality of Service* (QoS) is typically used, which is defined in ITU Recommendation E.800<sup>4</sup> as

"The totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service."

Obviously, there is some natural relation between technical performance of a system and user perception, and between QoS and QoE. However, many research studies on quality perception and quality modeling in the past have shown that this link is not a simple one-to-one relation. Varela et al.<sup>5</sup> discuss in more detail the differences between QoS and QoE, emphasizing that both concepts differ in their conceptual scope (utilitarian vs. utilitarian & hedonic), application scope (telecommunication services vs. all ICT and not-necessarily networked based services), perspective (system vs. user), and assessment methods (technology-oriented vs. multi-disciplinary).

<sup>1</sup> Marie-Neige Garcia. *Parametric Packet-based Audiovisual Quality model for IPTV services*. Springer, 2014

<sup>2</sup> Sebastian Möller and Alexander Raake. "Motivation and Introduction". In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 3–9

<sup>3</sup> European Network on Quality of Experience in Multimedia Systems and Services. *Qualinet White Paper on Definitions of Quality of Experience*. White Paper. Version Version 1.2. Patrick Le Callet, Sebastian Möller and Andrew Perkis (Eds.) Lausanne, Switzerland: (COST Action IC 1003), Mar. 2013

<sup>4</sup> ITU-T. *Recommendation E.800 - Definitions of terms related to quality of service*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2008

<sup>5</sup> Martín Varela, Lea Skorin-Kapov, and Touradj Ebrahimi. "Quality of Service Versus Quality of Experience". In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 85–96

#### 4.2.2 A Hierarchical Perspective on Quality

Another property of the definition of QoE is that it looks at quality from a holistic perspective, which in turn means that QoE can be actually considered as a construct of a number of individual aspects. There are a number of taxonomies<sup>6</sup> that identify those individual aspects and put them into relation for different services. The most relevant individual aspect for the present research is the concept of *Speech Communication Quality* by Möller<sup>7</sup>, which can be further separated into *Voice Transmission Quality*, *Ease of Communication*, and *Conversation Effectiveness*. According to the elaborations of Möller as well as Raake<sup>8</sup>, *Voice Transmission Quality* refers to the quality of the speech signal in terms of perceivable signal distortions, stemming from the technical system (e.g. packet loss) or the environment in which the interlocutor uses the system (e.g. background noise). To transfer this term to audiovisual communication, the present text uses the term *Media Transmission Quality*, whereas synonyms in relevant literature are *Media Quality*<sup>9</sup>, or *Multimedia Quality*<sup>10</sup>. Furthermore, *Media Transmission Quality* refers to the quality that can be perceived in a pure listening-only or viewing-only situation, whereas the other two aspects *Ease of Communication* and *Conversation Effectiveness* can be perceived only in a conversational situation. *Ease of Communication* refers to the system's capability of enabling a conversation, *Conversation Effectiveness* refers to the communication partners. Another term that essentially combines both aspects is *Conversational Quality*, a term that can often be found in relevant literature such as in ITU-T Recommendation P.805<sup>11</sup> or in Guéguin<sup>12</sup>. This set of different aspects of quality and the hierarchical interpretation from a holistic QoE down to Media Transmission Quality is highly relevant for the present research, in particular for Chapters 7 and 8, which strongly rely on a proper separation of the different aspects.

#### 4.2.3 A Process-Oriented Perspective on Quality

Coming back to the definition of QoE, the second sentence in that definition implies that quality is a result the user's perception of the system with respect to his or her expectations. This perspective has been formulated by Jekosch<sup>13</sup>, who defined quality as the

“result of [the] judgment of the perceived composition of an entity with respect to its desired composition.”

This implies that quality is the outcome of a perception and a comparison & judgment process. Jekosch developed a corresponding conceptual model, which has been further refined by Raake<sup>14</sup>, and which will be extensively discussed and adapted to the multiparty telemeeting scenario in Chapter 5.

<sup>6</sup> Sebastian Möller. *Assessment and Prediction of Speech Quality in Telecommunications*. Kluwer Academic publishers, 2000

Sebastian Möller. *Quality of Telephone-Based Spoken Dialogue Systems*. New York, USA: Springer, 2005

Andreas Silzle. “Quality Taxonomies for Auditory Virtual Environments”. In: *Proceedings of the 122nd AES Convention*. May 2007

Sebastian Möller et al. “A Taxonomy of Quality of Service and Quality of Experience of Multimodal Human-Machine-Interaction”. In: *Proceedings of the International Workshop on Quality of Multimedia Experience QoMEX*. 2009

<sup>7</sup> Sebastian Möller. *Assessment and Prediction of Speech Quality in Telecommunications*. Kluwer Academic publishers, 2000

<sup>8</sup> Alexander Raake. *Speech Quality of VoIP – Assessment and Prediction*. Chichester, West Sussex, UK: Wiley, 2006

<sup>9</sup> Martín Varela, Lea Skorin-Kapov, and Touradj Ebrahimi. “Quality of Service Versus Quality of Experience”. In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 85–96

Benjamin Weiss et al. “Temporal Development of Quality of Experience”. In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 133–147

<sup>10</sup> Markus Vaalgamaa and Benjamin Belmudez. “Audiovisual Communication”. In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 195–212

<sup>11</sup> ITU-T. *Recommendation P.805 - Subjective evaluation of conversational quality*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2007

<sup>12</sup> Marie Guéguin et al. “On the evaluation of the conversational speech quality in telecommunications”. In: *Eurasip Journal on Applied Signal Processing* (2008), pp. 185–248

<sup>13</sup> Ute Jekosch. *Voice and Speech Quality Perception – Assessment and Evaluation*. Berlin, Germany: Springer, 2005

<sup>14</sup> Alexander Raake. *Speech Quality of VoIP – Assessment and Prediction*. Chichester, West Sussex, UK: Wiley, 2006

Alexander Raake and Sebastian Egger. “Quality and Quality of Experience”. In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 11–34

### 4.3 Characterizing Quality: Quality Elements and Quality Features

The different approaches to define quality are of a rather abstract and conceptual nature. To transfer those conceptual considerations into a more concrete characterization of quality, the notion of *Quality Elements* and *Quality Features* is in the author's view a useful tool for that purpose.

#### 4.3.1 Quality Features

According to Jekosch<sup>15</sup>, *Quality Features* constitute the *Perceived* and *Desired Compositions* of the entity that she is referring to in her definition of quality. In that sense, Jekosch defined a *Quality Feature* as a "recognizable and nameable characteristic of an entity that is relevant to the entity's quality." Bearing in mind Jekosch's basic quality formation process and applying this to the telemeeting context, *Quality Features* are those characteristics of a telemeeting that a user is considering when forming a quality judgment about the telemeeting. In that respect, *Quality Features* reflect the user-centric perspective of quality, and they provide a means to characterize individual aspects of quality, or – as Jekosch formulated it – "a quality feature is the analyzed result of the perceived, designed entity and is therefore the basis of any description of its quality."

The list of possible *Quality Features* is vast. Möller et al.<sup>16</sup> propose to organize the features according to five levels and provide corresponding example features, shown in Table 4.1. In addition, Garcia<sup>17</sup> compiled a comprehensive list of *Quality Features* that are relevant for IP-based television services but which are also applicable for audiovisual communication; Belmudez<sup>18</sup> described the most relevant *Quality Features* for video telephony; Raake<sup>19</sup> discussed in detail the different *Quality Features* for Voice over IP telephony; and Wältermann<sup>20</sup> identified from a multidimensional perspective of quality further *Quality Features* – or in this case *Quality Dimensions*, see below – of transmitted speech. All those features are added in Table 4.1 as well, to give an overview about the variety of individual perceived aspects that can constitute a quality judgment.

*A note on Quality Dimensions* Next to *Quality Features*, the concept of *Quality Dimensions* is an additional method to describe the individual aspects of quality from a perceptual perspective. The idea is to describe quality as a multidimensional construct, consisting of a number of perceptual dimensions.

In the author's view, one can argue about the boundary between *Quality Features* and *Quality Dimensions*, as both refer to individual aspects of quality from a perceptual perspective. For that reason, Table 4.1 is further extended with additional examples, which were originally labeled as *Quality Dimensions*. Those features (dimensions) stem from a literature overview by Wältermann<sup>21</sup>, who updated an

<sup>15</sup> Ute Jekosch. "Sprache hören und beurteilen: Ein Ansatz zur Grundlegung der Sprachqualitätsbeurteilung". Habilitation thesis. Universität/Gesamthochschule Essen, Germany, 2003

Ute Jekosch. *Voice and Speech Quality Perception – Assessment and Evaluation*. Berlin, Germany: Springer, 2005

<sup>16</sup> Sebastian Möller, Marcel Wältermann, and Marie-Neige Garcia. "Features of Quality of Experience". In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 73–84

<sup>17</sup> Marie-Neige Garcia. *Parametric Packet-based Audiovisual Quality model for IPTV services*. Springer, 2014

<sup>18</sup> Benjamin Belmudez. *Assessment and Prediction of Audiovisual Quality for Videotelephony*. Springer, 2009

<sup>19</sup> Alexander Raake. *Speech Quality of VoIP – Assessment and Prediction*. Chichester, West Sussex, UK: Wiley, 2006

<sup>20</sup> Marcel Wältermann. *Dimension-based Quality Modelling of Transmitted Speech*. Springer, 2013

<sup>21</sup> Marcel Wältermann. *Dimension-based Quality Modelling of Transmitted Speech*. Springer, 2013

Feature category	Modality	Example features
level of direct perception	audio & speech	localization <sup>1</sup> (blurred spatial position, i.e. loss of stereo image) <sup>2</sup> , timbre <sup>1,4</sup> (timbre distortion, i.e. birdies) <sup>2</sup> , coloration <sup>1,4</sup> (muffled audio, band-limited artifacts) <sup>2,5</sup> , loudness <sup>1,4,5</sup> , continuity <sup>1</sup> , smoothness <sup>5</sup> , interruptions <sup>2,5</sup> , quantization artifacts <sup>2</sup> , pre-echo <sup>2</sup> , aliasing artifacts <sup>2</sup> , binaural-unmasking distortion <sup>2</sup> , tone-trembling/sparking <sup>2</sup> , tone-shift <sup>2</sup> , noise overflow <sup>2</sup> , tone spike <sup>2</sup> , sawtooth <sup>2</sup> , beat-artifact <sup>2</sup> , tone leakage <sup>2</sup> , clicking <sup>2</sup> , frame repetition <sup>2</sup> , spatial-fidelity- and timbral-fidelity- related-attributes <sup>2</sup> , background noise and noise reduction artifacts <sup>4</sup> , clipping <sup>4</sup> , clearness (or intelligibility) <sup>4</sup> , color of sound (attributes related to brightness, sharpness and fullness) <sup>4,5</sup> , naturalness <sup>4</sup> , noisiness <sup>1,5</sup> (uncorrelated and correlated with speech, background noise, residual noise, i.e. musical tones) <sup>4</sup> , amount of low frequency-components <sup>4,5</sup> , amount of high-frequency components <sup>5</sup> , listener echo <sup>4</sup> , residual echo <sup>4</sup> , non-linear distortion <sup>4</sup> , clipping/switching <sup>4</sup> , bubbling <sup>5</sup> , diffusivity <sup>5</sup> , signal versus background distortion <sup>5</sup>
	video	sharpness <sup>1,2</sup> , darkness <sup>1,2</sup> , brightness <sup>1,2</sup> , contrast <sup>1,2</sup> , flicker <sup>1,2</sup> , distortion <sup>1</sup> (smear/geometrical distortion for CRT displays) <sup>2</sup> , color perception <sup>1</sup> (color bleeding <sup>2,3</sup> , color space conversion artifacts <sup>2,3</sup> , chroma subsampling artefacts <sup>3</sup> , lack of deep black <sup>2</sup> , color naturalness <sup>2</sup> ), smearing <sup>2,3</sup> , jerkiness <sup>2,3</sup> , motion blur <sup>2</sup> , noisiness <sup>2,3</sup> , camera shake <sup>3</sup> , blockiness <sup>2,3</sup> , blurriness <sup>2,3</sup> , quantization noise <sup>2</sup> , mosquito effect <sup>2,3</sup> , staircase effect <sup>2,3</sup> , ringing <sup>3</sup> , mosaic pattern effect <sup>2,3</sup> , DCT bases-image effect <sup>2</sup> , slicing <sup>2</sup> , freezing <sup>2</sup> , ghosting <sup>2,3</sup> , combing effect (de-interlacing artifacts) <sup>2</sup> , trade-off image size/compression artifacts <sup>2</sup> , motion blur <sup>2</sup> , no depth feel (for LCD displays) <sup>2</sup>
	audiovisual	balance and synchronism <sup>1</sup> (audio-video asynchrony, lip-synchronization) <sup>2</sup>
level of action	audio & speech	perception of sidetone <sup>1,4</sup> , echo <sup>1</sup> , double-talk degradations <sup>1</sup> , talker echo <sup>4</sup>
	video	involvement and immersion <sup>1</sup> , perception of space <sup>1</sup> , perception of own motions <sup>1</sup>
level of interaction	all	responsiveness <sup>1</sup> , naturalness of interaction <sup>1</sup> , communication efficiency <sup>1</sup> , conversation effectiveness <sup>1</sup> , communicability <sup>4</sup> , perception of lack of attention <sup>4</sup> , decreased interruptability <sup>4</sup>
level of the usage instance	all	learnability <sup>1</sup> , intuitivity <sup>1</sup> , effectiveness and efficiency for reaching a particular goal <sup>1</sup> , amount of misunderstanding <sup>4</sup> , ease of use <sup>1</sup> , non-functional features such as the personality of the interaction partner (human or machine) <sup>1</sup> , aesthetics <sup>1</sup>

Notes: Quality Features mentioned in:

<sup>1</sup>: Sebastian Möller, Marcel Wältermann, and Marie-Neige Garcia. "Features of Quality of Experience". In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 73–84

<sup>2</sup>: Marie-Neige Garcia. *Parametric Packet-based Audiovisual Quality model for IPTV services*. Springer, 2014

<sup>3</sup>: Benjamin Belmudez. *Assessment and Prediction of Audiovisual Quality for Videotelephony*. Springer, 2009

<sup>4</sup>: Alexander Raake. *Speech Quality of VoIP – Assessment and Prediction*. Chichester, West Sussex, UK: Wiley, 2006

<sup>5</sup>: Marcel Wältermann. *Dimension-based Quality Modelling of Transmitted Speech*. Springer, 2013

Table 4.1: List of example quality features for the categories according to Möller et al., extended with features mentioned in Garcia, Belmudez, Raake, and Wältermann.

earlier list originally compiled by Raake<sup>22</sup>.

Despite including *Quality Dimensions* into the list of *Quality Features*, the author emphasizes that three aspects can be used to – at least roughly – differentiate these two concepts. First, research on *Quality Dimensions* is concerned with identifying the smallest appropriate number of dimensions. This means, *Quality Dimensions* represent a decomposition of *Quality Features* into a reduced number of dimensions. In other words, *Quality Dimensions* and *Quality Features* – and actually also *Quality Elements*, see below – can be considered to have a hierarchical relation to each other. This approach – from quality elements to quality features to quality dimensions – is for instance very thoroughly presented in Garcia<sup>23</sup>.

Second, *Quality Dimensions* are considered to be orthogonal, whereas *Quality Features* do not need to fulfill this requirement. This perspective is for instance taken in Wältermann<sup>24</sup>.

Third, *Quality Features* are for the user nameable aspects of the perceived quality, which means that users are able to describe them, while *Quality Dimensions* do not necessarily need to be nameable by the users. This refers to the two typical methodologies applied in the relevant research. In one method, the *Quality Dimensions* are extracted from a list of attributes or attribute pairs (semantic differentials) by means of principle component or factor analysis. Here, users usually do not rate the dimensions directly, except if the used attributes already represent the dimensions, as it was done for instance in Wältermann. In the other method, the *Quality Dimensions* are extracted from distance ratings between pairs of stimuli, which require no attributes at all.

#### 4.3.2 *Quality Elements*

According to Jekosch<sup>25</sup>, *Quality Elements* refer to the technical characteristics of an ICT system that contribute to a certain level of quality. In that sense, Jekosch defined a *Quality Element* as a “contribution to the quality” whereas this contribution is “of an immaterial or a material product as the result of an action/activity or a process in one of the planning, execution or usage phases” and/or this contribution is “of an action or of a process as the result of an element in the course of this action or process.” *Quality Elements* reflect the technology-oriented perspective of the service provider or system manufacturer, and they provide a means to characterize individual aspects of quality, or – as Jekosch formulated it – “[a quality element] is the building block for designing an entity.”

Similar to the *Quality Features*, the list of possible *Quality Elements* is vast, especially when considering the technical complexity of ICT systems. One typical approach followed by authors such as Möller, Raake, Wältermann, Garcia, or Belmudez is to identify as *Quality Elements* all system stages, signal operations and environmental factors that can alter the audio and video signals, as well as source signal characteristics that can interact with the technology. In that line of

<sup>22</sup> Alexander Raake. *Speech Quality of VoIP – Assessment and Prediction*. Chichester, West Sussex, UK: Wiley, 2006

<sup>23</sup> Marie-Neige Garcia. *Parametric Packet-based Audiovisual Quality model for IPTV services*. Springer, 2014

<sup>24</sup> Marcel Wältermann. *Dimension-based Quality Modelling of Transmitted Speech*. Springer, 2013

<sup>25</sup> Ute Jekosch. “Sprache hören und beurteilen: Ein Ansatz zur Grundlegung der Sprachqualitätsbeurteilung”. Habilitation thesis. Universität/Gesamthochschule Essen, Germany, 2003

Ute Jekosch. *Voice and Speech Quality Perception – Assessment and Evaluation*. Berlin, Germany: Springer, 2005

thought, the telemeeting components described in Chapter 3 can be seen as *Quality Elements* of a telemeeting system.

Using a different terminology stemming from the Qualinet group<sup>26</sup>, Reiter et al.<sup>27</sup> give an overview of *Influence Factors*, which “have an influence on the Quality of Experience for the user.” More specifically, the Qualinet group and Reiter et al. consider three types of *Influence Factors: Human, System and Context Influence Factors*. In that concept, *Quality Elements* comprise the *System Influence Factors* as well as one part of the *Context Influence Factors* that Reiter et al. refer to as *Physical Context*.

#### 4.4 Perceptual Assessment of Quality

It is known in the field that the perception of quality can be influenced by various factors, which are for instance summarized by Reiter et al.<sup>28</sup>. One consequence for the perceptual assessment of quality by means of experiments is that the experimental setup in terms of instructions, conditions, questionnaires, and general context, can influence results. In order to nevertheless achieve a high reproducibility and comparability of results, the common approach is to conduct standardized quality assessment tests following agreed-upon protocols. The International Telecommunication Union (ITU), and other standardization bodies, such as European Broadcast Union (EBU) or European Telecommunications Standards Institute (ETSI), have developed such test protocols for various ICT applications, systems and services.

*Overview of Test Methods* The existing standardized test methods differ in a number of aspects, each method optimally designed for a specific test purpose. Table 4.2 gives an overview of the main characteristics for the most common test protocols.

In that table, *Test Modality* refers to the modality under test, i.e. audio quality, video quality or audiovisual quality. *Test Paradigm* refers to the interactivity of the test participants during the test, defining two levels: non-interactive tests in which test participants rate test stimuli in a listening-only or viewing-only task, and conversation tests, in which participants perform a conversation task via the system under test. *Presentation Mode* refers to the number of stimuli that is presented per assessment trial and the manner in which the stimuli are presented. In this context, assessment trial is one individual unit within a test, in which a certain stimulus or stimulus combination is presented and the subject is asked to give a rating for this unit. Thus, a whole test session per participants usually consists of a number of such test trials. The most common methods differentiate between three *Presentation Modes*: single stimulus, in which the participant rates directly the stimulus under test; multiple stimuli in temporal sequence, in which the participant gives ratings for (part of) the multiple stimuli or in which the participant judges one stimulus in light of the other stimuli; multiple stimuli simultaneously presented,

<sup>26</sup> European Network on Quality of Experience in Multimedia Systems and Services. *Qualinet White Paper on Definitions of Quality of Experience*. White Paper. Version 1.2. Patrick Le Callet, Sebastian Möller and Andrew Perkis (Eds.) Lausanne, Switzerland: (COST Action IC 1003), Mar. 2013

<sup>27</sup> Ulrich Reiter et al. “Factors Influencing Quality of Experience”. In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 55-72

<sup>28</sup> Ulrich Reiter et al. “Factors Influencing Quality of Experience”. In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 55-72

which differs from above by presenting the multiple stimuli simultaneously and the participant can switch between those during a trial (for audio & video) or can see all spatially separated (for video). *Reference Presentation* refers to the aspect whether the test method requires an explicit reference, and whether participants are aware which stimuli are the reference (“known”) or not (“hidden”). This does not mean that other methods may not include stimuli that the experimenter considers – from his or her expert perspective – as a reference, but those methods do not require the inclusion of them into the test protocol. *Rating Mode* refers to the manner in which participants rate the stimuli: retrospective, i.e. the participants give a rating directly after the stimulus, retrospective (but during stimuli presentation), i.e. the participants are rating the whole stimulus but are allowed to enter the ratings already during the presentation; and instantaneous/continuous, i.e. the participants give ratings all the time, usually realized by a slider. *Rating Scale Type* refers to the aspect whether the test method asks for quality (e.g. from bad to excellent), for quality impairments (e.g. not perceivable to perceivable and very annoying), or for other rating attributes. *Rating Scale Values* refer to the granularity of the ratings, separating discrete categorical scales and continuous scales. *Rating Scale Labels* refers to the aspect whether the used scale employs textual labels at intermediate points on the scale or not. *Rating Scale Numbers* refers to the aspect, whether the used scale employs numbers as numerical labels, alternatively or in addition to any textual labels.

*Further Aspects of Test Methods* Table 4.2 intends to provide a first overview of the most relevant test methods for the present research; it is not exhaustive in two aspects. First, there are more characteristics that differentiate test methods: experimental design, stimuli characteristics, test environment, subject profiles, etc. For those characteristics, the author refers to the individual standard documents. Second, there are many more standardized test methods for additional test scenarios such as noise suppression algorithms<sup>29</sup>, spoken dialogue systems<sup>30</sup>, or HD television<sup>31</sup>. As starting point for further standardized assessment test methods, the author refers to ITU-T Recommendation G.1011<sup>32</sup> as well as to the P-Series<sup>33</sup>, BS-Series<sup>34</sup>, and BT-Series<sup>35</sup> of the International Telecommunication Union.

#### 4.5 Instrumental Assessment and Prediction of Quality

The term *Instrumental Assessment* refers to the algorithmic estimation of quality scores without the need to conduct a perceptual test; common synonyms for *Instrumental Assessment* are *Quality Prediction* or *Objective Models*. As many authors emphasize, such quality prediction models provide only a rough estimate of the quality that real test participants would perceive, and such models should be applied with care. The reason is that quality perception is known to be influenced by the participant’s personality and internal state as well

<sup>29</sup> ITU-T. *Recommendation P.835 - Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2003

<sup>30</sup> ITU-T. *Recommendation P.851 - Subjective quality evaluation of telephone services based on spoken dialogue systems*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2003

<sup>31</sup> ITU-R. *Recommendation BT.710-4 - Subjective assessment methods for image quality in high-definition television*. International Standard. Geneva, Switzerland: International Telecommunication Union, 1998

<sup>32</sup> ITU-T. *Recommendation G.1011 - Reference guide to quality of experience assessment methodologies*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2015

<sup>33</sup> ITU-T. *ITU-T Recommendations – P-Series: Terminals and subjective and objective assessment methods*. retrieved on August 19, 2015. International Telecommunication Union. 2015. URL: <https://www.itu.int/rec/T-REC-P/en>

<sup>34</sup> ITU-T. *ITU-R Recommendations – BS-Series: Broadcasting service (sound)*. retrieved on August 19, 2015. International Telecommunication Union. 2015. URL: <https://www.itu.int/rec/R-REC-BS/en>

<sup>35</sup> ITU-T. *ITU-R Recommendations – BT-Series: Broadcasting service (television)*. retrieved on August 19, 2015. International Telecommunication Union. 2015. URL: <https://www.itu.int/rec/R-REC-BT/en>

Test Characteristic		Common Test Methods								
		P.800	P.805	P.910	P.911	P.920	BS.1435	BS.1116	BT.500	BT.1788
Test Modality	Audio	✓	✓				✓	✓		
	Video			✓					✓	✓
	Audiovisual				✓	✓				
Test Paradigm	Non-interactive (listening-only, viewing-only)	✓ in Annex B		✓	✓		✓	✓	✓	✓
	Conversation	✓ in Annex A	✓			✓				
Presentation Mode	Single stimulus	✓ in Annex A, B	✓	✓ in Clause 6.1, 6.2	✓ in Clause 6.1, 6.4	✓			✓ in Clause 6.1, 6.3	
	Multiple stimuli in temporal sequence	✓ in Annex D		✓ in Clause 6.3, 6.4	✓ in Clause 6.2, 6.3				✓ in Clause 4, (6.2)	
	Multiple stimuli presented simultaneously			✓ in Annex C			✓	✓	✓ in Clause 5, (6.2), 6.4	✓
Reference Presentation	None	✓ in Annex A, B	✓	✓ in Clause 6.1, 6.4	✓ in Clause 6.1, 6.3, 6.4	✓			✓ in Clause 6.1	
	Hidden	✓ in Annex E		✓ in Clause 6.2	✓ in Clause 6.2		✓	✓	✓ in Clause 5	✓
	Known	✓ in Annex D		✓ in Clause 6.3			✓	✓	✓ in Clause 4, 6.4	✓
Rating Mode	Retrospective (after stimuli presentation)	✓	✓	✓	✓ in Clause 6.1, 6.2, 6.3	✓				
	Retrospective (but during presentation)						✓	✓	✓ in Clause 4, 5, 6.1	✓
	Instantaneous / continuous (during presentation)				✓ in Clause 6.4				✓ in Clause 6.1, 6.4	
Rating Scale: Type	Quality	✓ in Annex A, B	✓	✓ in Clause 6.1, 6.2	✓ in Clause 6.1, 6.4	(✓)	✓		✓ in Clause 5, 6.1, 6.3, 6.4	✓
	Impairment	✓ in Annex D, E	✓	✓ in Clause 6.2, 6.3				✓	✓ in Clause 4, 6.1	
	Other		✓						✓ in Clause 6.2	
Rating Scale: Values	Discrete / Categorical	✓		✓	✓ in Clause 6.1, 6.2, 6.3	(✓)		✓	✓ in Clause 4	
	Continuous			✓ in Annex B	✓ in Clause 6.4		✓		✓ in Clause 5, 6.3, 6.4	✓
Rating Scale: Labels	No								(✓ in Clause 6.1)	
	Yes	✓		✓	✓	(✓)	✓	✓	✓ in Clause 4, 5, (6.1)	✓
Rating Scale: Numbers	No								(✓ in Clause 6.1)	
	Yes	✓		✓	✓	(✓)	✓	✓	✓ in Clause 4, 5, (6.1), 6.3, 6.4	✓

Bibliographic details of test methods:

ITU-T. *Recommendation P.800 - Methods for objective and subjective assessment of quality*. International Standard. Geneva, Switzerland: International Telecommunication Union, 1996

ITU-T. *Recommendation P.805 - Subjective evaluation of conversational quality*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2007

ITU-T. *Recommendation P.910 - Subjective video quality assessment methods for multimedia applications*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2008

ITU-T. *Recommendation P.911 - Subjective audiovisual quality assessment methods for multimedia applications*. International Standard. Geneva, Switzerland: International Telecommunication Union, 1998

ITU-T. *Recommendation P.920 - Interactive test methods for audiovisual communications*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2000

ITU-R. *Recommendation BS.1534-1 - Method for the subjective assessment of intermediate quality levels of coding systems*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2003

ITU-R. *Recommendation BS.1116-1 - Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*. International Standard. Geneva, Switzerland: International Telecommunication Union, 1997

ITU-R. *Recommendation BT.500-13 - Methodology for the subjective assessment of the quality of television pictures*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2012

ITU-R. *Recommendation BT.1788 - Methodology for the subjective assessment of video quality in multimedia applications*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2007

Table 4.2: Overview of the main characteristics for the most common perceptual quality assessment methods.

as the context. This is already accounted for in the definition for *Quality of Experience*<sup>36</sup>, and Reiter et al.<sup>37</sup> for instance elaborate on the relevant *Human Influence Factors* and *Context Influence Factors*. As a first consequence, the typical approach is to circumvent individual influences by estimating a quality score averaged over a population of test participants, referred to as Mean Opinion Score (MOS). In other words, common quality prediction models do not consider individual participants or internal aspects such as the personal state. As a second consequence, the typical approach is to address the context dependency by developing quality prediction models for well-defined and rather specific test purposes, and to standardize those models to ensure proper implementation and application across test laboratories. In other words, a quality prediction model is strongly linked to a specific perceptual test protocol or set of test protocols described in the previous section.

*Overview of Quality Prediction Models* In order to select an appropriate quality prediction model, the key characteristics of the model need to match the test goals. For that purpose, Table 4.3 provides an overview of such key characteristics for the most common models standardized by the ITU. This table is inspired by two overview articles by Möller et al.<sup>38</sup> and Raake et al.<sup>39</sup>.

In that table, *Test Protocol* refers to the perceptual test method that served as the basis for the model development. Usually, the standards describing the model explicitly refer to the test method with phrases such as “the model predicts quality in terms of Mean Opinion Scores, see Recommendation XYZ”. However, some standards do not explicitly mention a particular test method; here Table 4.3 shows the most obvious methods and provides a corresponding remark. *Prediction Modality* refers to the modality that is predicted, and refers thus to the test modality of the corresponding perceptual test, i.e. audio quality, video quality or audiovisual quality. *Prediction Paradigm* refers to the type of quality in terms of the interactivity of users, and refers thus to the interactivity of test participants in the corresponding perceptual test. The *Prediction Paradigm* defines two levels: non-interactive, in which test participants rate test stimuli a listening-only or viewing-only task, and conversation, in which participants perform a conversation task via the system under test. *Reference Type* defines whether the model requires a reference signal or not: defining Full-Reference, Reduced-Reference, and No-Reference models. In case of a Full-Reference model, the reference is the original input signal to the system under test. In case of a Reduced-Reference model, features extracted from the original input signal serve as reference. *Input Source* refers to the source and the type of information that the model is using as input. Figure 4.1 visualizes seven different combinations of *Input Source* and *Reference Type* that are applied in existing standardized models. The model input can stem from the system as such, from the packet data stream or the audio/video signals. In case of the system as source, the model input consists of

<sup>36</sup> European Network on Quality of Experience in Multimedia Systems and Services. *Qualinet White Paper on Definitions of Quality of Experience*. White Paper. Version Version 1.2. Patrick Le Callet, Sebastian Möller and Andrew Perkis (Eds.) Lausanne, Switzerland: (COST Action IC 1003), Mar. 2013

<sup>37</sup> Ulrich Reiter et al. “Factors Influencing Quality of Experience”. In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 55–72

<sup>38</sup> Sebastian Möller et al. “Speech Quality Estimation”. In: *IEEE Signal Processing Magazine* 28.6 (Nov. 2011), pp. 18–28

<sup>39</sup> Alexander Raake et al. “IP-Based Mobile and Fixed Network Audiovisual Media Services”. In: *IEEE Signal Processing Magazine* 28.6 (Nov. 2011), pp. 68–79

parameters that are either assumed in case the model is applied in the planning/design phase or measured for an existing system. In case of the signals as source, the model input consists of the raw output and – in case of Full-Reference models – also the input signals. In case of Reduced-Reference video quality models, i.e. ITU-T J.246<sup>40</sup>, not the raw input reference signal but features from that reference are used as input. However, that means that the model requires still the signal, hence it is included in Table 4.3 as a special case of the *Signal* category. Furthermore, there are No-Reference video quality models that use an own internal decoder and not the decoder of the original transmission path, i.e. ITU-T P.1202 (Mode 2)<sup>41</sup>, which means they use video signals as input, which are not necessarily the video signals that the end user is receiving. For that reason, Raake et al.<sup>42</sup> refer to such models as bit-stream models and not as signal-based models, even if the model input is essentially a signal and accordingly included in Table 4.3 as a special case of the *Signal* category. In case of the packet data stream as source, the model input can consist of packet header information or packet payload information (without fully decoding the payload).

<sup>40</sup> ITU-T. *Recommendation J.246 - Perceptual visual quality measurement techniques for multimedia services over digital cable television networks in the presence of a reduced bandwidth reference*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2008

<sup>41</sup> ITU-T. *Recommendation P.1202 - Parametric non-intrusive bitstream assessment of video media streaming quality*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2012

<sup>42</sup> Alexander Raake et al. "IP-Based Mobile and Fixed Network Audiovisual Media Services". In: *IEEE Signal Processing Magazine* 28.6 (Nov. 2011), pp. 68–79

Model Characteristic		Standardized Models																
		BS.1387	G.107	G.107.1	G.1070	J.144	J.246	J.247	J.341	P.562	P.563	P.564	P.862	P.862.2	P.863	P.1201	P.1202	
Test Protocol	BS.1116	✓																
	BT.500					✓												
	P.800		✓	✓	✓ <sup>2</sup>					✓ <sup>3</sup>	✓	✓	✓	✓	✓			
	P.805		✓	✓	✓ <sup>2</sup>					✓ <sup>3</sup>								
	P.910				✓ <sup>2</sup>		✓	✓	✓								✓	✓
	P.911																✓	
	P.920		✓	✓	✓ <sup>2</sup>													
	Prediction	Speech		✓	✓						✓	✓	✓	✓	✓	✓		
Modality	Audio	✓															✓	
	Video					✓	✓	✓	✓								✓	✓
	Audiovisual				✓												✓	
Prediction	Non-interactive	✓	✓ <sup>1</sup>	✓ <sup>1</sup>	✓ <sup>1</sup>	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	
Paradigm	Conversation		✓	✓	✓					✓								
Reference	Full-Reference	✓				✓		✓	✓				✓	✓	✓			
Type	Reduced-Reference																	
	No-Reference		✓	✓	✓					✓	✓	✓					✓	✓
Input Source	System	Parameters	✓	✓	✓													
	Data	Packet Header															✓	
		Packet Payload																✓
	Signal	Raw Signal	✓				✓	✓	✓	✓	✓	✓	✓	✓	✓			
		Features							✓									
Model Internal Signal																	✓	

## Remarks:

✓<sup>1</sup>: Even though G.107, G.107.1 & G.1070 address Conversational Quality, it is possible to conduct Listening-Only Tests for deriving some of the parameters, i.e. equipment impairment factors, by following the dedicated protocols P.833 & P.833.1, which in turn are based on P.800

ITU-T. Recommendation P.833 - Methodology for derivation of equipment impairment factors from subjective listening-only tests. International Standard. Geneva, Switzerland: International Telecommunication Union, 2001

ITU-T. Recommendation P.833.1 - Methodology for the derivation of equipment impairment factors from subjective listening-only tests for wideband speech codecs. International Standard. Geneva, Switzerland: International Telecommunication Union, 2009

✓<sup>2</sup>: G.1070 is not specific on the actual test paradigm. For speech G.1070 is essentially a part of G.107 and is thus based on the same procedures; the video part follows the idea of equipment impairment factors, which may be derived from non-interactive tests such as P.910, and the video part includes the impact of delay, which may be derived from a conversational test such as P.920.

✓<sup>3</sup>: Actually, P.562 mentions P.800 only in terms of the quality scale on which the model produces an output, making P.800 the preferable test method. Since this model is used for conversational quality, P.805 may be used as well.

## Bibliographic details of models:

ITU-R. Recommendation BS.1387-1 - Method for objective measurements of perceived audio quality. International Standard. Geneva, Switzerland: International Telecommunication Union, 2001

ITU-T. Recommendation G.107 - The E-model: a computational model for use in transmission planning. International Standard. Geneva, Switzerland: International Telecommunication Union, 2014

ITU-T. Recommendation G.107.1 - Wideband E-model. International Standard. Geneva, Switzerland: International Telecommunication Union, 2015

ITU-T. Recommendation G.1070 - Opinion model for video-telephony applications. International Standard. Geneva, Switzerland: International Telecommunication Union, 2012

ITU-T. Recommendation J.144 - Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference. International Standard. Geneva, Switzerland: International Telecommunication Union, 2004

ITU-T. Recommendation J.246 - Perceptual visual quality measurement techniques for multimedia services over digital cable television networks in the presence of a reduced bandwidth reference. International Standard. Geneva, Switzerland: International Telecommunication Union, 2008

ITU-T. Recommendation J.247 - Objective perceptual multimedia video quality measurement in the presence of a full reference. International Standard. Geneva, Switzerland: International Telecommunication Union, 2008

ITU-T. Recommendation J.341 - Objective perceptual multimedia video quality measurement of HDTV for digital cable television in the presence of a full reference. International Standard. Geneva, Switzerland: International Telecommunication Union, 2011

ITU-T. Recommendation P.562 - Analysis and interpretation of INMD voice-service measurements. International Standard. Geneva, Switzerland: International Telecommunication Union, 2004

ITU-T. Recommendation P.563 - Single-ended method for objective speech quality assessment in narrow-band telephony applications. International Standard. Geneva, Switzerland: International Telecommunication Union, 2004

ITU-T. Recommendation P.564 - Conformance testing for voice over IP transmission quality assessment models. International Standard. Geneva, Switzerland: International Telecommunication Union, 2007

ITU-T. Recommendation P.862 - Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. International Standard. Geneva, Switzerland: International Telecommunication Union, 2001

ITU-T. Recommendation P.862.2 - Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs. International Standard. Geneva, Switzerland: International Telecommunication Union, 2007

ITU-T. Recommendation P.863 - Perceptual objective listening quality assessment. International Standard. Geneva, Switzerland: International Telecommunication Union, 2014

ITU-T. Recommendation P.1201 - Parametric non-intrusive assessment of audiovisual media streaming quality. International Standard. Geneva, Switzerland: International Telecommunication Union, 2012

ITU-T. Recommendation P.1202 - Parametric non-intrusive bitstream assessment of video media streaming quality. International Standard. Geneva, Switzerland: International Telecommunication Union, 2012

Table 4.3: Overview of the main characteristics for the most common standardized instrumental quality models.

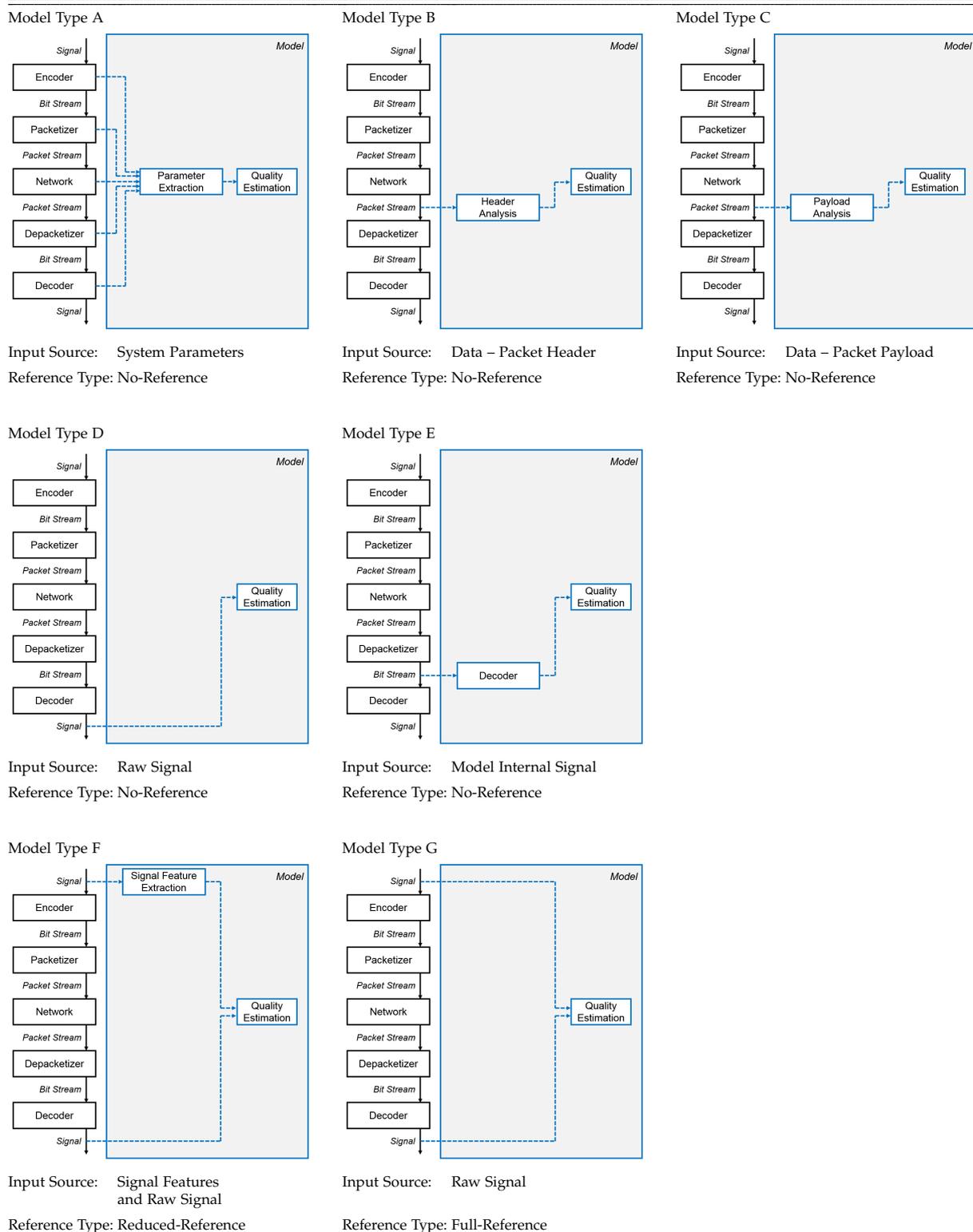


Figure 4.1: Block-diagrams visualizing the different types of information that are used by quality models.

*Further Aspects of Quality Prediction Models* Table 4.3 intends to provide a first overview of the most relevant quality prediction models for the present research; it is not exhaustive in two aspects. First, there are more characteristics that differentiate quality prediction models: *impairment sets, output variables, and ground truth data.*

*Impairment sets* refers to the list of impairments that the model can handle. Here, each model was developed and evaluated only for a specific set of different impairments, and the standard documents usually list those impairment sets, sometimes labeled as test factors, and sometimes augmented with explicit impairment sets for which the models are not working properly.

*Output variables* refers to the type of quality score that the model is actually computing. Most models estimate Mean Opinion Scores in terms of the 5-point Absolute Category Rating Scale defined in ITU-T P.800<sup>43</sup> and P.910<sup>44</sup>. Other models compute quality-related scores which can be transformed into Mean Opinion Score, such as the Transmission Rating Factor computed in ITU-T G.107<sup>45</sup>, G.107.1<sup>46</sup>, & G.1070<sup>47</sup>. Further models compute alternative quality scores which rely on explicit comparisons to a reference, such as the 5-grade Impairment Scale defined in ITU-R BS.1284<sup>48</sup>, which is for instance computed in BS.1387<sup>49</sup>.

*Ground truth data* refers to the question, whether the model is directly estimating perceptual quality ratings, or whether the model is actually estimating quality scores computed by another model, which is deemed as sufficiently correct to serve as a reference model. The motivation for this intermediate step of using predictions from an instrumental reference model as ground truth is to circumvent for the new model the need of extensive perceptual tests, while the new model is supposed to circumvent certain principle limitations of the reference model. The prominent example for such an approach is P.564<sup>50</sup>, which uses P.862<sup>51</sup> predictions as ground-truth, whereas P.564 is a No-Reference model and does not have the limitation of requiring the reference signal.

Second, there are more standardized and non-standardized quality prediction models, such as dimension-based approaches or models for speech and noise quality. For more details on those characteristics and pointers for further models, the author refers to the individual standard documents, to ITU-T Recommendation G.1011<sup>52</sup>, as well as the overview articles of Möller et al.<sup>53</sup> and Raake et al.<sup>54</sup>.

*A note on Quality Prediction Models for Multiparty Telemeetings* To the author's knowledge, no quality prediction model for the specific purpose of multiparty telemeetings has been standardized or published yet. There is one publication by Adel et al.<sup>55</sup>, however, that goes into the direction of a multiparty telemeeting quality model. Adel et al. validated the performance of the E-Model (i.e. ITU-T G.107) against PESQ (i.e. ITU-T P.862) predictions for system configurations that occur in telemeetings, and they proposed an update of the E-Model.

Unfortunately, the study did not address the multiparty aspect

<sup>43</sup> ITU-T. *Recommendation P.800 - Methods for objective and subjective assessment of quality*. International Standard. Geneva, Switzerland: International Telecommunication Union, 1996

<sup>44</sup> ITU-T. *Recommendation P.910 - Subjective video quality assessment methods for multimedia applications*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2008

<sup>45</sup> ITU-T. *Recommendation G.107 - The E-model: a computational model for use in transmission planning*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2014

<sup>46</sup> ITU-T. *Recommendation G.107.1 - Wideband E-model*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2015

<sup>47</sup> ITU-T. *Recommendation G.1070 - Opinion model for video-telephony applications*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2012

<sup>48</sup> ITU-R. *Recommendation BS.1284-1 - General methods for the subjective assessment of sound quality*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2003

<sup>49</sup> ITU-R. *Recommendation BS.1387-1 - Method for objective measurements of perceived audio quality*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2001

<sup>50</sup> ITU-T. *Recommendation P.564 - Conformance testing for voice over IP transmission quality assessment models*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2007

<sup>51</sup> ITU-T. *Recommendation P.862 - Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2001

<sup>52</sup> ITU-T. *Recommendation G.1011 - Reference guide to quality of experience assessment methodologies*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2015

<sup>53</sup> Sebastian Möller et al. "Speech Quality Estimation". In: *IEEE Signal Processing Magazine* 28.6 (Nov. 2011), pp. 18–28

<sup>54</sup> Alexander Raake et al. "IP-Based Mobile and Fixed Network Audiovisual Media Services". In: *IEEE Signal Processing Magazine* 28.6 (Nov. 2011), pp. 68–79

<sup>55</sup> Mohamed Adel et al. "Improved E-Model for Monitoring Quality of Multiparty VoIP communications". In: *Proceedings of the IEEE Globecom Workshops*. Atlanta, USA, 2013, pp. 1180–1185

as only the connection from one speaker to one listener was evaluated. Instead, the study essentially investigated the quality impact of a tandem of two links that occur between two interlocutors in a central-bridge topology: one link from speaker to bridge and one link from bridge to listener. The test conditions comprised different packet loss rates on each link, and this was tested for three different codecs and for two bridging processes *audio signal mixing* and *packet stream mixing* (see Section 3.6), which means conditions with and without additional decoding and encoding inside the bridge. In that respect, the proposed update of the E-Model is not a true *multiparty* telemeeting quality model, but it presents a first step towards such a model as it considers tandem-effects that occur in telemeeting scenarios.

### Summary

The concept of quality can be defined from different perspectives. The complementary approaches are to take a user's perspective and a technical perspective; the first perspective leading to the concepts of Quality of Experience and Quality Features, the second to the concepts of Quality of Service and Quality Elements. Another perspective on quality is to consider quality from a holistic perspective or from a more analytical perspective in which quality is considered as a construct of individual aspects, such as Speech Communication Quality. Furthermore, a process-oriented perspective is to consider quality as a result of the user's perception of the system with respect to his or her expectations.

Since quality perception can be influenced by various factors, a typical approach in literature is to apply standardized assessment methods. Those methods can be categorized into perceptual assessment methods and instrumental assessment methods.

Perceptual assessment methods, that is the assessment of a system's quality by means of human ratings, follow stringent and standardized test protocols. Those protocols can be distinguished by a number of aspects. The most important aspects are *Test Modality* (audio, video, audiovisual), *Test Paradigm* (non-interactive, conversation), *Presentation Mode* (single or multiple stimuli per rating), *Presentation of Reference* (none, hidden or known to participant), *Rating Mode* (retrospective, instantaneous), *Rating Scale Type* (quality, impairment, other), *Rating Scale Values* (discrete, continuous), and usage of *Labels and Numbers on the Rating Scales*.

Instrumental assessment methods, that is the estimation of a system's quality by means of a computational model, are limited to a certain number of test cases. For that reason, the key characteristics of such models need to match the test case at hand. The most important characteristics are *Test Protocol*, which served as the basis for the model development, *Prediction Modality* (audio, video, audiovisual), *Prediction Paradigm* (non-interactive, conversation), *Reference Type* (Full-Reference, Reduced-Reference, No-Reference), and *Input Source* (system parameters, signals, packet data stream).

## **Part III**

# **Contribution**

## 5

# Conceptual Model of Telemeeting Quality

---

This chapter contains text passages that either stem from previous texts of the author or that are based on such source texts. For readability purposes those passages are not marked as quotations. The main source documents used for this chapter are:

- Janto Skowronek, Katrin Schoenenberg, and Gunilla Berndtsson. "Multimedia Conferencing and Telemeetings". In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 217–228
- Janto Skowronek et al. "Method and Apparatus for Computing the Perceived Quality of a Multiparty Audio or Audiovisual Telecommunication Service or System". Patent application WO/2016/041593 (EP). Deutsche Telekom AG. Mar. 2016
- Janto Skowronek and Alexander Raake. "Conceptual model of multiparty conferencing and telemeeting quality". In: *Proceedings of the 7th International Workshop on Quality of Multimedia Experience (QoMEX 2015)*. Pilos, Greece, May 2015, pp. 1–6

Since additional documents served as further background material, see Appendix A for the full list of source documents and their contributions to this chapter.

---

### What this chapter is about

When it comes to the assessment of telemeeting quality, a number of ambiguities inherent in this task can cause misinterpretation of results. The goal of this chapter is to dissolve such ambiguities by providing a conceptual model for telemeeting quality. This conceptual model could serve as a tool to precisely specify all those aspects of telemeeting quality that a researcher or practitioner plans to investigate. For the present text, it will serve as a theoretical basis for the empirical research described in the next chapters.

### 5.1 Introduction

Quality of Experience is apparently a non-trivial issue because the process from perception to quality judgment inside a user appears to be rather complex, which is, in addition, influenced by many factors. In order to structure and explain this complex matter, conceptual models of the quality formation process have been developed in the literature, which can be more specifically described as taxonomies<sup>1</sup>, categorizations<sup>2</sup> and process models<sup>3</sup>. Quality of Experience of a telemeeting – in short telemeeting quality – is even more complex because there are a number of ambiguities when it comes to the proper interpretation of telemeeting quality.

One such ambiguity is that telemeeting quality can refer to the quality perceived by each individual telemeeting participant, or to the quality perceived by the whole group of telemeeting participants.

---

<sup>1</sup> Sebastian Möller. *Assessment and Prediction of Speech Quality in Telecommunications*. Kluwer Academic publishers, 2000

Sebastian Möller. *Quality of Telephone-Based Spoken Dialogue Systems*. New York, USA: Springer, 2005

Alexander Raake. *Speech Quality of VoIP – Assessment and Prediction*. Chichester, West Sussex, UK: Wiley, 2006

Andreas Silzle. "Quality Taxonomies for Auditory Virtual Environments". In: *Proceedings of the 122nd AES Convention*. May 2007

Sebastian Möller et al. "A Taxonomy of Quality of Service and Quality of Experience of Multimodal Human-Machine-Interaction". In: *Proceedings of the International Workshop on Quality of Multimedia Experience QoMEX*. 2009

<sup>2</sup> Sebastian Möller, Marcel Wältermann, and Marie-Neige Garcia. "Features of Quality of Experience". In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 73–84

Ulrich Reiter et al. "Factors Influencing Quality of Experience". In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 55–72

<sup>3</sup> Alexander Raake. *Speech Quality of VoIP – Assessment and Prediction*. Chichester, West Sussex, UK: Wiley, 2006

Alexander Raake and Sebastian Egger. "Quality and Quality of Experience". In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 11–34

Focusing on the perspective of an individual participant, telemeeting quality can refer either to the quality of the whole telemeeting or to the quality of the individual connections between participants. Notice that in the following these related ambiguities are addressed as the *Different Aggregation Levels of Telemeeting Quality*.

Another ambiguity is that telemeeting quality can focus on media transmission quality, e.g. perception of speech or video signal distortions, on group communication aspects, e.g. perceived conversation flow or comfort, or on both. Notice that in the following these different foci are addressed as the *Telecommunication & Group-Communication Components of Telemeeting Quality*.

Apparently, there are different facets of telemeeting quality which require a solid description in order to avoid misinterpretation of research results. For that purpose, the following sections will describe how the aforementioned telemeeting specific concepts *Quality Aggregation Levels* and *Telecommunication & Group-Communication Components* influence the perception of telemeeting quality. To this aim, Section 5.2 will re-synthesize and extend an existing model of the quality formation process inside a person; while Sections 5.3 to 5.4 will each describe one of the telemeeting specific concepts in more detail and will then include them into the process model.

## 5.2 Telemeeting Quality Formation Process

This section describes a conceptual model of the quality formation process inside a telemeeting participant. The model is based on a comprehensive model of the quality formation process proposed by Raake & Egger<sup>4</sup>, which in turn builds on a prior model developed by Raake<sup>5</sup> and the concepts of Jekosch<sup>6</sup>.

While the model of Raake & Egger is a very generic model that covers various aspects of the quality formation process inside a person, the present text will re-synthesize that model for the specific case of telemeetings. To this aim, the model presented here will focus on a limited number of parts of the Raake & Egger model that are essential in the telemeeting context. In addition, the telemeeting-specific model will elaborate on a number of those essential parts in more detail than it has been done for the Raake & Egger model. Given the complexity of the process model, the present text will explain step-by-step the individual model components, starting from the basic principle and ending at a detailed level sufficient to include later the telemeeting specific aspects *Quality Aggregation Levels* and *Telecommunication & Group-Communication Components*.

### 5.2.1 Basic principle

Starting from the definitions of Jekosch, Figure 5.1 visualizes the basic principle of the quality formation process: Telemeeting quality is the result of a judgment process in which the perceived composition of the telemeeting is compared against its desired composition. In this

<sup>4</sup> Alexander Raake and Sebastian Egger. "Quality and Quality of Experience". In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 11–34

<sup>5</sup> Alexander Raake. *Speech Quality of VoIP – Assessment and Prediction*. Chichester, West Sussex, UK: Wiley, 2006

<sup>6</sup> Ute Jekosch. "Sprache hören und beurteilen: Ein Ansatz zur Grundlegung der Sprachqualitätsbeurteilung". Habilitation thesis. Universität/Gesamthochschule Essen, Germany, 2003

Ute Jekosch. *Voice and Speech Quality Perception – Assessment and Evaluation*. Berlin, Germany: Springer, 2005

process, the perceived composition is the totality of quality-relevant features that are perceived from the telemeeting, which requires a perception process in the model.

Furthermore, it is known that various influencing factors have an impact on the quality formation process, e.g. Reiter et al.<sup>7</sup>, as well as the perception process, e.g. Raake & Egger<sup>8</sup>. Such factors can be organized in three categories:

1. Human Influencing Factors, some of them are stemming from inside the person, such as the persons' emotional state or attitude, others are stemming from outside the person, such as the personality or interaction behavior of the other interlocutors<sup>9</sup>.
2. System Influencing Factors, covering essentially all technical aspects of the system that contribute to the quality judgment.
3. Context Influencing Factors, covering all contextual aspects of the situation in which a quality judgment is made, including the environment, use case, tasks, social context, economic context.

Now, considering the specific case of a quality assessment test, this principle can be extended in four aspects according to Figure 5.1: First, a coding process is included in which the person is transforming his or her internal quality judgment into a description that is available to the outside world, i.e. the investigator. Second, such descriptions are usually ratings on a scale provided to the person during the assessment test, but they can in principle also be of a verbal nature. In other words that quality assessment test is directly influencing the coding process. Third, the fact that a quality assessment test is taking place determines the influencing factors, e.g. the person as a test participant is in a certain state, the system properties are defined in form of different conditions, and the contextual situation is determined by test environment, tasks and so forth. Fourth, Jekosch discussed that the person is performing a "controlled quality evaluation", thus the quality assessment test as such can also directly trigger the quality formation process.

<sup>7</sup> Ulrich Reiter et al. "Factors Influencing Quality of Experience". In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 55-72

<sup>8</sup> Alexander Raake and Sebastian Egger. "Quality and Quality of Experience". In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 11-34

<sup>9</sup> Janto Skowronek, Katrin Schoenberg, and Gunilla Berndtsson. "Multi-media Conferencing and Telemeetings". In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 217-228

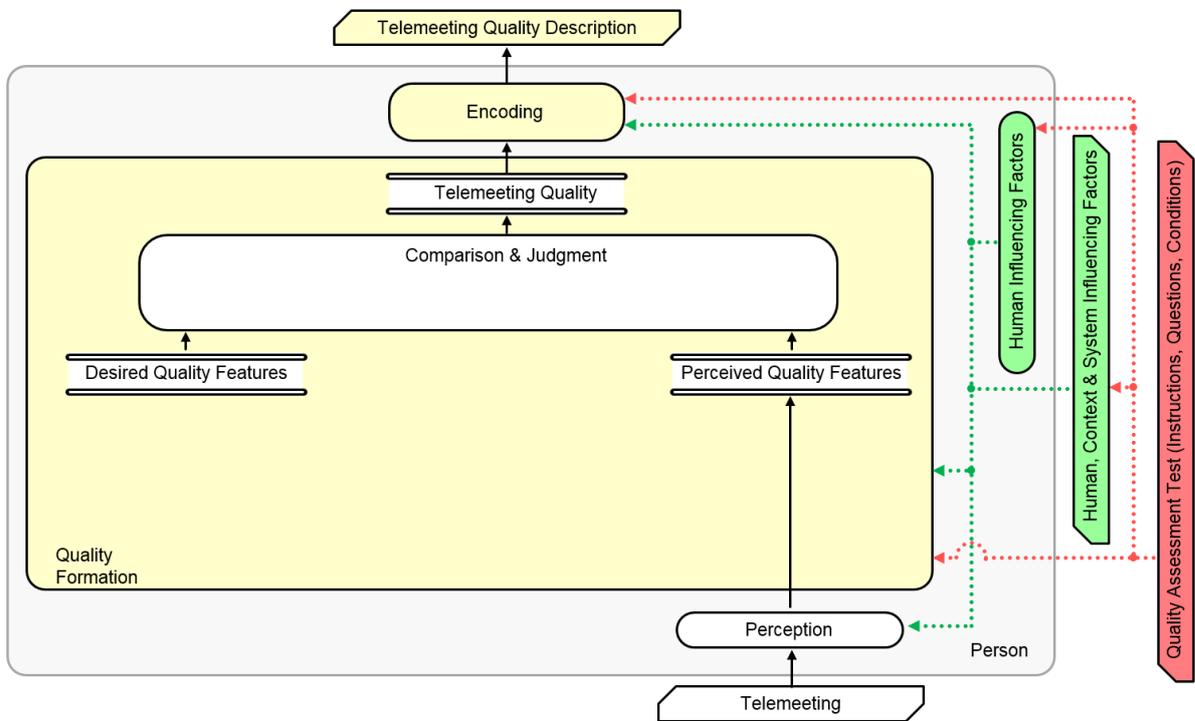


Figure 5.1: Quality formation process inside a telemeeting participant: Basic process model according to the principle concept of Jekosch for the situation of a quality assessment test.

Ute Jekosch. *Voice and Speech Quality Perception – Assessment and Evaluation*. Berlin, Germany: Springer, 2005

5.2.2 Reflection and Attribution

Raake<sup>10</sup> specified the principle of Jekosch in more detail by introducing an additional reflection stage: This stage emphasizes that from the totality of all characteristics of an item, only those characteristics are used in the judgment process that are quality relevant. Furthermore, the more recent version by Raake & Egger<sup>11</sup> calls this stage reflection and attribution in order to account for the possibility that any quality-relevant features are also actually attributed to the item under consideration and not to something else such as the environment or the interlocutors<sup>12</sup>. Figure 5.2 visualizes that additional stage, whereas the two aspects reflection and attribution are drawn as separate processing blocks: The combination of both blocks separate from the perceived character, i.e. totality of all perceived features, only the perceived quality features, i.e. those features that are quality-relevant and attributed to the telemeeting. In other words, these stages can be interpreted as a selection process. Similarly to the perceived character, also the desired character, i.e. the totality of features that characterize the quality reference, is filtered, leading to a set of relevant desired quality features.

<sup>10</sup> Alexander Raake. *Speech Quality of VoIP – Assessment and Prediction*. Chichester, West Sussex, UK: Wiley, 2006

<sup>11</sup> Alexander Raake and Sebastian Egger. “Quality and Quality of Experience”. In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 11–34

<sup>12</sup> This aspect of attribution is particularly important in the context of delay, as Raake for instance discusses that under certain conditions delay may be attributed to the interlocutors and not to the system.

Alexander Raake. *Speech Quality of VoIP – Assessment and Prediction*. Chichester, West Sussex, UK: Wiley, 2006

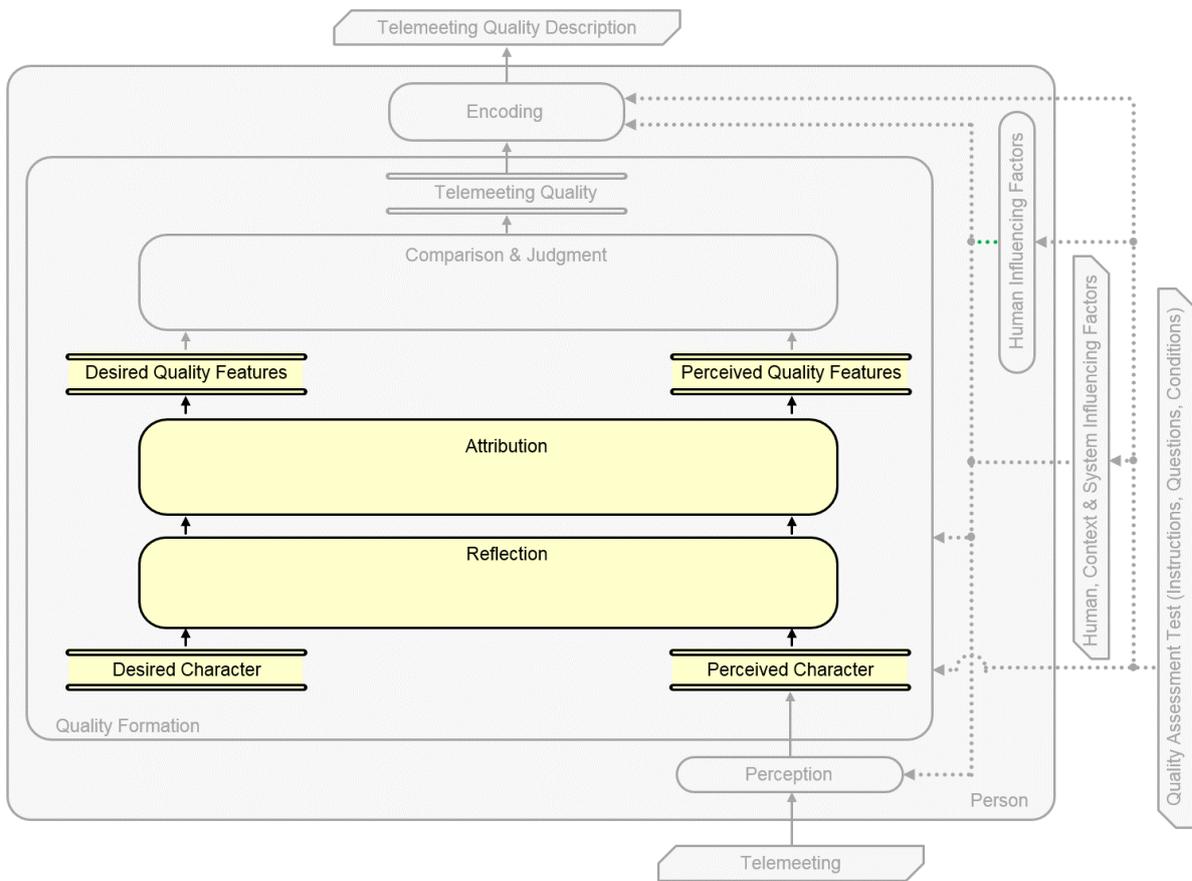


Figure 5.2: Quality formation process inside a telemeeting participant: Extension of the process model with a reflection and attribution stage, according to the models from Raake and Raake & Egger.

### 5.2.3 *A more detailed process of perception*

The model of Raake & Egger<sup>13</sup> provides more details of the perception process, from which some aspects will be relevant later when it comes to the multiparty-specific extensions. For that reason, Figure 5-3 visualizes the relevant details of perception.

First, Raake & Egger distinguish between perception and experiencing by inserting an experiencing stage as a subsequent process to the perception stage. This extra stage reflects the view of Raake & Egger that a quality judgment is made on experiencing the item under considerations, whereas the authors define experience as the one “individual stream of perceptions (of feelings, sensory percepts and concepts) that occurs in a particular situation of reference.” In other words, experiencing refers on the one hand to a subset of all objects in the perceptual world of a person (i.e. individual stream); on the other hand it refers to a certain encounter with the item under consideration (i.e. situation of reference).

Second, Raake & Egger describe that perception is actually an iterative process combining bottom-up and top-down processes: The physical input from the person’s sensory organs is processed and sent into higher cognitive processes to form perceptual events. Then those events are matched against anticipated hypotheses, which are generated in a top-down manner based on internal perceptual references.

Third, Raake & Egger suggest that the perceptual references are also used as input for the reflection and the subsequent attribution stage, which essentially determines the desired character.

<sup>13</sup> Alexander Raake and Sebastian Egger. “Quality and Quality of Experience”. In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 11–34

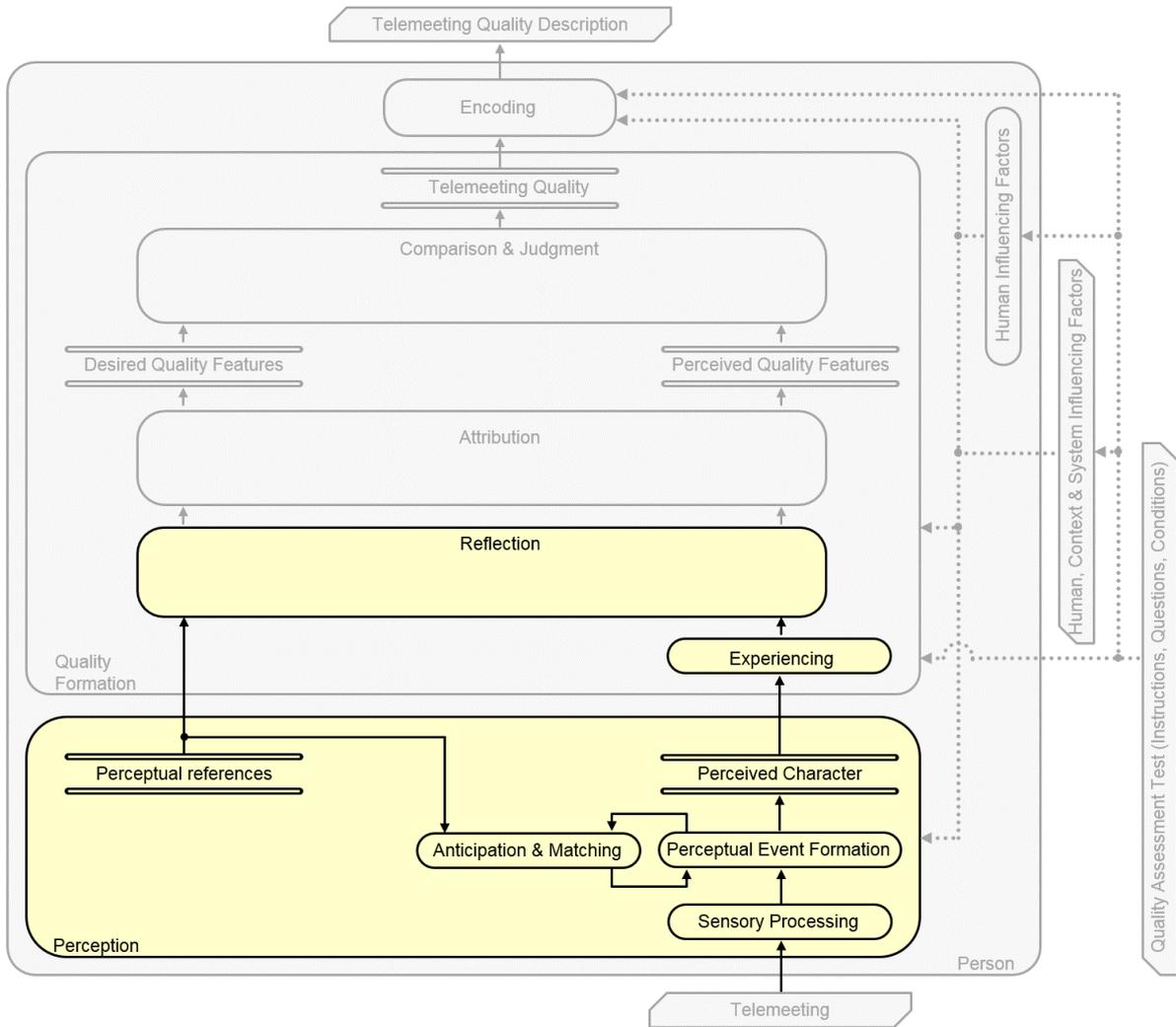


Figure 5.3: Quality formation process inside a telemeeting participant: Extension of the process model with a more detailed description of the perception and experiencing process, according to the model from Raake & Egger.

Alexander Raake and Sebastian Egger. "Quality and Quality of Experience". In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 11-34

#### 5.2.4 Quality Awareness and Quality Attention Focus

Another process block proposed by Raake & Egger<sup>14</sup> is a quality awareness stage, which is introduced as a trigger mechanism to actually start a quality formation process of an experience. Figure 5.4 visualizes the triggering character of this quality awareness stage by means of switches placed before the reflection process stage. This emphasizes that a quality judgment takes only place if certain internal or external triggers are activated. An internal trigger would be that certain assumptions that the person has of the perceived object are not met, for example when the character of speech signals in the telemeting is different than what the person is used to. This actually means that the results of the “Anticipation and Matching” stage are examined to trigger the quality formation process. An external trigger would be an instruction given to the person to make a quality judgment, for example in case of a quality assessment test. Thus, the quality assessment test as such can control the quality awareness stage. Correspondingly, Figure 5.4 visualizes the two possible input paths to the quality awareness stage.

A new refinement to the process model is to insert an additional stage representing the quality attention focus (see Figure 5.4). Quality attention focus is a term introduced here to describe a cognitive process that focuses the attention to a limited number of quality relevant aspects, i.e. the set of quality features, that a user is actually considering when making the quality judgment. Quality attention focus<sup>15</sup> refers to what is focused on during the quality judgment process, while quality awareness refers to the act of triggering the quality judgment process.

While such a stage could be simply seen as part of the reflection and attribution stage, the introduction as an extra stage emphasizes that the interpretation of the reflection and attribution stage as a selection process requires two inputs: on the one hand a mechanism to trigger the selection (“when to select”), as it is represented by the quality awareness stage; on the other hand a mechanism to define the selection (“what to select”), as it is presented by the new stage.

As for the quality awareness stage, the quality attention focus stage uses as input either internal or external information in order to define which set of quality features shall be selected. The external input would be again the instructions given in a quality assessment test, as shown in Figure 5.4. Concerning the internal input, those characteristics of the perceptual object that are different from the assumptions and that are actually triggering the quality awareness would be the primary set of characteristics that shall be reflected upon. However, the quality attention focus may be more complex by also identifying additional characteristics that should be reflected upon. This accounts for the author’s view that single characteristics can be sufficient to trigger a quality judgment, while the quality judgment as such can be nevertheless a multidimensional construct. An example to explain this perspective: The perception of a clearly audible speech

<sup>14</sup> Alexander Raake and Sebastian Egger. “Quality and Quality of Experience”. In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 11–34

<sup>15</sup> Attention is a term widely used to describe the human ability to concentrate on a limited set of information to perform a certain task. More precisely, in psychology, attention can refer to sensory perception or to higher cognitive processes and motor control and can be related to working memory and learning. For a comprehensive overview of these different facets, the interested reader may be referred to the work of Pashler. However, for the present text, the interpretation of the term as a general concept is sufficient.

Harold E. Pashler. *The psychology of attention*. Cambridge, MA, USA: The MIT Press, 1998

distortion during the telemeeting triggers the participant to reflect on quality. Then it might happen of course that only this distortion would be used in the judgment process, independently from whether additional aspects of the telemeeting, such as conversational flow or intelligibility, are affected by that distortion or not ("I heard that clear distortion, so the telemeeting is just bad"). However, it might also happen that the other aspects are taken into account as well ("I heard that distortion, but I still could understand and follow the discussion, so it was not too bad"), even if those aspects alone would not trigger a quality judgment ("Conversational flow and intelligibility are as always, so no need to judge the quality"). In order to allow the process model to use additional perceptual characteristics than those used to trigger quality awareness, Figure 5.4 shows that the input of the quality attention stage is not stemming from the quality awareness stage but directly from the perception process.

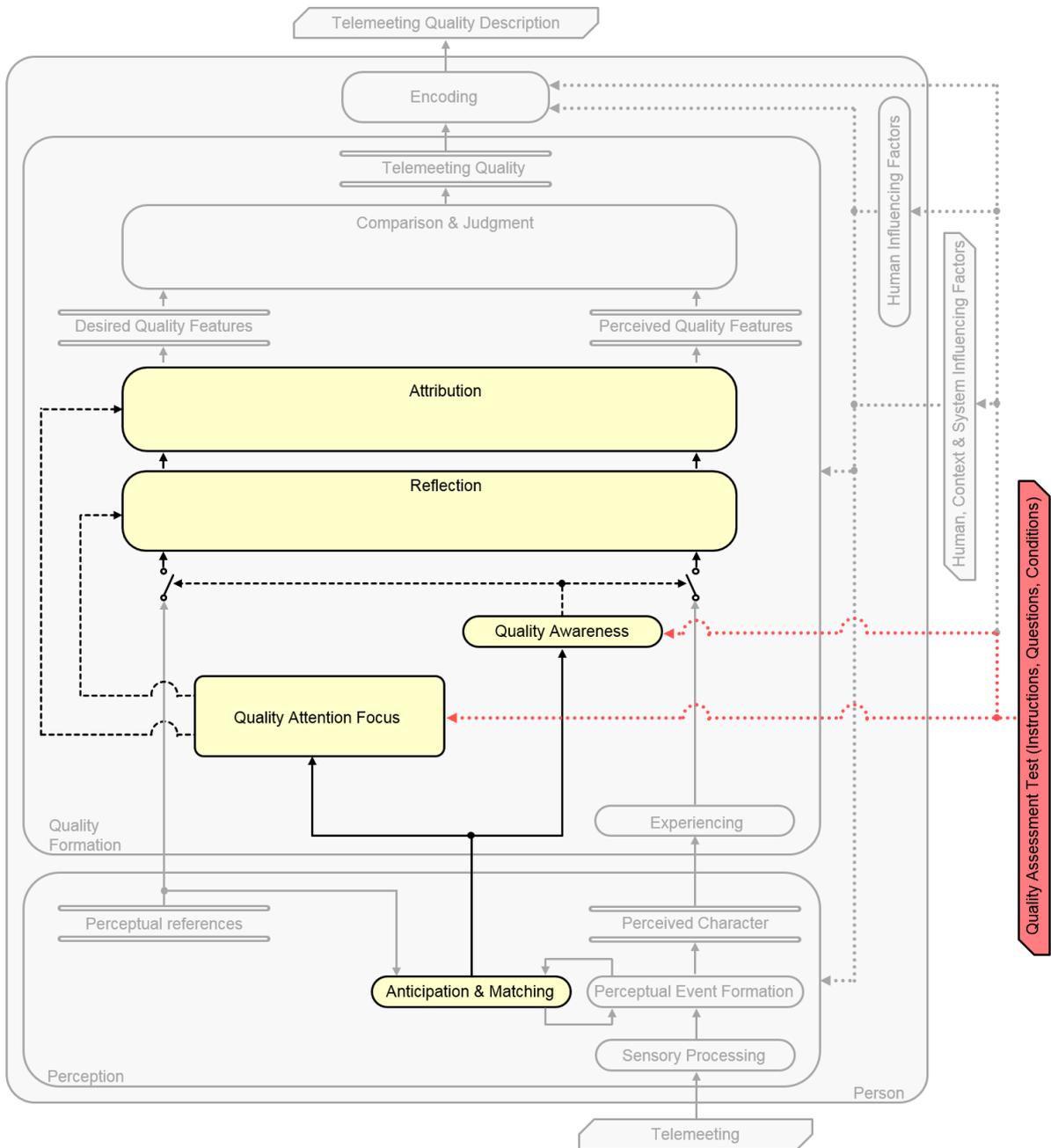


Figure 5.4: Quality formation process inside a telemeeting participant: Extension of process model with a quality awareness stage, according to the model from Raake & Egger, and a newly introduced quality attention focus stage.

Alexander Raake and Sebastian Egger. "Quality and Quality of Experience". In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 11-34

5.2.5 Summary of process model

Figure 5.4 visualizes the whole process model developed in the previous sections. This model version does not include the multiparty-specific concepts *Quality Aggregation Levels* and *Telecommunication & Group-Communication Components*. Those will be included in the next sections, after the concepts have been discussed in more detail.

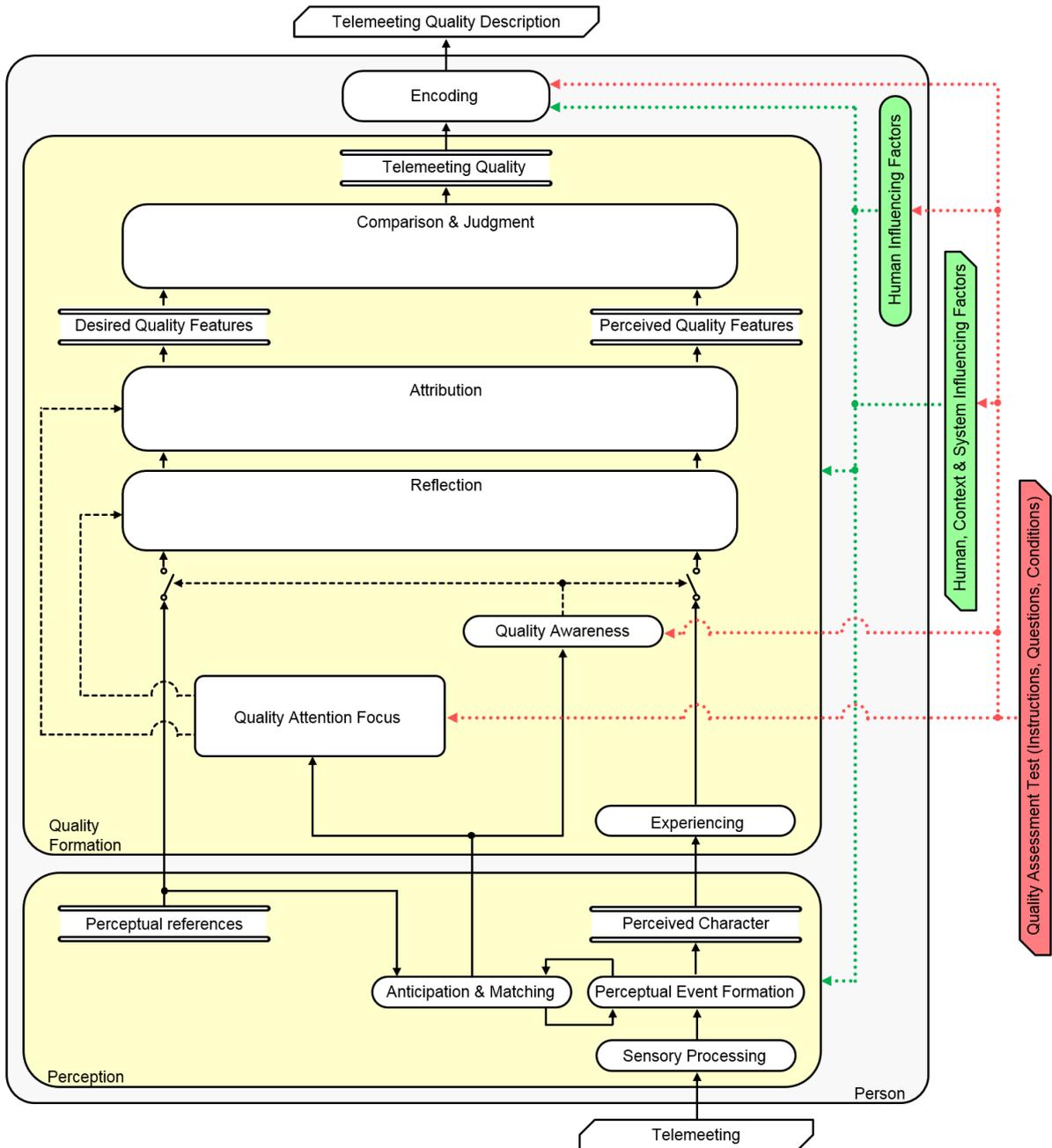


Figure 5.5: Quality formation process inside a telemeeting participant: Summary of the process model.

### 5.3 Different Aggregation Levels of Telemeeting Quality

A telemeeting system essentially provides a set of individual but simultaneous communication channels between multiple interlocutors, henceforth referred to as individual connections. The possibility to encounter asymmetric conditions, i.e. interlocutors are connected with different equipment and connection properties, raises the question to which extent the individual connections contribute to the overall system quality. This refers to the mental models that interlocutors form about the telemeeting system while communicating over those systems<sup>16</sup>. An example in which interlocutors may become aware of the individual connections is an asymmetric setting with strong quality impairments for one single interlocutor and no impairments for the other interlocutors. An example in which interlocutors may not consider individual connections is a symmetric setting in which no particular differences between interlocutors are perceived.

As a tool to investigate such potential differences between asymmetric and symmetric conditions, the present section introduces three different quality aggregation levels of telemeeting quality. The first level reflects how each interlocutor perceives the quality of these individual connections, henceforth referred to as *Individual Connection Quality* and abbreviated as  $Q_{ij}$ . The second level reflects how each interlocutor perceives the quality of the overall telemeeting, i.e. the composition of the individual connections, henceforth referred to as *Single-perspective Telemeeting Quality* and abbreviated as  $Q_i$ . The third level reflects how the whole group of interlocutors perceives the quality of the overall telemeeting, i.e. the combination of the judgments of all interlocutors, henceforth referred to as *Group-perspective Telemeeting Quality* and abbreviated as  $Q_{all}$ .

These three aggregation levels show that telemeeting quality is a function of perspective (perspective of individual interlocutors or of the whole group) and focus (focusing on the whole telemeeting, or on individual connections). This does not only concern the reporting and interpretation of results, it has also some non-trivial implications for the conduction of perceptual quality assessment tests: test methods need to be appropriate to collect the quality judgments from the right perspective, i.e. the perspective of the whole group or of individual participants; and test instructions need to be formulated in such a way that the subjects' focus is triggered as desired, i.e. focusing on individual connections or the overall composition of those connections.

This is especially relevant for Chapter 8, in which both symmetric and asymmetric conditions will be investigated and in which test participants will be explicitly asked to rate *Single-perspective Telemeeting Quality*  $Q_i$  and *Individual Connection Quality*  $Q_{ij}$ . For that reason, the remainder of this section provides a precise specification of the targeted quality aggregation levels.

<sup>16</sup> According to Norman, mental models are considered to be generated and modified during the interaction with a system; they are functional but need not be technically correct; and they depend on the user's technical background and experience with similar systems.

Donald A. Norman. "Some Observations on Mental Models". In: *Mental Models*. Ed. by Dedre Gentner and Albert L. Stevens. Psychology Press, 1983, pp. 7-14

### 5.3.1 Concept of Individual Connections and Individual Connection Quality

Figure 5.6 illustrates the individual connections in the form of signal paths between three interlocutors ( $IL_1$ ,  $IL_2$ ,  $IL_3$ ) who are connected via a telemeeting system. As the figure shows for such a three-party call, audio and video signals are sent from every interlocutor to each other interlocutor. In addition, each interlocutor may also get back some feedback signals, that are possibly altered versions of the audio or video signals that he or she is sending into the telemeeting. Examples from the audio domain are the presence of side tones or echoes, an example from the video domain is a window on the video screen showing the own camera picture.

As Figure 5.6 suggests, the exact interpretation of *individual connections* may focus on each individual signal path per direction between two interlocutors, may include both directions, and may include or exclude those feedback signal paths. An additional interpretation possibility is to consider an individual connection not between two interlocutors, but between an interlocutor and the telemeeting. This means that the concept of *Individual Connection Quality* is subject to different interpretations, which needs to be reflected by the employed test method, e.g. by proper test instructions. However, before these interpretations are explained in more detail, two aspects will be mentioned first, namely the notion that there exists a mutual influence of individual connection qualities and that two different categories of quality features may be distinguished.

### 5.3.2 Mutual Influence of Individual Connection Qualities

There is one additional aspect concerning *Individual Connection Quality*  $Q_{ij}$  that requires some clarification. At first glance,  $Q_{ij}$  appears to correspond to the quality of a conventional two-party connection. However, in Skowronek et al.<sup>17</sup> we discussed that

...both from a technical and a perceptual point of view, this is not entirely true. Technically it is possible that a degradation occurring in one connection can also affect the other connections. For example, an acoustic echo at the device of one interlocutor does not only mean that the other interlocutors hear an echo of their own voices (i.e. talker echo), but they also hear the echoed voices of all other interlocutors that are also fed back via that one echo-causing device (i.e. listener echo). Perceptually it is known that the context can influence quality judgments; hence it cannot be assumed that the quality perception of the considered individual connection in the multiparty call will be the same as the perception of the technically same connection in a two-party call.

That means, there is some mutual influence of individual connections taking place in a multiparty setting, which would not be there in a two-party setting. In the following sections, this mutual influence is inherently taken into account by analyzing the contributions of the individual signal paths to  $Q_{ij}$  directly in the multiparty context.

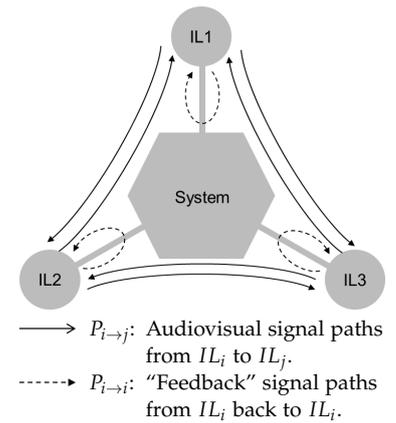


Figure 5.6: Concept of individual connections between telemeeting participants on the example of three participants. The individual connections are expressed by audiovisual signal paths between the interlocutors.

<sup>17</sup> Janto Skowronek, Katrin Schoenenberg, and Gunilla Berndtsson. "Multimedia Conferencing and Telemeetings". In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 217–228

### 5.3.3 Signal-based and Cue-based Quality Features

A quality rating is the outcome of a comparison process, in which the perceived quality features of the item are compared with its desired quality features. Those quality features can be grouped into different categories, see Section 4.3. In this context of *Individual Connection Quality*, however, the author defines two categories of quality features from a slightly different perspective: quality features extracted from the audio and video signals on that connection, here referred to as *signal-based features*, and quality features inferred from non-signal cues about that connection such as content or auxiliary information, here referred to as *cue-based features*. Examples for the first category are to hear or see distortions in the audio or video signal; examples for the second category are comments made by the other interlocutors such as “I can’t see you well”, “I do hear my own voice” or “Could you repeat, please?”

Which combination of those two categories of quality features will be used in the judgment process, especially in the experiments in Chapter 8, depends on two constraints. The first constraint is the question which individual connection is assessed by which interlocutor. The signal-based features can only be extracted from those signals that an interlocutor is actually receiving; for all other signal paths, he or she may consider only cue-based quality features. Figure 5.7 illustrates this on the situation for interlocutor  $IL_1$ . The second constraint is the question if there is any quality relevant information available at all. Regarding the signal-based features, this concerns the feedback signals as they may or may not bear information that a participant can use, e.g. presence or absence of echo. Regarding to the cue-based features, interlocutors may or may not say something from which the assessing participant can infer such cues. Thus, the former type of quality features depends on the technical conditions, the latter type depends on the technical condition and the behavior of interlocutors.

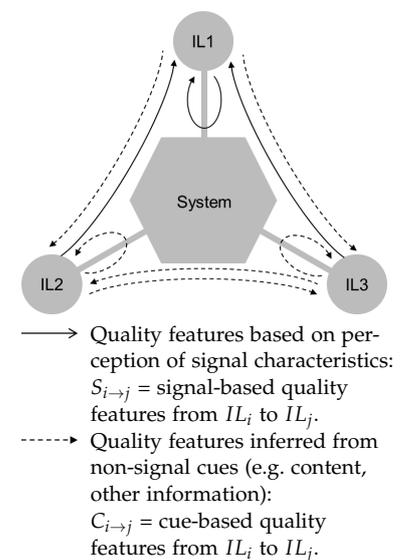


Figure 5.7: Concept of two types of quality features on the example of interlocutor  $IL_1$ .

### 5.3.4 Interpretations of Individual Connection Quality $Q_{ij}$

There are eight different possibilities to interpret the signal paths from Figure 5.6 as individual connections, given the three combinatory variables discussed above: *one direction vs. two directions, with feedback signals vs. without feedback signals, and connection between two interlocutors vs. connection between interlocutor and the system*. As a consequence, participants in a quality assessment test may have a different understanding of Individual Connection Quality  $Q_{ij}$  than the researcher. This could lead to an increased measurement noise or even to misinterpretation of results by the researcher.

To minimize such risks, a researcher may take two measures: First, proper test instructions and experimental designs may increase the probability that a certain interpretation is – at least predominantly – used by participants. Second, the interpretation of results acknowledges that different interpretations of an Individual Connection Quality  $Q_{ij}$  are possible and it includes a discussion, which interpretation

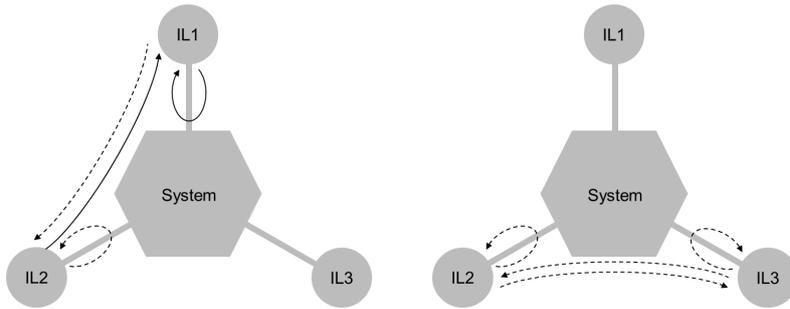
or mixture of interpretations were most likely used by participants, or which were at least intended in light of the actual experimental setup. To facilitate such interpretations of results, the following will examine in more detail those two of the eight different interpretations in which both directions and the feedback signals are included.

*Interpretation 1:*  $Q_{a,ij}$  Quality of the connection in both directions between  $IL_i$  and  $IL_j$ , seen from the perspective of  $IL_a$ .

The contributing paths are the two going from one interlocutor to the other and the feedback paths of the two interlocutors. Thus,  $Q_{a,ij}$  is a function of the relevant quality features on those four paths.

Figure 5.8 visualizes two examples from the perspective of  $IL_1$  that exemplarily show the relevant paths and quality features. From these examples the following generic rules can be extracted:

**If**  $a = i$  (observer is part of the connection),  
**then**  $Q_{a,ij} = Q_{a,aj} = f(S_{j \rightarrow a}, S_{a \rightarrow a}, C_{a \rightarrow j}, C_{j \rightarrow j})$ .  
**If**  $a \neq i \neq j$  (observer is not part of the connection),  
**then**  $Q_{a,ij} = f(C_{i \rightarrow j}, C_{j \rightarrow i}, C_{i \rightarrow i}, C_{j \rightarrow j})$ .



Example 1: Quality of the connection in both directions between  $IL_1$  and  $IL_2$ , seen from the perspective of  $IL_1$ :  
 $Q_{1,21} = f(S_{2 \rightarrow 1}, C_{1 \rightarrow 2}, S_{1 \rightarrow 1}, C_{2 \rightarrow 2})$

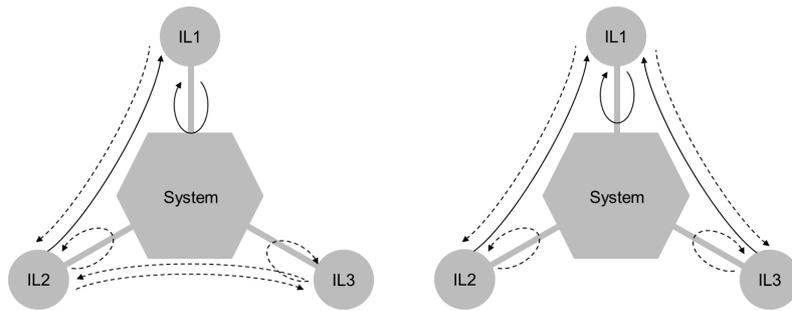
Example 2: Quality of the connection in both directions between  $IL_2$  and  $IL_3$ , seen from the perspective of  $IL_1$ :  
 $Q_{1,23} = f(C_{2 \rightarrow 3}, C_{3 \rightarrow 2}, C_{2 \rightarrow 2}, C_{3 \rightarrow 3})$

Figure 5.8:  $Q_{a,ij}$ : Quality of the connection in both directions between  $IL_i$  and  $IL_j$ , seen from the perspective of  $IL_a$ . Two examples from the perspective of interlocutor  $IL_1$ .

*Interpretation 2:*  $Q_{a,iT}$  Quality of the connection in both directions between an interlocutor  $IL_i$  and the telemeeting  $T$ , seen from the perspective of  $IL_a$ . The contributing paths for  $Q_{a,iT}$  are all outgoing paths from  $IL_i$ , all incoming paths to  $IL_i$ , and the feedback paths of all interlocutors. Thus  $Q_{a,iT}$  is a function of the relevant quality features on those seven paths.

Figure 5.9 visualizes two examples from the perspective of  $IL_1$  that exemplarily show the relevant paths and quality features. From these examples the following generic rules can be extracted:

**If**  $a = i$  (observer is part of the connection),  
**then**  $Q_{a,iT} = Q_{a,aT} = f(C_{a \rightarrow j}, C_{a \rightarrow k}, S_{j \rightarrow a}, S_{k \rightarrow a}, S_{a \rightarrow a}, C_{j \rightarrow j}, C_{k \rightarrow k})$ .  
**If**  $a \neq i$  (observer is not part of the connection),  
**then**  $Q_{a,iT} = f(S_{i \rightarrow a}, C_{i \rightarrow j}, C_{a \rightarrow i}, C_{j \rightarrow i}, S_{a \rightarrow a}, C_{i \rightarrow i}, C_{j \rightarrow j})$ .



Example 1: Quality of the connection in both directions between an interlocutor  $IL_2$  and the telemeeting  $T$ , seen from the perspective of  $IL_1$ :  $Q_{1,2T} = f(S_{2 \rightarrow 1}, C_{1 \rightarrow 2}, C_{2 \rightarrow 3}, C_{3 \rightarrow 2}, S_{1 \rightarrow 1}, C_{2 \rightarrow 2}, C_{3 \rightarrow 3})$

Example 2: Quality of the connection in both directions between an interlocutor  $IL_1$  and the telemeeting  $T$ , seen from the perspective of  $IL_1$ :  $Q_{1,1T} = f(S_{2 \rightarrow 1}, C_{1 \rightarrow 2}, S_{3 \rightarrow 1}, C_{1 \rightarrow 3}, S_{1 \rightarrow 1}, C_{2 \rightarrow 2}, C_{3 \rightarrow 3})$

Figure 5.9:  $Q_{a,ij}$ : Quality of the connection in both directions between  $IL_i$  and  $IL_j$ , seen from the perspective of  $IL_a$ . Two examples from the perspective of interlocutor  $IL_1$ .

*Implication for this work* The experimental studies in Chapter 8 target Interpretation 2 by means of dedicated instructions and corresponding phrasing of the questions in the questionnaire (see Chapter 8 for details).

### 5.3.5 Single-Perspective Telemeeting Quality $Q_i$

The second level of telemeeting quality – the *Single-Perspective Telemeeting Quality*  $Q_i$  – refers to the quality that a participant perceives from the whole telemeeting as such. In terms of quality aggregation, two concepts are possible, depending on the mental model that a participant has formed about the telemeeting system. One concept is that  $Q_i$  is a function of the *Individual Connection Qualities*  $Q_{ij}$ . This means that participants are first forming quality judgments of the individual connections and then combine them to an overall judgment. The other concept is that  $Q_i$  is a function of the individual signal-based or cue-based quality features that participants extract from the individual connections.

The difference between the two processes relates to whether either the quality features are first transformed into quality judgments and those are then aggregated over the individual connections, or whether the quality features are first aggregated over the individual connections and then transformed into one quality judgment. Although the difference between both processes may be rather small, it has some implications in many practical contexts, both from a perceptual and an engineering perspective.

From a perceptual perspective, the first concept presumes that participants have indeed a mental model of the individual connections in mind; otherwise they could not form a quality judgment of those individual connections first. The second concept does not require such a mental model of individual connections as participants may directly form one quality judgment for the whole telemeeting based on all quality features that they perceive, without the need to first assign those quality features to individual connections.

From an engineering perspective, both concepts would lead to different modeling approaches. The first concept suggests to model first the *Individual Connection Qualities*  $Q_{ij}$  based on existing modeling approaches for one-to-one conversations, and then the *Single-Perspective Quality*  $Q_i$  from the predictions of  $Q_{ij}$  by means of an appropriate aggregation function. The second concept suggests to directly model *Single-Perspective Telemeeting Quality*  $Q_i$  from the available input information representing the quality features. This means to calculate the aggregation of quality features by directly changing the internal computations of existing models or developing new models from scratch.

### 5.3.6 *Group-Perspective Telemeeting Quality* $Q_{all}$

The third level of telemeeting quality – the *Group-Perspective Telemeeting Quality*  $Q_{all}$  – refers to the quality that the whole group of participants is perceiving from the telemeeting. In terms of quality aggregation,  $Q_{all}$  is a function that represents a joint quality judgment from all participants. For modeling,  $Q_{all}$  can be calculated either from the *Single-Perspective Qualities*  $Q_i$ , from the *Individual Connections Qualities*  $Q_{ij}$ , directly from the quality features or from a combination of any of these three types of information.

Independently from the exact modeling approach, there is a conceptual difference between modeling the *Group-Perspective Telemeeting Quality*  $Q_{all}$  and modeling the *Single-Perspective Telemeeting Quality*  $Q_i$ . The computation of  $Q_i$ , which is a quality judgment of a single participant, is based on information from the perspective of that single participant ( $Q_{ij}$ , quality features). Extending a corresponding discussion in Skowronek et.al<sup>18</sup>, this means that the relation between  $Q_i$  and  $Q_{ij}$  or the quality features, respectively, can be determined from a perspective from *inside* the telemeeting. The computation of  $Q_{all}$ , which represents a joint quality judgment of a group of participants, is based on information from different participants. Thus determining  $Q_{all}$  requires a perspective *outside* of the telemeeting, as now different perceptions of the same system need to be integrated.

Here the term *outside* does not necessarily mean that  $Q_{all}$  is determined by someone who was not part of the telemeeting. Instead it means that determining  $Q_{all}$  is a process, in which all participants come to that joint quality judgment by changing from their personal perspective to a joint perspective. Such a process could be for instance that all are describing their personal experiences and then discussing the common experience. However,  $Q_{all}$  may be also determined by someone who was indeed not part of the telemeeting, such as the service provider who could consider  $Q_{all}$  as an abstract measure for benchmarking or optimization.

### 5.3.7 *Extension of the Quality Formation Process Model*

This section will describe how the concept of *Quality Aggregation Levels* relates to the quality formation process within a telemeeting

<sup>18</sup> Janto Skowronek, Katrin Schoenenberg, and Gunilla Berndtsson. "Multimedia Conferencing and Telemeetings". In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 217–228

participant by extending the process model from Section 5.2. To be more precise, only the two levels *Single-Perspective Telemeeting Quality*  $Q_i$  and the *Individual Connection Quality*  $Q_{ij}$  are incorporated, since the model concerns the process inside an individual interlocutor.

The model extension has three main components: (1) a process that decides which aggregation level is actually considered, henceforth referred to as *Quality Aggregation Level Selection*; (2) a process that decomposes the quality perception of the telemeeting into quality perception of individual connections, henceforth referred to as *Multiparty Decomposition*; and (3) a process that aggregates the quality perception of individual connections to a telemeeting quality perception, henceforth referred to as *Multiparty Aggregation*. This approach with three stages provides the flexibility required to account for the fact that there are four different scenarios possible in which the aggregation levels contribute to the quality formation process. The four scenarios may differ from person to person and even case-by-case, and they differ in the questions whether a decomposition takes place at all, whether an aggregation takes place at all, and in case it takes place, whether the aggregation is actually happening before or after forming a quality judgment. The scenarios are characterized in more detail as follows:

1. The person is judging the *Single-Perspective Telemeeting Quality*  $Q_i$  without decomposing the telemeeting into individual connections. In this scenario, the person has no need, possibility or capability to identify individual connections, and is forming a quality judgment using directly all quality features ("I perceive some slight distortions. Thus, the overall quality is moderate"). In mathematical terms this means:

$$Q = Q_i^a = D^a(\overrightarrow{X}_{des}^a, \overrightarrow{X}_{per}^a) \quad (5.1)$$

with  $D^a$  : quality judgment function comparing the desired quality feature vector  $\overrightarrow{X}_{des}^a$  and the perceived quality feature vector  $\overrightarrow{X}_{per}^a$ .

2. The person is judging the *Single-Perspective Telemeeting Quality*  $Q_i$  by decomposing the telemeeting into individual connections on a quality-feature level and aggregating those features without forming explicit judgments on the *Individual Connection Quality*  $Q_{ij}$ . In this scenario, the person indeed identifies the quality features of individual connections, but will form a quality judgment directly using all those quality features ("I perceive one participant distorted and the others without distortions. Thus the overall quality is moderate"). In mathematical terms this means:

$$Q = Q_i^b = D^b(A^b(\overrightarrow{X}_{des,ij}^b), A^b(\overrightarrow{X}_{per,ij}^b)) \quad (5.2)$$

with  $A^b$  an aggregation function over the individual connections of the quality feature vectors  $\overrightarrow{X}_{des,ij}^b$  and  $\overrightarrow{X}_{per,ij}^b$  per individual connection. The difference to the previous scenario is that the person might now focus on a different set of quality features, might weight

the quality features of the individual connections differently and might perform a certain probably non-linear aggregation operation. In mathematical terms, this difference can be expressed by  $\overrightarrow{X_{des}^a} \neq A^b(\overrightarrow{X_{des,ij}^b})$  and  $\overrightarrow{X_{per}^a} \neq A^b(\overrightarrow{X_{per,ij}^b})$ .

3. The person is judging the *Single-Perspective Telemeeting Quality*  $Q_i$  by decomposing the telemeeting into individual connections, forming judgments on the *Individual Connection Quality*  $Q_{ij}$  for each connection and then aggregating those judgments. In this scenario, the person forms quality judgments of the individual connections and then aggregates them into an overall judgment (“I perceive one participant distorted – so he has a bad connection – and the others without distortions – so they have good connections. Thus, the overall quality is moderate”). In mathematical terms this means:

$$Q = Q_i^c = A^c(D^c(\overrightarrow{X_{des,ij}^b}, \overrightarrow{X_{per,ij}^b})) \quad (5.3)$$

with  $A^c$  an aggregation function over the *Individual Connection Quality* scores  $Q_{ij}$ . The difference to the previous scenario is that the person might now use a different aggregation function over individual connections, or that the judgment function is actually different for individual connections and for the whole telemeeting. In mathematical terms, this can be expressed by  $A^c \neq A^b$  and  $D^c \neq D^b$ .

4. The person is judging an *Individual Connection Quality*  $Q_{ij}$ , meaning that no aggregation across individual connections takes place. In this scenario, the person is directly focusing on one individual connection (“I perceive one participant distorted – so she has a bad connection”). In mathematical terms this means:

$$Q = Q_{ij} = D^d(\overrightarrow{X_{des,ij}^d}, \overrightarrow{X_{per,ij}^d}) \quad (5.4)$$

With these four scenarios in mind, the inclusion of the three stages *Quality Aggregation Level Selection*, *Multiparty Decomposition*, and *Multiparty Aggregation* into the process model can now be motivated and described, see also Figure 5.10.

*Multiparty Decomposition and Multiparty Aggregation stages* The four scenarios suggest that the activation of these stages and their exact location in the process model depends on each scenario. For Scenario 1 both stages need to be deactivated; Scenario 2 requires the aggregation process before the *Comparison & Judgment* process; Scenario 3 requires the aggregation process after the *Comparison & Judgment* process; and for Scenario 4 the aggregation process needs to be deactivated. In order to have only one common visualization for these different cases, Figure 5.10 indicates the two stages as sub-processes: the *Multiparty Decomposition* is considered as part of the *Attribution* stage, the *Multiparty Aggregation* as part of the *Comparison & Judgment* stage.

The argumentation for including the *Multiparty Decomposition* into the *Attribution* stage is as follows: In the *Attribution* stage, the person

is judging whether the perceived quality-relevant features coming from the *Reflection* stage are subject to the item under consideration, i.e. the telemeeting. This means the person is comparing the perceived quality-relevant features with a mental model of the telemeeting system. Now, if the person is considering individual connections of the telemeeting, then those individual connections will constitute – or at least become part of – that mental model. In particular, the person is – more or less subconsciously – answering the question for himself or herself whether the perceived quality-relevant features could be attributed to the individual connections, to the overall system or to something else, such as the interlocutors, environments etc. Hence, the *Attribution* process appears to be the appropriate place to integrate the *Multiparty Decomposition* stage.

The argumentation for including the *Multiparty Aggregation* into the *Comparison & Judgment* stage is as follows: During the *Comparison & Judgment* process, the desired and perceived quality features are compared and transformed into a quality judgment. Considering that the sets of desired and perceived quality features are assumed to consist of more than one individual feature, quality judgments can be considered as multi-dimensional constructs (see for instance Wältermann<sup>19</sup> for a more in-depth discussion on dimensions of quality).

However, quality judgments are often – for instance in a quality assessment test – expressed in a single-dimensional manner, also referred to as *Integral Quality*<sup>20</sup>. That means, some aggregation across quality dimensions may take place during the *Comparison & Judgment* stage. Now, considering the quality features of the individual connections as a number of such dimensions, the multiparty aggregation essentially represents such a similar mapping of dimensions onto a single-dimensional scale. Hence, from a mathematical perspective, there is no conceptual difference between the aggregation of individual features when forming a single-dimensional quality judgment, and the aggregation of features over individual connections. Thus, the *Comparison & Judgment* stage appears to be the appropriate place to integrate the *Multiparty Aggregation* stage.

*Quality Aggregation Level Selection stage* The four scenarios suggest a process to control the other two stages in terms of activation and deactivation and in terms of interaction with their higher-level processes *Attribution* and *Comparison & Judgment*. To this aim, a *Quality Aggregation Level Selection* stage essentially defines which *Quality Aggregation Level* shall be used, or in other words, which level the person is focusing on. For that reason, this stage is included into the model as a subprocess of the *Quality Attention Focus* stage, as it covers a multiparty-specific aspect of the person's attention focus. Furthermore, being part of the *Quality Attention Focus*, the *Quality Aggregation Level Selection* is essentially using the same input information, which either stems from the *Anticipation & Matching* stage or from the quality assessment test. However, the path in Figure 5.10 from the quality assessment test is directly drawn to the *Quality Aggregation Level*

<sup>19</sup> Marcel Wältermann. *Dimension-based Quality Modelling of Transmitted Speech*. Springer, 2013

<sup>20</sup> Sebastian Möller. *Assessment and Prediction of Speech Quality in Telecommunications*. Kluwer Academic publishers, 2000

Marcel Wältermann. *Dimension-based Quality Modelling of Transmitted Speech*. Springer, 2013

Selection stage, and not to its superordinate stage *Quality Attention Focus*. This is done to emphasize that a telemeeting quality assessment test can particularly be designed to assess a specific *Quality Aggregation Level*, as this is addressed in the empirical research described in Chapter 8.

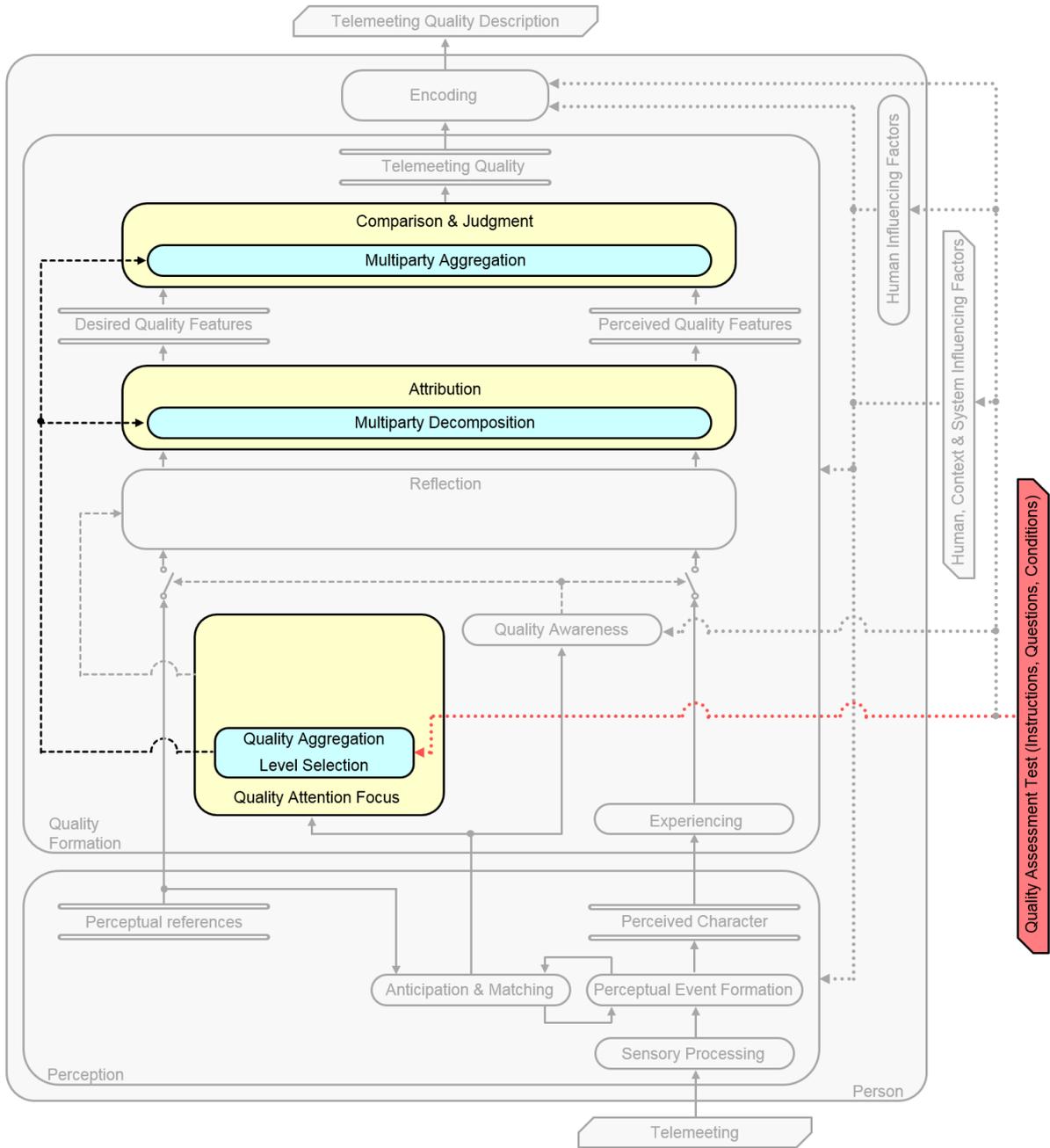


Figure 5.10: Quality formation process inside a telemeeting participant: Extension of process model with the concept of Quality Aggregation Levels, realized in three stages. One stage decomposes the telemeeting into individual connections, which is considered as a sub-process of the Attribution stage; one stage aggregates the individual connections, which is considered as a sub-process of the Comparison & Judgment stage; and one stage to control the decomposition and aggregation. See text for further details.

#### 5.4 *Telecommunication and Group-Communication Components of Telemeeting Quality*

A telemeeting system enables a communication between a group of usually more than two interlocutors who are located at physically remote locations. Thus, on the one hand a telemeeting is a group conversation, on the other hand it is a conversation over a telecommunication system. With regard to a conceptual telemeeting quality model, the author proposes to consider these two aspects as two components of telemeeting quality: a *Telecommunication Component* and a *Group-Communication Component*. The term *Telecommunication Component* refers to the user's perception of the system's performance in transmitting information such as audio and video signals. The term *Group-Communication Component* refers to the user's perception of the system's performance in facilitating a group conversation.

##### 5.4.1 *Characterization of the Telecommunication and Group-Communication Components*

The two components can more precisely be defined by the quality features that are used in the quality judgment process. The *Telecommunication Component* refers to features that can directly be perceived from impairments of the audio and video signals, whereas such impairments may be perceived when listening or viewing but also when talking and gesticulating. In terms of the quality feature categorization proposed in Möller et al.<sup>21</sup>, such features encompass the feature levels of direct perception and action.

The *Group-Communication Component* refers to features that can be perceived from impairments of the group conversation as such. In terms of the quality feature categorization proposed in Möller et al., such features encompass the feature levels of interaction (e.g. conversational aspects) and usage instance (e.g. effectiveness and efficiency in reaching a certain goal or aesthetics). Additional features proposed by the author are inspired by the exhaustive literature on human group-communication (see Chapter 2). Those features address group-communication aspects such as the perceived effort to establish a common ground (e.g. speaker identification effort, focal assurance, perception of non-verbal feedback signals, and topic comprehension effort), the perception of group-conversation dynamics (e.g. negotiating of who's turn is next), and the perception of cognitive load (e.g. concentration effort and fatigue).

Möller et al. provide for each of their categories an example list of quality features. Table 5.1 assigns that list of examples to the two categories discussed here, augmented with the proposed additional features addressing group-communication aspects.

<sup>21</sup> Sebastian Möller, Marcel Wältermann, and Marie-Neige Garcia. "Features of Quality of Experience". In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 73–84

Quality Features that constitute the Telecommunication Component		
Feature category	Modality	Example features
level of direct perception <sup>1</sup>	audio & speech	localization, timbre, coloration, noisiness, loudness, continuity
	video	sharpness, darkness, brightness, contrast, flicker, distortion, color perception
	audiovisual	balance and synchronism
level of action <sup>1</sup>	audio & speech	perception of sidetone, echo, or double-talk degradations
	video	involvement and immersion, perception of space, perception of own motions
Quality Features that constitute the Group-Communication Component		
Feature category	Modality	Example features
level of interaction <sup>1</sup>	all	responsiveness, naturalness of interaction, communication efficiency, and conversation effectiveness
level of the usage instance <sup>1</sup>	all	learnability, intuitivity, effectiveness and efficiency for reaching a particular goal, ease of use, non-functional features such as the personality of the interaction partner (human or machine) or aesthetics
level of establishing common ground <sup>2</sup>	all	speaker identification effort, focal assurance, perception of non-verbal feedback signals, topic comprehension effort
level of group-conversation dynamics <sup>2</sup>	all	negotiation of who's turn is next
level of cognitive load <sup>2</sup>	all	concentration effort, fatigue
Remarks:		
<sup>1</sup> : Feature category according to Möller et al.: Sebastian Möller, Marcel Wältermann, and Marie-Neige Garcia. "Features of Quality of Experience". In: <i>Quality of Experience - Advanced Concepts, Applications, Methods</i> . Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 73–84		
<sup>2</sup> : Additional feature category specific to multiparty communication, see Chapter 2.		

Table 5.1: List of example quality features for the two categories transmission-related and group-conversation-related features.

#### 5.4.2 Influencing the Quality Attention Focus

Whether a user is concerned about the telecommunication component, the group-communication component, or a mixture of both, plays an important role in telemeeting quality assessment. The reason is that this question influences the quality formation process inside a user. More specifically, the two components influence the *Quality Attention Focus*, which has been introduced in Section 5.2 as that cognitive process that focuses the attention to a limited number of quality-relevant aspects, i.e. the set of quality features, that a user is actually considering when making the quality judgment. This influence will be now explained in more detail for the specific scenario of a quality assessment test.

#### 5.4.3 Quality Attention Focus in Assessment Tests

This section discusses a number of aspects that can influence the *Quality Attention Focus* of persons taking part in a quality assessment test. While the list of those aspects stems from the more general considerations made in Skowronek et al.<sup>22</sup> on the conduction of multiparty quality assessment tests, the discussion here will concentrate

<sup>22</sup> Janto Skowronek, Katrin Schoenenberg, and Gunilla Berndtsson. "Multi-media Conferencing and Telemeetings". In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 217–228

on the influence of those aspects on the *Quality Attention Focus*.

Before going into the details, two aspects should be mentioned here. First, the investigator needs to decide to which extent he or she wants to control for the *Quality Attention Focus*, depending on the actual quality assessment goals. That means that the following discussions are intended to provide an overview of the possibilities (“can do”) and should not be interpreted as strict requirements (“must do”). Second, reviewing the approaches presented below, the critical reader will note that an exact control of the *Quality Attention Focus* is nevertheless very limited, because the mentioned effects of the aspects on the *Quality Attention Focus* are hard to formally validate. Therefore, great care must be taken when designing the quality assessment test, either by attempting to control for the *Quality Attention Focus*, or at least by properly considering the *Quality Attention Focus* when reporting and interpreting the results.

*Questionnaire* The probably most important aspect is the questionnaire, more specifically the choice of measurement variables and the exact wording of the corresponding questions or scale labels. One approach to control the *Quality Attention Focus* is to choose measurement variables that are directly referring to individual quality features, and to phrase the corresponding questions or scale labels such that the test participants understand those quality features. While this appears to be a straight-forward approach, the practical implementation is not trivial; a matter well known in the research field of multidimensional quality assessment, which essentially seeks to assess quality in terms of individual quality features. As an example, Wältermann<sup>23</sup> describes a thorough process of refining and simplifying the choice of measurement variables and rating labels. Transferring this to the telemeeting case, the approach of asking directly for individual quality features appears to be feasible due to the existing methods<sup>24</sup> on multidimensional quality assessment.

Another approach to control the *Quality Attention Focus* is necessary when the desired measurement variables should cover quality in a more holistic or utilitarian sense, which is often referred to as *overall quality*<sup>25</sup>. If a simple question such as “What was the overall quality?” is asked in a telemeeting quality assessment test, then there is the possibility that test participants interpret this question quite differently, depending on the *Quality Attention Focus* that each test participant has. If the researcher wants to limit the interpretation of the overall quality question to a certain *Quality Attention Focus*, there are two possibilities, which can be also combined.

First, a more specific phrasing of the question could steer the *Quality Attention Focus*. For example, instead of “What was the overall quality?” one could ask “What was the overall quality of the system’s technical condition?” or “What was the overall quality of the telemeeting in terms of facilitating the communication?”. The first alternative would likely steer the *Quality Attention Focus* to the *Telecommunication Component*, due to the keyword “technical condition”; the

<sup>23</sup> Marcel Wältermann. *Dimension-based Quality Modelling of Transmitted Speech*. Springer, 2013

<sup>24</sup> Marcel Wältermann. *Dimension-based Quality Modelling of Transmitted Speech*. Springer, 2013

ITU-T. *Recommendation P.806 - A Subjective Quality Test Methodology using Multiple Rating Scales*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2014  
Friedemann Köster and Sebastian Möller. “Perceptual Speech Quality Dimensions in a Conversational Situation”. In: *Proceedings of the 16th Annual Conference of the Speech Communication Association (Interspeech)*. Dresden, Germany, Sept. 2015, pp. 2544–2548

<sup>25</sup> The terms mouth-to-ear quality or end-to-end quality are often used as synonyms for overall quality in order to emphasize that the whole transmission path is considered.

The term integral quality is used when such overall quality is considered as the composition of individual quality dimensions. For more details, see:

Sebastian Möller. *Assessment and Prediction of Speech Quality in Telecommunications*. Kluwer Academic publishers, 2000

Alexander Raake. *Speech Quality of VoIP – Assessment and Prediction*. Chichester, West Sussex, UK: Wiley, 2006

Marcel Wältermann. *Dimension-based Quality Modelling of Transmitted Speech*. Springer, 2013

second alternative would likely steer the *Quality Attention Focus* to the *Group-Communication Component*, due to the keyword “communication”. However, there is a disadvantage to this approach: using self-defined questions bears the danger that different researchers are actually measuring something different, even though they might report results under the same term *overall quality*. Unfortunately, there is no well-defined and validated set of such questions specifically for the telemeeting context available yet. As an alternative to phrasing own questions, one could use the available standardized mean opinion scores (MOS) according to ITU-T Rec. P.800<sup>26</sup> for listening-only tests (“Quality of the speech”) and conversation tests (“What is your opinion of the connection you have just been using?”). However, then it is necessary to consider how these questions would determine the *Quality Attention Focus* in terms of the *Telecommunication* and *Group-Communication Components*, in order to appropriately report and interpret the results.

A second possibility to limit the interpretation of the overall quality question to a certain *Quality Attention Focus* is the presentation of a small set of additional questions, each asking for different individual aspects of quality. Such an approach is also known in the field; for instance, the standards ITU-T Rec. P.800<sup>27</sup> and especially ITU-T Rec. P.805<sup>28</sup> propose to ask additional questions in order to cover different quality aspects. While the idea to use such more detailed questions essentially represents the approach of asking for individual quality features, the argumentation here in terms of overall quality is the following: If the questionnaire contains both a more loosely defined question asking for overall quality and a number of dedicated questions asking for individual quality features, then it is likely that participants form their overall judgment on the basis of the individual questions: not necessarily in a strictly mathematical sense, but at least participants will consider those individual aspects. In such a case, the ratio between questions on quality features targeting the *Telecommunication Component* and the *Group-Communication Component* can influence the *Quality Attention Focus*, when subjects answer the overall quality question.

*Test instructions* Next to the questionnaire, the test instructions – both verbal and written – can be explicitly designed to influence the *Quality Attention Focus*. If the test instructions clearly explain that test participants should focus on the transmission characteristics or on the communication as such, or even on both, the *Quality Attention Focus* will be steered accordingly. However, two aspects should be considered in such an approach. First, the instructions should be formulated such that subjects are able to understand what they should focus on; using examples might be useful here. Second, the instructions should be supported by a proper choice and wording of the questions as discussed above; otherwise test participants start to forget the exact instructions throughout the test.

<sup>26</sup> ITU-T. *Recommendation P.800 - Methods for objective and subjective assessment of quality*. International Standard. Geneva, Switzerland: International Telecommunication Union, 1996

<sup>27</sup> ITU-T. *Recommendation P.800 - Methods for objective and subjective assessment of quality*. International Standard. Geneva, Switzerland: International Telecommunication Union, 1996

<sup>28</sup> ITU-T. *Recommendation P.805 - Subjective evaluation of conversational quality*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2007

*Stimuli and choice of conditions* Another aspect that can influence the *Quality Attention Focus* is the choice of test stimuli in listening-only or viewing-only tests, or the choice of system conditions in conversation tests. More precisely, the type of degradations and the strength of those degradations need to be carefully chosen, if it is desired to control the *Quality Attention Focus* in the experiment. Alternatively, the type and strength of degradations should at least be reviewed in terms of their potential influence on the *Quality Attention Focus*, in order to properly report and interpret results.

An example for the impact of the strength of a degradation is packet loss. If the signal distortions are minor due to a rather low loss rate, then the *Quality Attention Focus* will likely be on the telecommunication-related features (perception of some temporal discontinuity). If the signal distortions are more severe due to a rather high loss rate and affect speech intelligibility or even speaker identification, then the focus will likely be on the group-communication-related features. An example for the impact of the choice of degradations is delay, see for instance Schoenenberg<sup>29</sup>. Delay is known to affect the conversation, as it triggers changes and interruptions of the conversation, while it is not perceived as a signal distortion, as the speech signal as such is not degraded. Thus, judging the quality of delay in a telemeeting context requires that the *Quality Attention Focus* is on the group-communication-related features.

*Conversation task* If the telemeeting quality test shall consist of a conversation test, the considerations and recommendations on conducting such tests in Skowronek et al.<sup>30</sup> and ITU-T Rec. P.1301<sup>31</sup> emphasize the need for a deliberate choice of the conversation task. In terms of the *Quality Attention Focus*, this means that the investigator should consider, how critical the chosen conversation task is for the perception of telecommunication-related and group-communication-related quality features. This is actually strongly linked with the above mentioned choice of the conditions. For example, if a test contains only rather minor distortions in order to limit the *Quality Attention Focus* towards the *Telecommunication Component*, then the investigator might want to choose the conversation task such that those minor distortions are actually perceived, i.e. the conversation task should not be too cognitively demanding. As another example, if the test contains rather strong distortions affecting speech intelligibility or speaker identification in order to steer the *Quality Attention Focus* towards the *Group-Communication Component*, then the investigator might want to choose the conversation task such that test participants strongly rely on intelligibility and speaker identification.

#### 5.4.4 Extension of the Quality Formation Process Model

This section describes how the concept of *Telecommunication* and *Group-Communication Components* relates to the quality formation process in the mind of a telemeeting participant by extending the process

<sup>29</sup> Katrin Schoenenberg. "The Quality of Mediated-Conversations under Transmission Delay". PhD Thesis. Technische Universität Berlin, Germany, 2016

<sup>30</sup> Janto Skowronek, Katrin Schoenenberg, and Gunilla Berndtsson. "Multi-media Conferencing and Telemeetings". In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 217–228

<sup>31</sup> ITU-T. *Recommendation P.1301 - Subjective quality evaluation of audio and audiovisual telemeetings*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2012

model from Section 5.2. The model extension consists of a process that decides which weighting of the two components is actually used, see also Figure 5.11. This *Telecommunication & Group-Communication Component Weighting* stage is integrated into the process model as a sub-process of the *Quality Attention Focus* stage.

The argumentation for this way of including the new stage is as follows: The previous sections discussed in detail how the concept of *Telecommunication* and *Group-Communication Components* is influencing the *Quality Attention Focus*, suggesting to place the new process as an extra stage before the *Quality Attention Focus* stage. However, in the author's view it is difficult to coherently separate these two processes. The *Quality Attention Focus* stage was introduced in Section 5.2 as a process to control which quality features shall be selected from all incoming features during the *Reflection* process. Now, the new *Telecommunication & Group-Communication Component Weighting* stage essentially determines as well, which quality features shall be selected. For that reason, the inclusion of the new stage as a sub-process appears to be more appropriate.

Furthermore the path in Figure 5.11 from the quality assessment test is directly drawn to the *Quality Attention Focus* stage. Similar to the *Quality Aggregation Level Selection* stage, this emphasizes the relevance of the assessment test design for the empirical research described in Chapter 7.

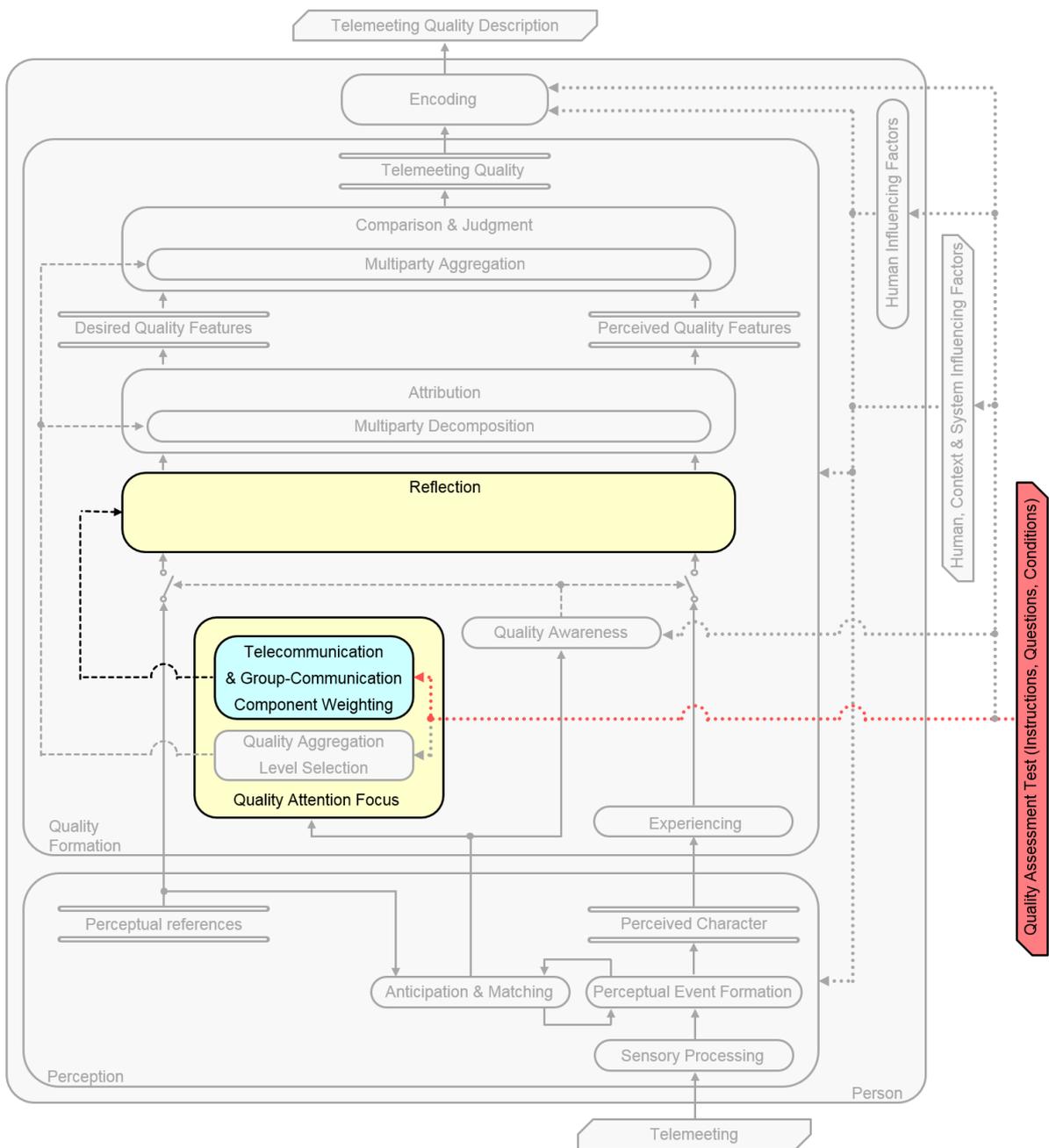


Figure 5.11: Quality formation process inside a telemeeting participant: Extension of process model with the concept of Telecommunication & Group-Communication Components, realized as a sub-process of the Quality Attention Focus stage.

## Summary

There are different facets of telemeeting quality that require a solid description in order to avoid misinterpretation of research results. Two such aspects are the concepts of *Quality Aggregation Levels* and *Telecommunication & Group-Communication Components*, which play an important role in the quality perception process.

*Quality Aggregation Levels* refer, on the one hand, to the perspective of either individual interlocutors or of the whole group. On the other hand, they refer to the focus on either the whole telemeeting or on individual connections. Therefore, this chapter defined three *Quality Aggregation Levels*:

1. *Individual Connection Quality*  $Q_{ij}$ : This reflects how each interlocutor perceives the quality of the individual connections between all interlocutors.
2. *Single-perspective Telemeeting Quality*  $Q_i$ : This reflects how each single interlocutor perceives the quality of the overall telemeeting.
3. *Group-perspective Telemeeting Quality*  $Q_{all}$ : This reflects how the whole group of interlocutors perceives the quality of the overall telemeeting, and may be used as an abstract tool for system designers.

In that context the exact interpretation of individual connections may focus on each individual signal path per direction between two interlocutors, may include both directions, and may include or exclude feedback signal paths (i.e. sidetone, echo, own camera picture). An additional interpretation possibility is to consider an individual connection not between two interlocutors, but between an interlocutor and the telemeeting. This means, also the concept of *Individual Connection Quality* is subject to different interpretations, which requires proper test instructions and reporting of results. Concerning the *Single-perspective Telemeeting Quality*  $Q_i$ , two principles of quality aggregation are possible. Either  $Q_i$  is an aggregation of the *Individual Connection Quality* scores  $Q_{ij}$ , which means that participants are first forming quality judgments of the individual connections and then combine them to an overall judgment. Or  $Q_i$  is an aggregation of the individual quality features that participants extract from the individual connections, without consciously forming a quality judgment about those individual connections.

The *Telecommunication & Group-Communication Components* of *Telemeeting Quality* reflect the duality that a telemeeting is, on the one hand, a group conversation, and, on the other hand, a conversation over a telecommunication system. In that respect, the *Telecommunication Component* refers to the user's perception of the system's performance in transmitting information such as audio and video signals; the *Group-Communication Component* refers to the user's perception of the system's performance in facilitating a group conversation. More specifically, the two components are constituted by the set of

quality features that the participant is using when forming a quality judgment. Furthermore, the two components influence the *Quality Attention Focus*, a term introduced here as a cognitive process that focuses the attention to a limited number of quality-relevant aspects, i.e. the set of quality features.

The impact of the *Quality Aggregation Levels* and the *Telecommunication & Group-Communication Components* on quality perception can best be characterized by adapting an existing model of the quality formation process in the person's mind. The model describes quality perception as a multi-stage process and is presented in Figure 5.12 as a block diagram with a signal flow from the bottom to the top: Sensory information is perceived using bottom-up and top-down processes (*Perception*) leading to a perceived character of the telemeeting. Then, during the act of *Experiencing*, and being triggered by some *Quality Awareness*, the person reflects on the perceived character and its desired character (*Reflection*) and then the person attributes the different features to either the system or to something else, such as the conversation partners or their environments (*Attribution*). The result of these processes is a set of quality-relevant aspects, i.e. *Quality Features*, for the perceived and desired nature of the telemeeting, which the person is then comparing against each other to form a quality judgment (*Comparison & Judgment*). Finally, an encoding process enables to describe the person's internal quality judgment to the outside world, e.g. by means of a rating on a scale.

To integrate the *Quality Aggregation Levels* and the *Telecommunication & Group-Communication Components*, the model is extended first by inserting the above mentioned process of *Quality Attention Focus* with two sub-processes *Quality Level Aggregation Selection* and *Telecommunication & Group-Communication Component Weighting*. Then the model is extended with two further sub-processes, a *Multiparty Decomposition* process as part of the *Attribution* stage, and a *Multiparty Aggregation* process as part of the *Comparison & Judgment* stage. In this framework, the *Quality Level Aggregation Selection* sub-process determines whether the person is focusing on individual connections and how they contribute in the person's mental model to the overall telemeeting system. This in turn means that the *Quality Aggregation Level Selection* controls the *Multiparty Decomposition* and *Multiparty Aggregation* processes. The *Telecommunication & Group-Communication Component Weighting* sub-process determines whether the person is more concerned about quality aspects of signal transmission or group-communication, meaning that it controls the *Reflection* stage.

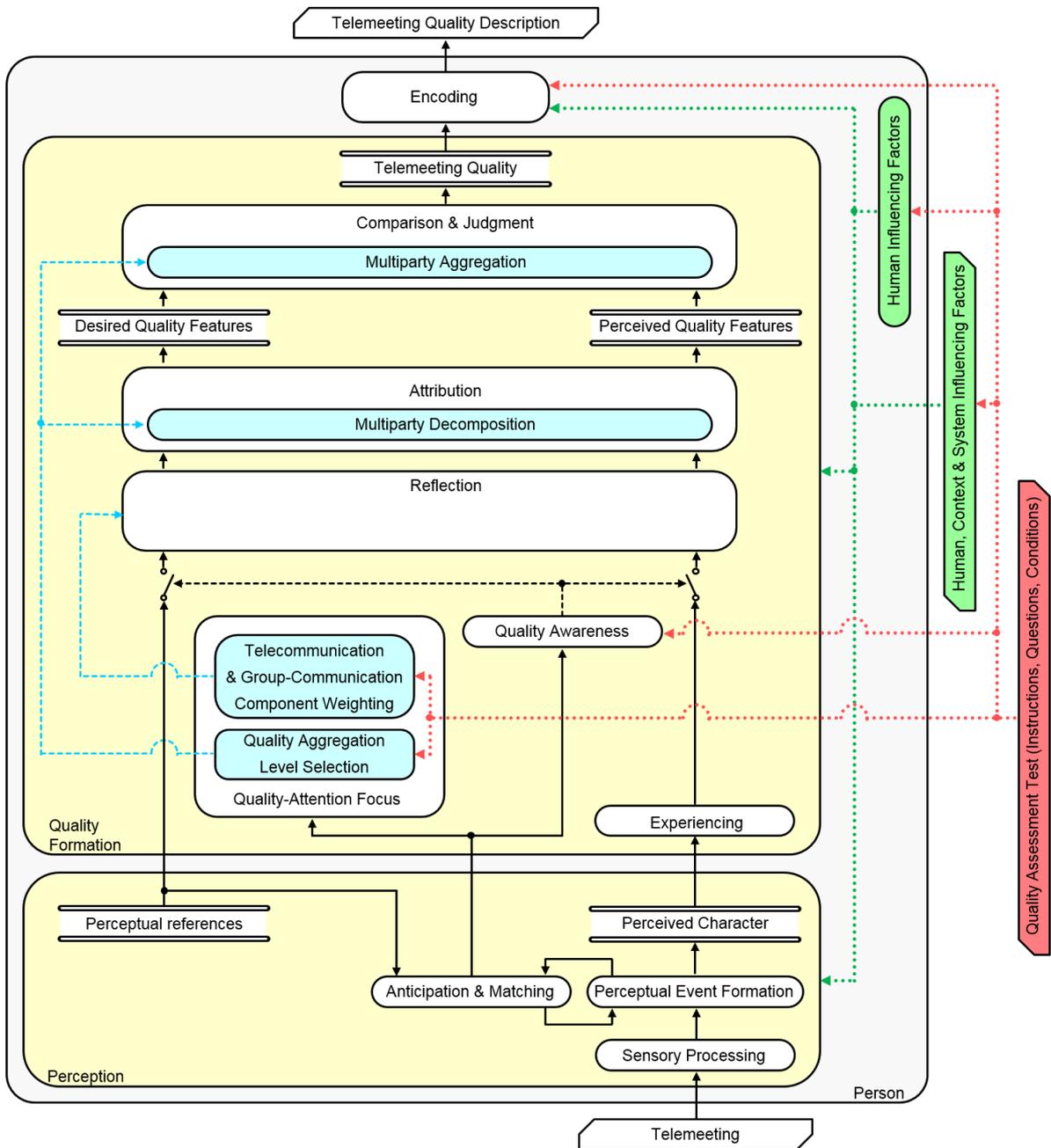


Figure 5.12: Full model of the quality formation process inside a telemeeting participant.

## 6

# *Methodology for the Perceptual Assessment of Telemeeting Quality*

---

---

This chapter contains text passages that either stem from previous texts of the author or that are based on such source texts. For readability purposes those passages are not marked as quotations. The main source documents used for this chapter are:

- Janto Skowronek, Julian Herlinghaus, and Alexander Raake. "Quality Assessment of Asymmetric Multiparty Telephone Conferences: A Systematic Method from Technical Degradations to Perceived Impairments". In: *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*. International Speech Communication Association. Lyon, France, Aug. 2013, pp. 2604–2608
- Janto Skowronek, Katrin Schoenberg, and Alexander Raake. "Experience with and insights about the new ITU-T standard on quality assessment of conferencing systems". In: *Proceedings of the international conference of acoustics (AIA-DAGA 2013)*. Merano, Mar. 2013, pp. 444–447
- Janto Skowronek and Alexander Raake. "Conceptual model of multiparty conferencing and telemeeting quality". In: *Proceedings of the 7th International Workshop on Quality of Multimedia Experience (QoMEX 2015)*. Pilos, Greece, May 2015, pp. 1–6
- Janto Skowronek et al. *Proposed main body and normative annex for draft recommendation P.AMT*. ITU-T Contribution COM 12 - C318 - E. Geneva, Switzerland: International Telecommunication Union, May 2012
- Janto Skowronek et al. *Proposed non-normative appendices for draft recommendation P.AMT*. ITU-T Contribution COM 12 - C319 - E. Geneva, Switzerland: International Telecommunication Union, May 2012
- Janto Skowronek and Alexander Raake. *Report on a study on asymmetric multiparty telephone conferences for a future update of P.1301*. ITU-T Contribution COM 12 - C134 - E. Geneva, Switzerland: International Telecommunication Union, Dec. 2013
- Janto Skowronek. *Document describing scenarios*. Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, Apr. 2013
- Janto Skowronek. *Improved model for audio-only communication*. Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, Dec. 2013
- Janto Skowronek. *Improved model for audiovisual communication*. Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, Nov. 2014
- Janto Skowronek. *Final Project Report – Quality of Multiparty Audio-Video Communication*. Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, Dec. 2014

Since additional documents served as further background material, see Appendix A for the full list of source documents and their contributions to this chapter.

---

---

### *What this chapter is about*

This chapter presents the fundamental test methodology that has been used to design and conduct perceptual multiparty quality assessment tests and to prepare the collected perceptual ratings for efficient statistical analysis and modeling. Further, this chapter provides a general guideline to conduct such multiparty tests and is written independently from the actual experiments reported in the following chapters.

## 6.1 Introduction

As already discussed in Chapter 5, the experimental setup in terms of instructions, conditions, questionnaires, and general context can influence results. In order to achieve a high reproducibility and comparability of results, the common approach is to conduct standardized quality assessment tests following agreed-upon protocols (see Section 4.4 for an overview). However, such protocols have been developed for specific evaluation scenarios, i.e. the systems under test and the evaluation goals, and may not be blindly used for new evaluation scenarios. Instead, the appropriateness of existing protocols for a new evaluation scenario needs to be verified and, if necessary, modified or even replaced.

In that line of thought, the quality assessment of telemeetings should be considered as such a new evaluation scenario given the two main differentiators between multiparty telemeetings and two-party audio or audiovisual telecommunication: special communicative situation and importance of asymmetric conditions. For that reason, the ITU-T as one major standardization body dealing with the quality assessment of telecommunication systems has – with contributions from the author – developed Recommendation P.1301<sup>1</sup> as a new test protocol specifically for the multiparty telemeeting scenario. Accordingly, this chapter will present the essential aspects of ITU-T Recommendation P.1301 and provide guidance on how to apply this method (Section 6.2).

While ITU-T Recommendation P.1301 is the first standard document on testing multiparty scenarios, this document is faced with the fact that a number of aspects on multiparty quality assessment have not yet been fully investigated. For such cases ITU-T Recommendation P.1301 can only give general advices and relies on more research and future updates of this standard. For that purpose, later parts of this chapter will address in more detail three such aspects that go beyond Recommendation P.1301: (1) the issue that in a multiparty setting different interlocutors may have different individual perspectives from the same technical degradations (Section 6.3); (2) the possibility to optimize the test design by focusing on those individual perspectives of the interlocutors (Section 6.4); and (3) the issue of efficiently structuring the perceptual data for proper analysis, in particular accounting for those individual perspectives (Section 6.5).

## 6.2 ITU-T Recommendation P.1301

### 6.2.1 Overview of ITU-T Recommendation P.1301

ITU-T Recommendation P.1301 has been developed with significant contributions from the author<sup>2</sup> and it describes a methodology to conduct perceptual quality evaluation tests for multiparty telemeeting systems. It consists of a main body and eight normative annexes, as well as seven informational appendixes and a bibliography. Fur-

<sup>1</sup> ITU-T. *Recommendation P.1301 - Subjective quality evaluation of audio and audiovisual telemeetings*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2012

<sup>2</sup> Janto Skowronek et al. *Proposed main body and normative annex for draft recommendation P.AMT*. ITU-T Contribution COM 12 - C318 - E. Geneva, Switzerland: International Telecommunication Union, May 2012

Janto Skowronek et al. *Proposed non-normative appendixes for draft recommendation P.AMT*. ITU-T Contribution COM 12 - C319 - E. Geneva, Switzerland: International Telecommunication Union, May 2012

thermore, the recommendation is augmented by an extra document<sup>3</sup> providing example conversation test scenarios.

At the beginning of writing ITU-T Recommendation P.1301, the responsible work group inside the ITU-T, Study Group 12 - Question 10, faced both challenges and opportunities. In Skowronek et al.<sup>4</sup> we discussed that

... [o]ne challenging aspect was that there is a huge variety of telemeeting systems such as high-end telemeeting rooms, classical telephone conference bridges, low-cost Voice over IP solutions, web conferencing applications, mobile solutions, etc. Obviously, such a variety makes it difficult to write a standard assessment method valid for all cases. Another challenging aspect was that many aspects of multiparty quality were and are still unknown, even though a number of published and ITU-T internal studies on the quality assessment of multiparty were available to serve as input for the new recommendation. ... One encouraging aspect was that the new recommendation could potentially build on the comprehensive amount of quality assessment methods that were available either for one-to-one telecommunication scenarios or for the individual components of telemeeting systems.

Due to this mixture of challenges and opportunities, the philosophy of ITU-T Recommendation P.1301 is three-fold. First, the recommendation is based on a set of twelve existing standardized test methods published by the ITU (see Table 6.1), which on the one hand exploits the existing knowledge in the field and on the other hand allows to address the vast variety of telemeeting systems. Second, the recommendation provides guidance to the methods by means of decision criteria and flowcharts, and it provides information about how the existing methods should be modified – if necessary – and applied to multiparty scenarios. Third, the recommendation gives specific advices whenever sufficient knowledge about a certain aspect is available and limits to more general suggestions if a certain aspect requires further study.

Due to this philosophy, the application of ITU-T Recommendation P.1301 consists of two main steps: (1) select an appropriate existing test method as the basic test method, (2) modify and apply this basic test method to the multiparty case. Both steps are explained in more detail in the next two paragraphs.

<sup>3</sup> ITU-T. *P-Series Supplement P.Sup26 - Scenarios for the subjective quality evaluation of audio and audiovisual multiparty telemeetings*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2012

<sup>4</sup> Janto Skowronek, Katrin Schoenberger, and Alexander Raake. "Experience with and insights about the new ITU-T standard on quality assessment of conferencing systems". In: *Proceedings of the international conference of acoustics (AIA-DAGA 2013)*. Merano, Mar. 2013, pp. 444-447

---



---

<p>ITU-T. <i>Recommendation P.800 - Methods for objective and subjective assessment of quality</i>. International Standard. Geneva, Switzerland: International Telecommunication Union, 1996</p> <p>ITU-T. <i>Recommendation P.805 - Subjective evaluation of conversational quality</i>. International Standard. Geneva, Switzerland: International Telecommunication Union, 2007</p> <p>ITU-T. <i>Recommendation P.880 - Continuous evaluation of time-varying speech quality</i>. International Standard. Geneva, Switzerland: International Telecommunication Union, 2004</p> <p>ITU-T. <i>Recommendation P.910 - Subjective video quality assessment methods for multimedia applications</i>. International Standard. Geneva, Switzerland: International Telecommunication Union, 2008</p> <p>ITU-T. <i>Recommendation P.911 - Subjective audiovisual quality assessment methods for multimedia applications</i>. International Standard. Geneva, Switzerland: International Telecommunication Union, 1998</p> <p>ITU-T. <i>Recommendation P.920 - Interactive test methods for audiovisual communications</i>. International Standard. Geneva, Switzerland: International Telecommunication Union, 2000</p>	<p>ITU-R. <i>Recommendation BS.1116-1 - Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems</i>. International Standard. Geneva, Switzerland: International Telecommunication Union, 1997</p> <p>ITU-R. <i>Recommendation BS.1285 - Pre-selection methods for the subjective assessment of small impairments in audio systems</i>. International Standard. Geneva, Switzerland: International Telecommunication Union, 1997</p> <p>ITU-R. <i>Recommendation BS.1534-1 - Method for the subjective assessment of intermediate quality levels of coding systems</i>. International Standard. Geneva, Switzerland: International Telecommunication Union, 2003</p> <p>ITU-R. <i>Recommendation BT.500-13 - Methodology for the subjective assessment of the quality of television pictures</i>. International Standard. Geneva, Switzerland: International Telecommunication Union, 2012</p> <p>ITU-R. <i>Recommendation BT.710-4 - Subjective assessment methods for image quality in high-definition television</i>. International Standard. Geneva, Switzerland: International Telecommunication Union, 1998</p> <p>ITU-R. <i>Recommendation BT.1788 - Methodology for the subjective assessment of video quality in multimedia applications</i>. International Standard. Geneva, Switzerland: International Telecommunication Union, 2007</p>
---	---

---



---

Table 6.1: List of ITU Recommendations used as basic test methods for ITU-T Recommendation P.1301.

### 6.2.2 Selection of a Basic Test Method according to ITU-T Rec. P.1301

In order to select the most appropriate basic test method for the evaluation task at hand, ITU-T Recommendation P.1301 provides flow charts that guide to the corresponding standard documents. An important aspect here is that the flow charts utilize decision criteria since there are different methods available for different evaluation scenarios.

While the recommendation text lists and explains those criteria, the investigator still needs to make the corresponding decisions according to the evaluation task at hand. This requires from the investigator a certain amount of understanding of the actual underlying problem, which motivates the need to conduct such a perceptual quality evaluation. Furthermore, a certain amount of experience in conducting quality assessment tests is beneficial, since the nature of the problem – being of a technical, conceptual or even economic nature – needs to be aligned with the different characteristics of the different possible test methods. This reflects also the conclusions by Möller<sup>5</sup>, who wrote:

It is very difficult to give general guidance on which assessment method or methodology is adequate for a specific quality assessment problem. [...] The main conclusion drawn is that there is no general best assessment method or methodology independent of the test task. The optimum method has to be found for each specific task by an experienced test designer.

Once the investigator has made the corresponding decisions, following the flow charts in the recommendation is straight-forward. If appropriate methods do not exist for the actual evaluation task,

<sup>5</sup> Sebastian Möller. *Assessment and Prediction of Speech Quality in Telecommunications*. Kluwer Academic publishers, 2000

then the flowcharts will terminate with the statement that a dedicated method needs to be developed. If appropriate methods exist for the actual evaluation task, then the flowcharts will point to the corresponding ITU recommendations. Furthermore, the flowcharts will also point to the relevant annexes of ITU-T Recommendation P.1301, in order to provide information about how the selected methods may be modified and applied in a multiparty context.

### *6.2.3 Modification and Application of a Basic Test Method according to ITU-T Rec. P.1301*

Once the basic test method is chosen with the flowcharts of ITU-T Recommendation P.1301, the next step is to apply this method in a multiparty context, whereas modifications of the basic method might be necessary. For that purpose, the flowcharts also point to those annexes of ITU-T Recommendation P.1301 that are relevant for the evaluation task at hand.

Always among those referred annexes is “Annex A - Set up of a multiparty telemeeting assessment test”. This Annex A provides detailed information about test facilities, tasks and stimuli, experiment design, test subject profiles, rating scales, test instructions, training phases, data collection, and data analysis. Important here is that this Annex A addresses only the multiparty-specific aspects, while it refers to the corresponding basic test method for any non-multiparty specific aspects. This essentially means that the investigator needs to go through the details of the basic test method and compares them with the details in Annex A in order to identify which test details need to be modified and which details can directly be applied. Similarly, the investigator needs to repeat this for any annexes of ITU-T Recommendation P.1301 that the flowcharts are additionally referring to, which happens in case of specific evaluation scenarios such as, spatial audio, 3D video, web conferencing functionalities, asymmetric conditions, etc.

At this stage, the investigator should act with caution given that in a number of cases ITU-T Recommendation P.1301 can give only general suggestions instead of detailed advices, which is due to the fact that for many detailed aspects of multiparty testing the amount of available knowledge and experience is still limited. This means, ITU-T Recommendation P.1301 permits the investigator a certain degree of freedom – which can even vary between different evaluation scenarios – in realizing the actual test protocol. For that reason, ITU-T Recommendation P.1301 suggests at various places that the investigator should conduct pilot tests in order to verify the appropriateness of the test protocol, and should properly document the actual test procedure in order to avoid false comparisons with other studies and corresponding misinterpretations of results.

6.2.4 Test Aspects beyond ITU-T Recommendation P.1301

ITU-T Recommendation P.1301 provides guidance on the implementation of a test protocol as precisely as possible given the current state of the art. However, there are a number of aspects in multiparty telemeeting quality that are not addressed in more detail in ITU-T Recommendation P.1301.

One aspect is that a degradation can have different effects for the individual interlocutors. Thus the individual perspectives of interlocutors need to be taken into account. This is especially important in the case of asymmetric conditions, i.e. interlocutors are connected with different equipment or different network conditions. While ITU-T Recommendation P.1301 has an Annex F dedicated to asymmetric conditions, the advices given in that annex are of a rather general nature. In particular, no systematic approach is described that allows to systematically analyze how the occurrence of technical degradations relates to perceptual impairments in terms of the individual perspectives of the interlocutors. Furthermore, no specific approach is described in more detail on the test design, i.e. assignment of conditions among interlocutors, except the advice to aim for a design in which every test participant should ideally be rating every technical condition. And finally, no specific approach on the data analysis is given either, except the advice to ask for both an overall judgment, i.e. *Single-Perspective Telemeeting Quality*  $Q_i$  from Chapter 5, and for judgments about the individual connections between interlocutors, i.e. *Individual Connection Quality*  $Q_{ij}$ . For that reason, the Sections 6.3 to 6.5 will complement the test methodology by addressing each of those three aspects.

6.3 Analysis Method to translate Technical Degradations into Possible Perceptual Impairments

6.3.1 Motivation and Basic Principle

In multiparty telemeetings, the perceptual impact of technical degradations can be quite complex, since it is possible that a degradation has different effects for the individual interlocutors. Whether such differences of the perceived impairments occur, depends on a number of factors, such as the type and strength of the degradations, the system topology, the number of interlocutors, and the locations in the different end-to-end paths where the degradations occur.

To explain this, consider two examples for a three-party telemeeting, which are visualized in Figure 6.1.

In the first example, a technical degradation in the end device of interlocutor IL2 occurs in send-direction, for example packet loss on the uplink to the network. Thus all outgoing signal-paths for this one interlocutor are degraded. Following the signal paths shows that interlocutor's IL1 and IL3 receive one degraded signal from IL2, but no degradations from each other, while IL2 does not receive any

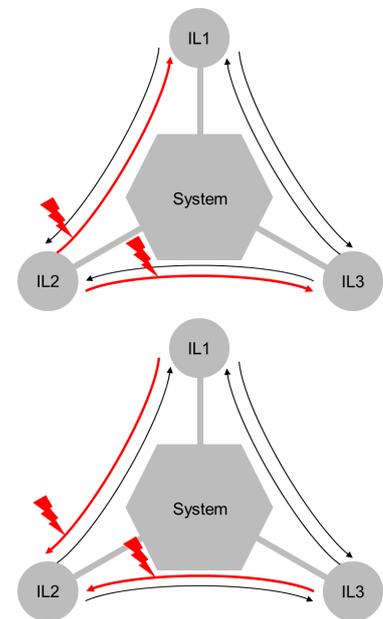


Figure 6.1: Examples of technical degradations in a three-party telemeeting. Top panel: Degradation occurs in the end device of interlocutor IL2 in send-direction, thus all outgoing signal-paths are degraded. Bottom panel: Degradation occurs in the end device of interlocutor IL2 in receive-direction, thus all ingoing signal-paths are degraded.

degraded signals at all. Concerning the possible perceptual impact for each interlocutor, this means, the perception of IL2 differs from the perceptions of IL1 and IL3: IL2 perceives no impairments at all; IL1 and IL3 perceive an impairment for one interlocutor and no impairment for the other interlocutor, which means IL1 and IL3 perceive an asymmetric impairment. In the second example, a technical degradation in the end device of interlocutor IL2 occurs in receive-direction, for example connection problems to external loudspeakers or an external video screen. Thus all incoming signal-paths for this one interlocutor are degraded. Following the signal paths shows that interlocutors IL1 and IL3 receive all signals without any degradations, and IL2 receives all signals with degradation. This means, the perception of IL2 differs again from the perceptions of IL1 and IL3: IL1 and IL3 perceive no impairments at all; IL2 perceives the same impairment for all interlocutors, which means IL2 perceives a symmetric impairment.

Apparently, a proper quality assessment method for multiparty telemeetings requires as a first step an analysis that translates the occurrence of technical degradations into the different possible perceptual impairments for each interlocutor. In order to facilitate such an analysis for a wide range of situations, it would be beneficial to have an approach available that is as systematic and as generic as possible. For that purpose, the following sections will present such a systematic method, which essentially performs an analysis of signal paths, in the form of a four-step approach<sup>6</sup>:

1. Description of the multiparty situation
2. Identification of *Degradation-Types* and *Degradation-Points*
3. Analysis of signal paths and deduction of possible perceptual impairments
4. Generation of a comprehensive representation for all interlocutors

### 6.3.2 Step 1: Description of the Multiparty Situation

The first step is to describe the multiparty situation in such a way that an appropriate signal path analysis is possible. In order to be as systematic and generic as possible, the proposal here is to use a graphical representation in a specific format that can easily be applied to different technical scenarios.

*Graphical representation* The previous section and its examples showed that the perceived impairments differ for degradations occurring on the send- or receive-side of an interlocutor. For that reason, the first requirement of the graphical representation is to facilitate a visually clear representation of the send- and receive-sides of the interlocutors, which, in addition, can easily be applied to an arbitrary number of interlocutors. This is achieved by replacing the more conventional illustration style shown in Figure 6.1 with a more signal-processing

<sup>6</sup> An earlier version of this approach has been presented in:

Janto Skowronek, Julian Herlinghaus, and Alexander Raake. "Quality Assessment of Asymmetric Multiparty Telephone Conferences: A Systematic Method from Technical Degradations to Perceived Impairments". In: *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*. International Speech Communication Association. Lyon, France, Aug. 2013, pp. 2604–2608

oriented style. Each interlocutor is shown twice, on the one side of the graph as sender, and on the other side as receiver.

Figure 6.2 shows this for a three- and four-interlocutor scenario. This representation can be read as a signal-flow diagram from left to right: Every interlocutor is in a certain environment (illustrated as dashed rectangle); he or she communicates via an end device (illustrated with a telephone handset), which sends the audio and video signals to the other interlocutors (illustrated as lines); on the receive side, every interlocutor is still in the same environment (dashed rectangle) using the same device (telephone handset).

*Extension with topologies* Since telemeeting systems can have different topologies, i.e. central bridge, peer-to-peer, or hybrid topologies, it would be beneficial for the further analysis to extend the graphical representation such that the topology is visualized as well. The approach chosen here is to use background shading to indicate which parts of the signal paths belong to the different main devices that constitute each topology (end devices, conference bridges, network links), while changing the signal paths as little as possible. Figure 6.3 shows this for a four-interlocutor scenario for the three topologies central-bridge, peer-to-peer, and hybrid. In the central-bridge topology (Figure 6.3, top panel), each end device (illustrated by a grey box around the telephone handset) sends the audio and video signals to one single point in the network, the central bridge (illustrated by a grey box around the mesh of signal paths), and each end device is receiving the signals from that single point in the network. In the peer-to-peer topology (Figure 6.3, middle panel), each end device as sender is responsible for duplicating the signals and sending them to all other end devices (illustrated by a grey box around the telephone handset and the lines until a duplication of the lines takes place). Similarly, each end device needs to mix all incoming signals together before the playback to the interlocutors (illustrated by a grey box including the telephone handset and the junction point of incoming lines). In the shown example of a hybrid topology (Figure 6.3, bottom panel), IL2 to IL4 are connected in a peer-to-peer topology, and the end device of IL2 acts as “central bridge” – or better say sub-bridge – for IL1, which is not peer-to-peer capable.

6.3.3 Step 2: Identification of Degradation-Types and Degradation-Points

The second step is to identify which kind of degradations, i.e. *Degradation-Types*, are occurring and to identify those points in the signal paths from sender to receiver in which the degradations are impacting the audio and video signals, i.e. *Degradation-Points*. The complexity of this identification step depends on the actual use case. In case of monitoring a system in real-life operation, a technical method is required that is capable to measure the occurrence of any such degradations at different points in the transmission chain from sender to receiver.

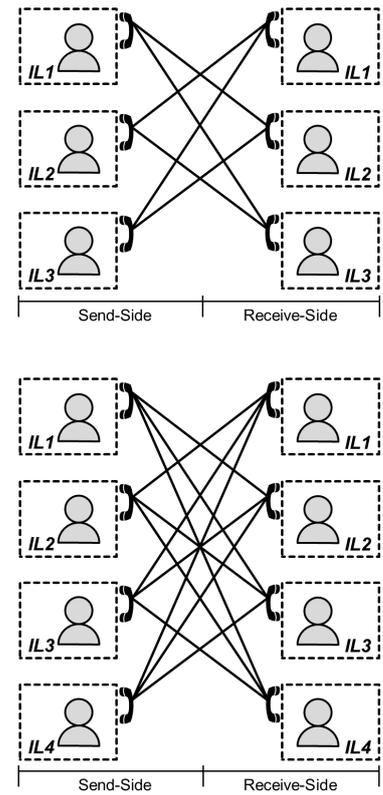


Figure 6.2: Principle of the graphical description of the multiparty situation: Each interlocutor is shown twice, as sender and as receiver. Top panel: Situation for three interlocutors. Bottom panel: Situation for four interlocutors.

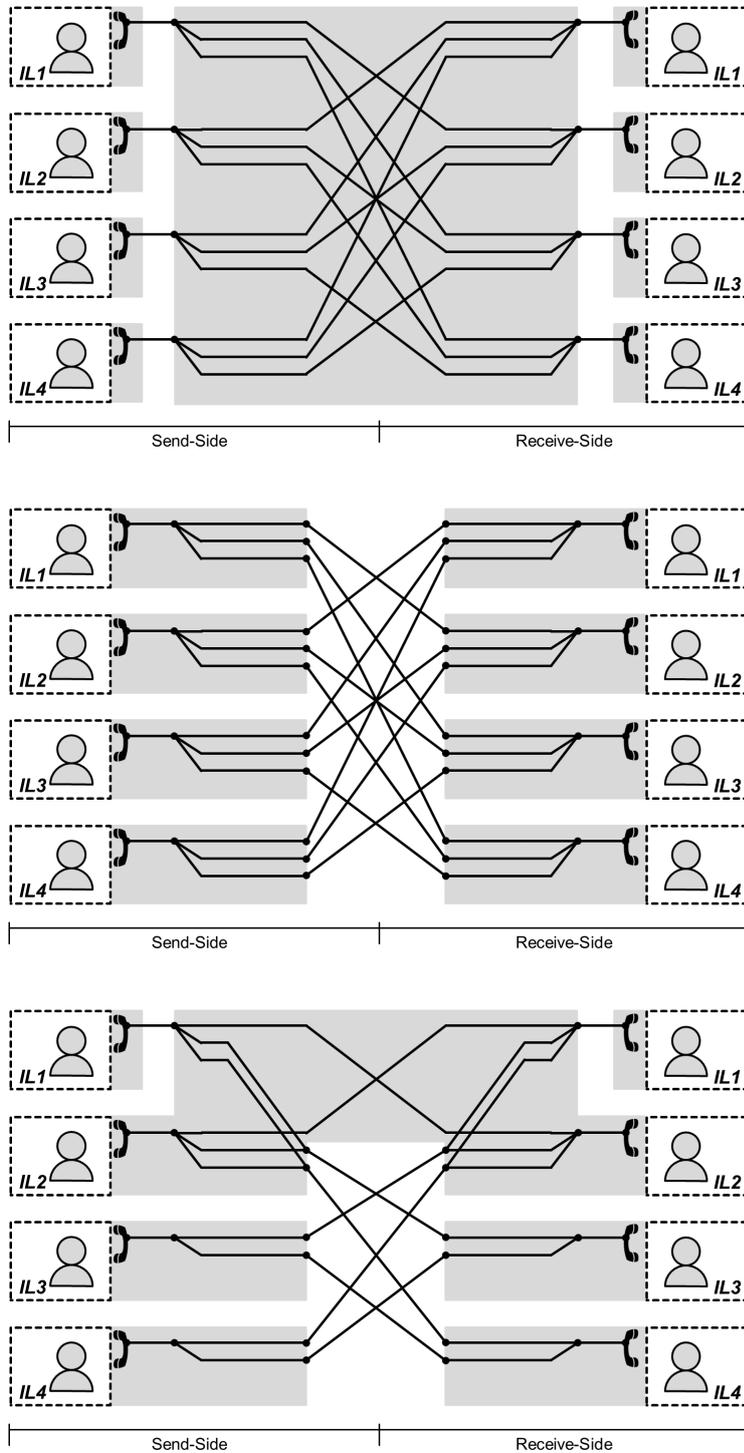


Figure 6.3: Representation of different telemeeting topologies using background shading to indicate, which parts of the signal paths are inside the main devices constituting the topology (end devices, conference bridges). Top panel: Central-bridge topology. Middle panel: Peer-to-peer topology. Bottom panel: Example of hybrid topology, a peer-to-peer topology between IL2, IL3 & IL4, with the end device of IL2 acting as “central bridge” for IL1, which is not peer-to-peer capable.

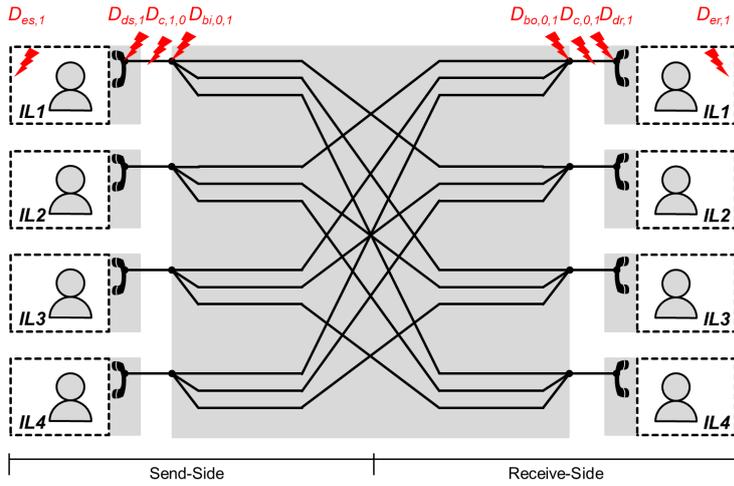
In case of planning a system or in case of a controlled quality assessment test, this step is simplified to selecting the *Degradation-Types* and deciding on the *Degradation-Points*.

*Definition of Degradation-Points* In order to be as generic as possible, the present approach defines the *Degradation-Points* on a more abstract level, independent from actual implementations. For example, in case of packet loss occurring in the network due to a certain router discarding packets due to traffic congestion, the approach does not consider that actual router as the *Degradation-Point*, but it considers the network connection as such as the *Degradation-Point*. While such an abstraction makes the analysis approach quite generic, a set of well-defined rules to construct such an abstraction improves the methodological soundness of the approach in terms of meaningfulness, reproducibility, and comparability of analysis results. For that reason, the present method uses the following well-defined set of *Degradation-Points* along the signal paths from sender to receiver, including a corresponding systematic notation scheme:

- $D_{es,x}$ : Degradation  $D$  in the environment  $e$  in send-side direction  $s$  of interlocutor  $x$  ( $x \in \mathbb{N}$ )
- $D_{ds,x}$ : Degradation  $D$  in the end device  $d$  in send-side direction  $s$  of interlocutor  $x$  ( $x \in \mathbb{N}$ )
- $D_{c,x,y}$ : Degradation  $D$  on the connection  $c$  from  $x$  to  $y$ , whereas  $x$  and  $y$  may refer to the interlocutors ( $x, y \in \mathbb{N}$ ) and to the central conference bridge ( $x, y \in [0]$ )
- $D_{bi,x,y}$ : Degradation  $D$  at the input  $bi$  of bridge  $x$ , with the signal coming from  $y$ , whereas  $x \in \mathbb{N}$  refers to bridges integrated inside the end devices of the interlocutors (peer-to-peer & hybrid topology),  $x \in [0]$  refers to the central bridge, and  $y \in \mathbb{N}_0$  may refer to the interlocutors or to the central bridge
- $D_{bo,x,y}$ : Degradation  $D$  at the output  $bo$  of bridge  $x$ , with the signal going to  $y$ , whereas  $x \in \mathbb{N}$  refers to bridges integrated inside the end devices of the interlocutors (peer-to-peer & hybrid topology),  $x \in [0]$  refers to the central bridge, and  $y \in \mathbb{N}_0$  may refer to the interlocutors or to the central bridge
- $D_{dr,x}$ : Degradation  $D$  in the end device  $d$  in receive-side direction  $r$  of interlocutor  $x$  ( $x \in \mathbb{N}$ )
- $D_{er,x}$ : Degradation  $D$  in the environment  $e$  in receive-side direction  $r$  of interlocutor  $x$  ( $x \in \mathbb{N}$ )

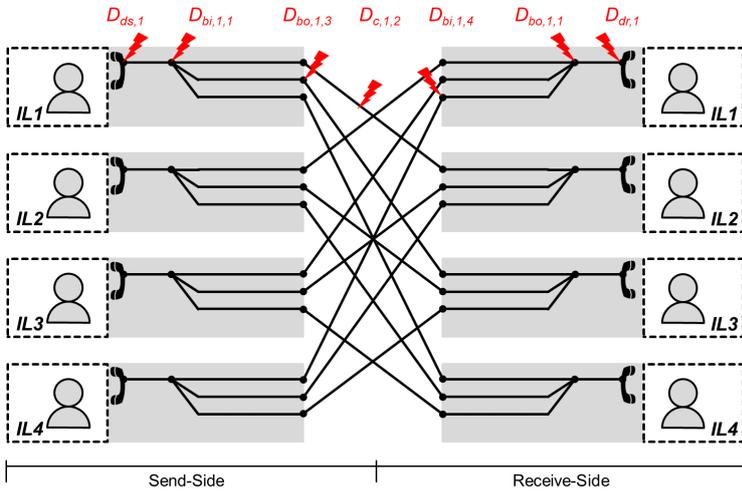
Figure 6.4 shows three example situations for the three topologies central-bridge, peer-to-peer and hybrid.

*Special cases: Acoustic echo, side-tone and camera-preview paths* There are three types of degradations that require an extension of the approach, as the signal paths for these degradations differ from the ones covered



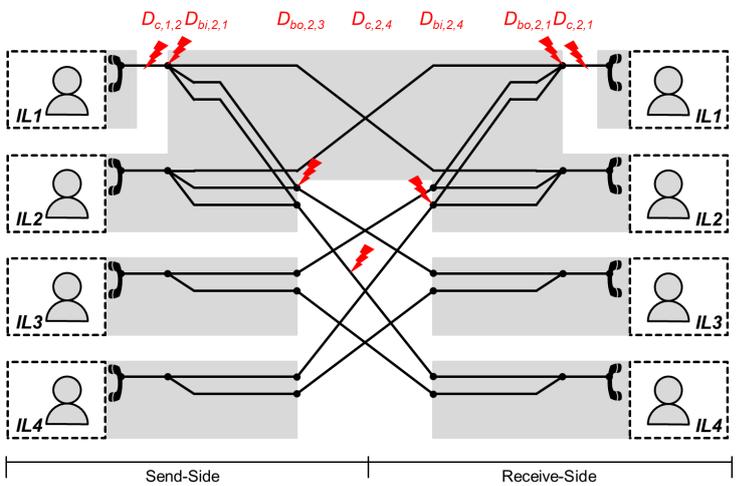
Examples of Degradation-Points for central-bridge topology:

- $D_{es,x} = D_{es,1}$ : send-side-relevant degradation (degr.) in the environment of IL1
- $D_{ds,x} = D_{ds,1}$ : degr. in the end device of IL1 in send-side direction
- $D_{c,x,y} = D_{c,1,0}$ : degr. on the connection from IL1 to the central bridge
- $D_{c,x,y} = D_{c,0,1}$ : degr. on the connection from the central bridge to IL1
- $D_{bi,x,y} = D_{bi,0,1}$ : degr. in the central bridge input, processing the stream of IL1
- $D_{bo,x,y} = D_{bo,0,1}$ : degr. in the central bridge output to IL1
- $D_{dr,x} = D_{dr,1}$ : degr. in the end device of IL1 in receive-side direction
- $D_{es,x} = D_{es,1}$ : receive-side-relevant degr. in the environment of IL1



Examples of Degradation-Points for peer-to-peer topology:

- $D_{ds,x} = D_{ds,1}$ : degr. in the end device of IL1 in send-side direction before the internal conference bridge
- $D_{bi,x,y} = D_{bi,1,1}$ : on the send-side of the integrated bridge of IL1, degr. in the bridge input processing the stream of IL1
- $D_{bo,x,y} = D_{bo,1,3}$ : on the send-side of the bridge of IL1, degr. in the bridge output to IL3
- $D_{c,x,y} = D_{c,1,2}$ : degr. on the connection from IL1 to IL2
- $D_{bi,x,y} = D_{bi,1,4}$ : on the receive-side of the bridge of IL1, degr. on the bridge input processing the stream from IL4
- $D_{bo,x,y} = D_{bo,1,1}$ : on the receive-side of the bridge of IL1, degr. on the bridge output to IL1
- $D_{dr,x} = D_{dr,1}$ : degr. in the end device of IL1 in receive-side direction after the internal bridge



Examples of Degradation-Points for hybrid topology:

- $D_{c,x,y} = D_{c,1,2}$ : degr. on the connection from IL1 to the integrated bridge of IL2
- $D_{bi,x,y} = D_{bi,2,1}$ : on the send-side of the bridge of IL2, degr. in the bridge input processing the stream of IL1
- $D_{bo,x,y} = D_{bo,2,3}$ : on the send-side of the bridge of IL2, degr. in the bridge output to IL3
- $D_{c,x,y} = D_{c,2,4}$ : degr. on the connection from IL2 to IL4
- $D_{bo,x,y} = D_{bo,2,1}$ : on the receive-side of the bridge of IL2, degr. on the bridge output to IL1
- $D_{c,x,y} = D_{c,2,1}$ : degr. on the connection from the integrated bridge of IL2 to IL1

Figure 6.4: Example visualization of Degradation-Points.

by the graphical representation so far: acoustic echo, distorted or non-optimal side-tone, and – for audiovisual communication – distorted camera-preview.

Acoustic echo is caused at the junction between end device and the environment of an interlocutor by a coupling of the loudspeaker audio signals into the microphone. In terms of the graphical representation used here, the audio signals at the receive-side of one interlocutor are coupled back into the send-side of that interlocutor, see Figure 6.5, top panel. Side-tone is a direct feedback of the microphone signal to the loudspeaker and is inserted into traditional telephone handsets to account for the fact that the handset is covering the ear and thus reducing the level of speech that a person is acoustically hearing from himself or herself. In terms of the graphical representation, the audio signal at the send-side of one interlocutor is coupled back into the receive-side of that interlocutor, see Figure 6.5, middle panel. Camera-preview is a feature often applied in audiovisual communication systems, in which the own picture taken by the camera is shown on the screen. In terms of the graphical representation, this case is essentially identical to the side-tone, since the video signal at the send-side is coupled back into the receive-side of that interlocutor.

This means, the graphical representation requires additional paths directly connecting the send- and receive-sides of the same interlocutor in both directions. This is realized in Figure 6.5, bottom panel, by the dotted line between send- and receive-side. Furthermore, the *Degradation-Points* for these two cases – echo on the one hand, and side-tone and camera-preview on the other hand – need to be defined as well. The decision is to define the *Degradation-Points* at those points in which the distorted signals would be added to the undistorted signals. In case of echo, this is the end device on the send-side; in case of side-tone and camera-preview, this is the end device on the receive-side; see Figure 6.5, bottom panel.

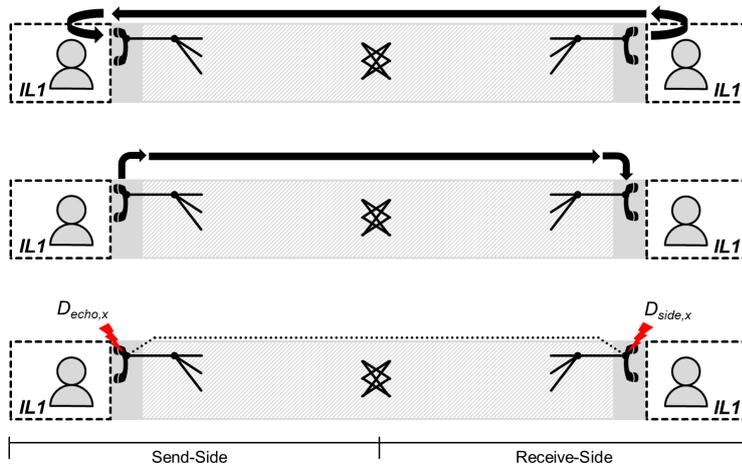


Figure 6.5: Special Degradation-Points for acoustic echo, side-tone and camera-preview. Top panel: The signal path for acoustic echo, which is a coupling from the receive-side to send-side of an interlocutor. Middle panel: The signal path for side-tone and camera-preview, which is a coupling from the send-side to the receive-side of an interlocutor. Bottom panel: Extension of graphical representation with a bidirectional coupling path and definition of Degradation-Points for Echo  $D_{echo,x}$  and Side-Tone & Camera-Preview  $D_{side,x}$ .

*Full set of Degradation-Points* The full set for the four-interlocutor scenario for the three topologies is shown in Figure 6.6.

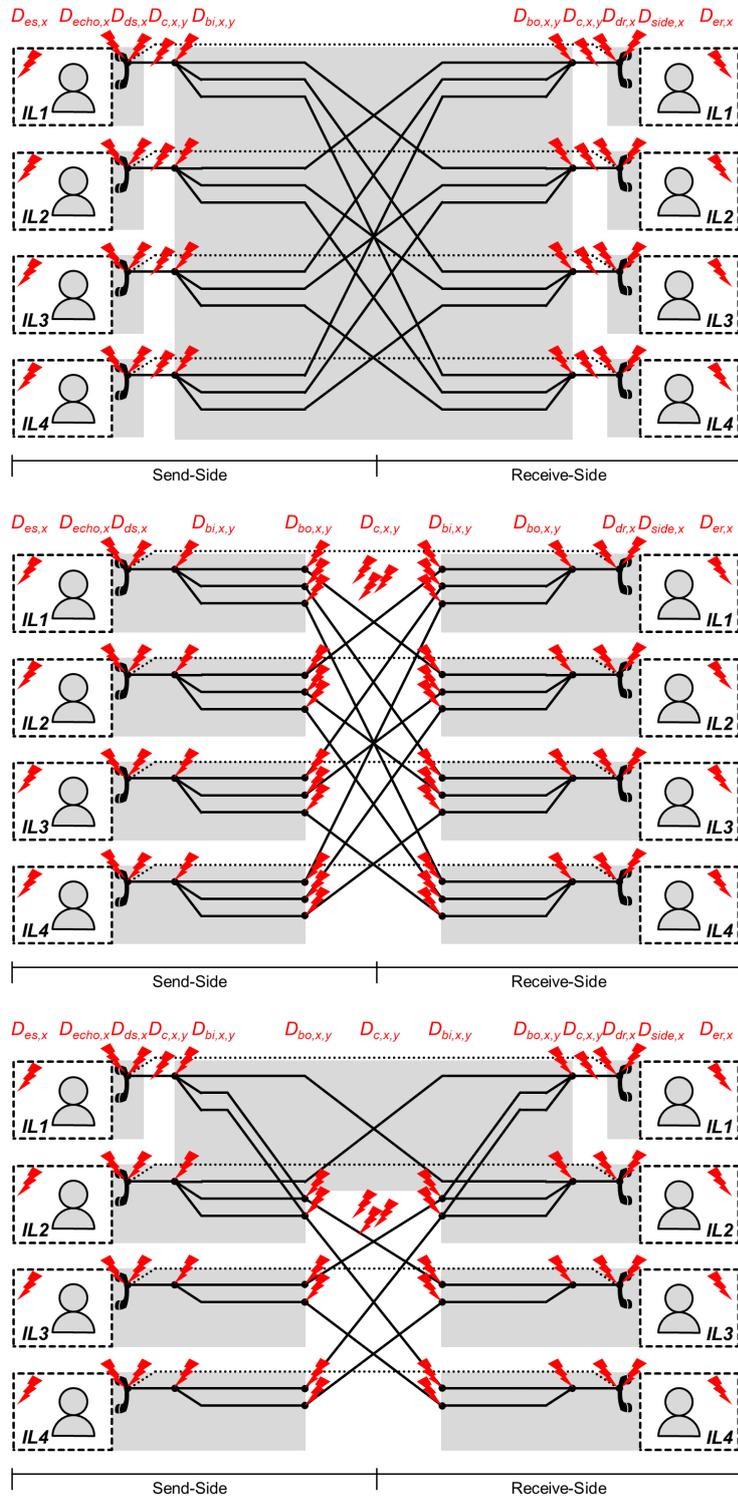


Figure 6.6: Full set of degradation points for the four-interlocutor scenario for the three topologies. Top panel: Central-bridge topology. Middle panel: Peer-to-peer topology. Bottom panel: Example of hybrid topology.

### 6.3.4 Step 3: Analysis of Signal Paths and Deduction of Possible Perceptual Impairments

The third step is to conduct the actual path analysis, whereas the procedure consists of two parts. The first part (Step 3.a) is to follow the signal paths from the send-side to the receive-side of each interlocutor and to identify any degradations on the individual signals. This is done by checking which *Degradation-Points* the signals are passing by and by employing technical knowledge to characterize how the degradation is affecting the signal. The second part (Step 3.b) is to determine for all combinations of sending and receiving interlocutors the possible perceptual impairments  $I_{x,y}$ , whereas  $I_{x,y}$  denotes that interlocutor  $x$  perceives at the receive-side an impairment of the audio or video signals coming from the send-side of interlocutor  $y$ . This is done by identifying which signals arriving at the receive side of each interlocutor are degraded or not (i.e. determine if  $I_{x,y} \neq 0$  or  $I_{x,y} = 0$ ), and – if possible – to employ technical and perceptual knowledge to characterize how those signal degradations are perceived (i.e. assign codes – short descriptors – to  $I_{x,y} \neq 0$  which allow to discriminate different degradations).

*Path analysis of individual degradations* Figures 6.7 to 6.9 provide a representative set of example path analyses that describe the translation of individual degradations  $D$  to the possible perceptual impairments  $I_{x,y}$  in more detail. For the purpose of explaining the method, the focus of the examples lies on the path analysis as such, that is the transmission of degradations from different *Degradation-Points* to the individual interlocutors. For that reason, the examples do not limit to certain actual *Degradation-Types* and therefore omit the aspects of how the effect of degradations on the signals can be characterized and how the identified impairments  $I_{x,y} \neq 0$  can be coded.

*Path analysis of multiple simultaneous degradations* The example path analyses in Figures 6.7 to 6.9 concern situations in which only one degradation occurs in the telemeeting. If multiple degradations occur simultaneously, the straight-forward approach is to conduct the analysis for each path between the interlocutors separately and then add the corresponding impairments at the receive-side. However, there are two aspects that require some caution.

The first, and more important, aspect is to consider that multiple degradations on the same signal path usually interact with each other, which in turn depends also on the order in which the degradations are affecting the signals. Consider as hypothetical example the combination of a coding artifact (some bandwidth limitation) and a packet loss artifact (some high-frequency distortions) on the path from IL1 to IL2. If the coding artifact occurs before the packet loss artifact, then the bandwidth limitation affects the original voice signal, but not the high-frequency distortions. If the coding artifact occurs after the packet loss artifact, then the bandwidth limitation also affects,

i.e. reduces, the high-frequency distortions. This means the strength of the high-frequency distortions is larger in one case than in the other. Concerning the analysis method, this means that while the analysis may be done for each individual path between the interlocutors, it may not be done separately for the individual degradations.

The second aspect refers to the possibility that non-linear signal processing is involved, which leads to non-linearities in the summation of individual signal streams. In such cases, the analysis of a summation of two signals, each with different signal degradations, may lead to different results than the separate analysis of each signal. However, to assess if this aspect is relevant requires an in-depth knowledge of the actual implementation of the system under consideration; no generally valid examples can be given here. Concerning the analysis method, this means that if such knowledge is available and demands an adjustment of the approach, then the analysis needs to be done for the relevant signal paths simultaneously. However, if such knowledge is not available, then the approach to conduct the analysis per individual path serves as an appropriate approximation of the reality.

#### 6.3.5 *Step 4: Generation of a Comprehensive Representation for all Interlocutors*

The fourth step is to generate a representation that summarizes the path analyses for all interlocutors. Ideally this representation should visualize the expected perceptual impairments for each considered technical degradation in a convenient format, such as, for instance, a table. Using an appropriate coding of the impairments and using Zero when no impairments are expected, such a table immediately visualizes the dependency of impairments and *Degradation-Points*. As examples, Figures 6.7 to 6.9 show such tables next to the corresponding path analyses.

#### 6.3.6 *Limitations*

The method presented here translates the occurrence of technical degradations into possible perceptual impairments, depending on the type and, predominantly, the location in which the degradations affect the signal paths between the interlocutors. The method's main limitation is inherently mentioned in the term "possible perceptual impairments": it determines whether impairments are theoretically possible at all, but it does not automatically provide the means to verify whether such impairments are actually perceived by the interlocutors. Such a verification can be done by either conducting corresponding perception tests or by applying instrumental models that predict in how far the degradations are actually perceived as impairments. The most appropriate place to insert this verification into the method is the path analysis in Step 3 in which technical and perceptual knowledge is already used to characterize how the signal







degradations arriving at the receive-side are – at least theoretically – perceived.

The study described in Skowronek et al.<sup>7</sup> is an example for conducting a corresponding perception test; the key facts are summarized in Table 6.2. This study revealed that such a verification is indeed necessary, as not all impairments were rated as the theoretical analysis suggested. In particular, the results for the tested background noise condition suggested that degradations need to be beyond a certain threshold in order to be perceived as quality-relevant impairments.

For that reason, possible first steps to realize such a verification is to develop for a sufficiently comprehensive set of degradations a full set of such perception thresholds in form of look-up tables. An alternative option is to develop dedicated instrumental models that apply corresponding thresholding operations.

<sup>7</sup> Janto Skowronek, Julian Herlinghaus, and Alexander Raake. "Quality Assessment of Asymmetric Multiparty Telephone Conferences: A Systematic Method from Technical Degradations to Perceived Impairments". In: *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*. International Speech Communication Association. Lyon, France, Aug. 2013, pp. 2604–2608

---



---

#### Study set up:

- Three-party conversation test, asking for quality ratings
- 47 test participants
- Three-party conversation scenarios, modified versions based on ITU-T P-supplement 26
  - Janto Skowronek. *3SCT - 3-party Short Conversation Test scenarios for conferencing assessment (Version 01)*. 2013. DOI: 10.5281/zenodo.16134
  - ITU-T. *P-Series Supplement P.Sup26 - Scenarios for the subjective quality evaluation of audio and audiovisual multiparty telemeetings*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2012
- Rating scale: 5-point Absolute Category Rating scale according to P.800
  - ITU-T. *Recommendation P.800 - Methods for objective and subjective assessment of quality*. International Standard. Geneva, Switzerland: International Telecommunication Union, 1996
- Test system: VoIP telephones, central-bridge
- Tested conditions: unimpaired (G.711, 150 ms one-way delay), loudness loss (20 dB SPL), background noise (65 dB SPL), echo (echo attenuation 10 dB SPL), packet loss (10% random loss)
- Further details: see Section 8.3, Experiment ACT<sub>1</sub>

---



---

#### Summary of results:

- Quotation: "In summary, the degradations were often judged in the way the structured technical analysis suggested, but some detailed findings were contrary to these expectations."
  - Examples of such detailed findings:
    - Loudness loss on another connection affected own connection rating.
    - Ratings for listener echo (hearing voice of one interlocutor twice) are not affected in the presence of talker echo (hearing echo of own voice).
    - In case of packet loss, also the unimpaired connection was affected, that is, was judged worse.
  - Quotation: "Apparently, some degradations have to be above a certain threshold, e.g. the level of transmitted background noise, before they are included in quality ratings."
- 
- 

Table 6.2: Key facts of the study described in

Janto Skowronek, Julian Herlinghaus, and Alexander Raake. "Quality Assessment of Asymmetric Multiparty Telephone Conferences: A Systematic Method from Technical Degradations to Perceived Impairments". In: *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*. International Speech Communication Association. Lyon, France, Aug. 2013, pp. 2604–2608

## 6.4 *Experimental Design Method distributing Possible Perceptual Impairments instead of Technical Degradations*

### 6.4.1 *Motivation*

Usually investigators choose first a set of conditions and then a design that ensures that all conditions are distributed across test calls, using one of various well-known design techniques (examples: full factorial design, balanced incomplete block design, within-subject design and across-subject design). This means, investigators assume in this process a technical or expert perspective: they define first the conditions in terms of the technical system properties and then distribute them across test calls according to the chosen design; in other words they ensure that the technical system properties are presented according to the chosen design.

When it comes to multiparty quality tests, however, this results in a non-optimal design because a certain technical condition can lead to the situation that at least one test participant does not perceive any impairment at all. Referring to the path analyses in Figures 6.7 to 6.9 in the previous section, consider the two following examples: First, if in one test call, one participant has significant packet loss in send direction, then the other interlocutors will perceive signal distortions for this one participant and no impairments for all other interlocutors. However, that one participant will perceive all interlocutors without any impairment. Second, if in another test call, one participant has significant packet loss in receive-side direction, then that one participant will perceive signal distortions from all interlocutors. However, the other interlocutors will perceive no impairments at all.

### 6.4.2 *Application*

To overcome such design inefficiencies, the novel design paradigm is not to distribute the technical conditions according to the design, e.g. have  $x$  calls with send side packet loss and  $y$  calls with receive side packet loss, but to distribute the perceptions according to the design, e.g. have  $x$  cases in which participants perceive packet loss distortions for one interlocutor and  $y$  cases in which participants perceive packet loss distortions for all interlocutors. This new paradigm means to shift the investigator's perspective from the technical towards the perceptual point of view.

The advantage is that this new perspective allows to come up with smart combinations of technical properties for individual participants in order to avoid unnecessary calls for participants without any impairment. An example showing the difference between these two design paradigms is visualized in Table 6.3. While the given example suggests a large room for optimizations in the order of 50% (three instead of six test calls), the actual amount of reduction of required test calls is usually lower and depends on two aspects: a) if the set of target impairments actually allows for combinations of different impairments into a single call without mutually influencing each

other, and b) if the actual test system is technically allowing to realize such combinations.

a) The conventional design paradigm, i.e. design technical degradation and obtain perceptual impairments, requires six test calls (table rows 1 to 6) to expose all interlocutors to each impairment. The disadvantage is that in every call, at least one interlocutor does not perceive any impairments.

Technical Degradations $D_x$ for Interlocutor $x$				Perceptual Impairments $I_{x,y}$ , from the perspective of Interlocutor $x$ with regard to Interlocutor $y$								
$D_1$	$D_2$	$D_3$		$I_{1,1}$	$I_{1,2}$	$I_{1,3}$	$I_{2,1}$	$I_{2,2}$	$I_{2,3}$	$I_{3,1}$	$I_{3,2}$	$I_{3,3}$
$D_{echo,1}$	0	0	$\Rightarrow$	0	0	0	0	te1	le1	0	le1	te1
0	$D_{echo,2}$	0		te1	0	le1	0	0	0	le1	0	te1
0	0	$D_{echo,3}$		te1	le1	0	le1	te1	0	0	0	0
$D_{dr,1}$	0	0		0	dev2	dev2	0	0	0	0	0	0
0	$D_{dr,2}$	0		0	0	0	dev2	0	dev2	0	0	0
0	0	$D_{dr,3}$		0	0	0	0	0	0	dev2	dev2	0

b) Applying the inverted design paradigm, i.e. design perceptual impairments and obtain technical degradations, allows to merge test cases such that the same impairments are covered with only three test calls (table rows 1 to 3).

Technical Degradations $D_x$ for Interlocutor $x$				Perceptual Impairments $I_{x,y}$ , from the perspective of Interlocutor $x$ with regard to Interlocutor $y$								
$D_1$	$D_2$	$D_3$		$I_{1,1}$	$I_{1,2}$	$I_{1,3}$	$I_{2,1}$	$I_{2,2}$	$I_{2,3}$	$I_{3,1}$	$I_{3,2}$	$I_{3,3}$
$D_{echo,1} + D_{dr,1}$	0	0	$\Leftarrow$	0	dev2	dev2	0	te1	le1	0	le1	te1
0	$D_{echo,2} + D_{dr,2}$	0		te1	0	le1	dev2	0	dev2	le1	0	te1
0	0	$D_{echo,3} + D_{dr,3}$		te1	le1	0	le1	te1	0	dev2	dev2	0

c) Explanation of table entries following the path analysis according to Section 6.3:  
 $D_{echo,x}$  inserted at Interlocutor  $x$  means that Interlocutor  $x$  does not perceive any impairments, while the other two interlocutors perceived talker echo (te1), i.e. echo of own voice, and listener echo (le1), i.e. echo of the other interlocutor's voice.  
 $D_{dr,x}$  inserted at Interlocutor  $x$  means that Interlocutor  $x$  perceives impairments of all received signals (dev2), i.e. of the other two interlocutors, while those interlocutors do not perceive any impairments.

Table 6.3: Examples of applying a conventional design paradigm and the novel design paradigm for a three interlocutor setting for two technical degradations: Echo  $D_{echo,x}$ , and Terminal Device Degradation at receive-side  $D_{dr,x}$ .

## 6.5 Method to organize Perceptual Data to account for the Individual Perspectives of Interlocutors

### 6.5.1 Motivation and Basic Principle

Since telemeeting participants can have different perceptions of the same call, not only the experimental design but also the data analysis needs to properly take these individual perspectives into account. The task is to organize the perceptual data according to the test conditions such that a statistical analysis is reasonable. Simply taking the ratings from all participants per test call and then grouping all test calls with the same technical degradation is not reasonable, as such sets could mix ratings referring to different perceptions of the same call. Instead, the quality ratings need to be grouped into individual sets sharing the same possible perceptual impairments.

The basic principle of such impairment-based grouping consists of two steps:

1. Starting from the actual experiment design, the sets are defined by the respective combinations of impairments on the individual connections. Those set definitions can be expressed in terms of

“Set  $x$  is defined by the fact that one connection has Impairment  $Ia$ , one connection Impairment  $Ib$ , and so on”. However, the actual definition of sets depends on the quality aggregation level (see Chapter 5), i.e. if the quality ratings refer to the *Individual Connection Quality* scores  $Q_{ij}$  or to the *Single-Perspective Telemeeting Quality*  $Q_i$ . For grouping the ratings of  $Q_i$ , each set of ratings is defined by the specific combination of impairments on the individual connections, i.e. how many individual connections have which impairments. For grouping the ratings of  $Q_{ij}$ , each set of ratings is defined by the specific impairment on that individual connection under the constraint of a specific combination of impairments on the other individual connections.

2. Then the ratings (per test participant and test call) can be assigned to the groups by applying selection criteria that match the set definitions. Such assignment can be expressed in terms of “Assign rating  $R$  to set  $x$ , if in the call the test participant had on one connection Impairment  $Ia$ , on one connection Impairment  $Ib$ , and so on”.

An example to explain the basic principle for ratings of  $Q_i$  and  $Q_{ij}$  is given in Table 6.4.

Next to enable a proper statistical analysis, it will be beneficial to extend the data organization to facilitate the modeling approaches that will be described in Chapter 9. The main goal of those modeling approaches is to estimate the *Single-Perspective Telemeeting Quality*  $Q_i$  based on the *Individual Connection Quality*  $Q_{ij}$ . For that reason the data organization needs to be extended such that when assigning the data into the different sets, the link between ratings of  $Q_i$  and  $Q_{ij}$  are maintained. This is done by forming additional groups that are defined in the same way as the groups for the  $Q_i$  ratings, but that contain the  $Q_i$  ratings and the corresponding  $Q_{ij}$  ratings of the same test participant for the same test call.

1. Example test design (one call per table row) for a three-interlocutor setting for two technical degradations: Echo  $D_{echo,i}$ , leading to the perceived impairments “te1” (talker echo) and “le1” (listener echo); and Terminal Device Degradation at receive-side  $D_{dr,i}$ , leading to the perceived impairment “dev2”. In case of evaluating the *Single-Perspective Telemeeting Quality*  $Q_i$ , one rating  $R_i$  per call per participant is collected. In case of evaluating the *Individual Connection Quality*  $Q_{ij}$ , one rating  $R_{ij}$  per connection and thus three ratings per call per participant are collected.

Technical Degradations $D_i$ for Interlocutor $i$			Perceptual Impairments $I_{i,j}$ , from the perspective of Interlocutor $i$ with regard to Interlocutor $j$								
$D_1$	$D_2$	$D_3$	$I_{1,1}$	$I_{1,2}$	$I_{1,3}$	$I_{2,1}$	$I_{2,2}$	$I_{2,3}$	$I_{3,1}$	$I_{3,2}$	$I_{3,3}$
$D_{echo,1} + D_{dr,1}$	o	o	o	dev2	dev2	o	te1	le1	o	le1	te1
o	$D_{echo,2} + D_{dr,2}$	o	te1	o	le1	dev2	o	dev2	le1	o	te1
o	o	$D_{echo,3} + D_{dr,3}$	te1	le1	o	le1	te1	o	dev2	dev2	o

2. Grouping principle for  $Q_i$ : each set of ratings is defined by the specific combination of impairments on the individual connections. For the given example this leads to the following two sets.

Set	Selection Criterion		Interpretation
1	$R_i   I_{i,a}=0 \wedge I_{i,b}=dev2 \wedge I_{i,c}=dev2$	with $a = i$ : own connection, $[b, c] \neq i$ connections to interlocutors	all ratings in which an interlocutor perceived no impairment of the own connection and the terminal-related impairment for the two interlocutors
2	$R_i   I_{i,a}=te1 \wedge I_{i,b}=0 \wedge I_{i,c}=le1$	with $a = i$ : own connection, $[b, c] \neq i$ connections to interlocutors	all ratings in which an interlocutor perceives the combination of talker echo, listener echo for one interlocutor and no echo for the other interlocutor

3. Grouping principle for  $Q_{ij}$ : each set of ratings is defined by the specific impairment of the currently considered connection under the constraint of a specific combination of impairments on the other individual connections. For the given example this leads to the following five sets.

Set	Selection Criterion		Interpretation
1a	$R_{ia}   I_{i,a}=0 \wedge I_{i,b}=dev2 \wedge I_{i,c}=dev2$	with $a = i$ : currently considered own connection, $[b, c] \neq i$ connections to interlocutors	all ratings of the own unimpaired connection, under the constraints that the other two connections have the impairment dev2
1b	$R_{ia}   I_{i,a}=dev2 \wedge I_{i,b}=0 \wedge I_{i,c}=dev2$	with $a \neq i$ : connection of currently considered interlocutor, $b = i$ : own connection, $c \neq i$ connection to remaining interlocutor	all ratings of a connection with the impairment dev2, under the constraints that the own connection has no impairment and the other connection has the impairment dev2
2a	$R_{ia}   I_{i,a}=te1 \wedge I_{i,b}=0 \wedge I_{i,c}=le1$	with $a = i$ : currently considered own connection, $[b, c] \neq i$ connections to interlocutors	all ratings of the own connection with talker echo te1, under the constraints that one connection has no impairment and the other connection has listener echo le1
2b	$R_{ia}   I_{i,a}=0 \wedge I_{i,b}=te1 \wedge I_{i,c}=le1$	with $a \neq i$ : connection of currently considered interlocutor, $b = i$ : own connection, $c \neq i$ connection to remaining interlocutor	all ratings of a connection with no impairment, under the constraints that the own connection has talker echo te1 and the other connection has listener echo le1
2c	$R_{ia}   I_{i,a}=le1 \wedge I_{i,b}=te1 \wedge I_{i,c}=0$	with $a \neq i$ : connection of currently considered interlocutor, $b = i$ : own connection, $c \neq i$ connection to remaining interlocutor	all ratings of a connection with listener echo le1, under the constraints that the own connection has talker echo te1 and the other connection has no impairment

4. This procedure results in the following assignments of the ratings into the different sets (each table row representing one set).

Sets of Telemeeting Quality Ratings $R_i$ for Interlocutor $i$			Sets of Individual Connection Quality Ratings $R_{i,j}$ , from the perspective of Interlocutor $i$ with regard to Interlocutor $j$								
$R_1$	$R_2$	$R_3$	$R_{1,1}$	$R_{1,2}$	$R_{1,3}$	$R_{2,1}$	$R_{2,2}$	$R_{2,3}$	$R_{3,1}$	$R_{3,2}$	$R_{3,3}$
1	2	2	1a	1b	1b	2b	2a	2c	2b	2c	2a
2	1	2	2a	2b	2c	1b	1a	1b	2c	2b	2a
2	2	1	2a	2c	2b	2c	2a	2b	1b	1b	1a

Table 6.4: Example for the basic principle to organize the perceptual data: starting from the experimental design, the sets of ratings are defined by the combinations of the impairments on the individual connections.

### 6.5.2 Application by means of an Algorithm

At first glance the two steps, definition of sets and assignment of data, appear to be straight-forward. However, going through the example in Table 6.4 suggests that conducting the data organization manually can become quite laborious and errorprone. For that reason, the next paragraphs describe an algorithm that automatically performs the data organization, whereas the presentation of the algorithm aims to be as generic as possible by means of an implementation-independent description.

*Specification of task, input and output* To enable a statistical analysis, the first and second tasks of the algorithm are to automatically conduct the two steps explained in the basic principle: the definition of sets and the corresponding assignment of the ratings for  $Q_i$  and  $Q_{ij}$ . To enable a modeling of  $Q_i$  as function of  $Q_{ij}$ , the third task of the algorithm is to maintain during the grouping process the link between ratings of  $Q_i$  and  $Q_{ij}$  per test participant and test call.

The input will be a data set that contains the information in which test calls which interlocutors rated which combination of impairments on individual connections with which ratings for  $Q_i$  and  $Q_{ij}$ . While the actual data format may be defined in another way, the proposal here is to use as input a data matrix with one row per test call. The format of that input data matrix and a corresponding example are given in Table 6.5.

The output will be a data set that contains the information which ratings of  $Q_i$  and  $Q_{ij}$  belong to which of the defined sets. While the actual data format may be defined in another way, the proposal here is to generate four data matrices: one with ratings of  $Q_i$ , one with ratings of the own connection  $Q_{ij}$  with  $i = j$ , one with ratings of the connections to the interlocutors  $Q_{ij}$  with  $i \neq j$ , and one with the linked ratings of  $Q_i$  and  $Q_{ij}$  used for modeling. The format of each output data matrix and corresponding examples are given in Table 6.6.

*Challenges for automatic definition of impairment sets* Concerning the automatic definition of the impairment sets, the algorithm needs to address three aspects that complicate the implementation, which will be explained in the following: treatment of own connections, sorting of connections to interlocutors, and identification of permutations.

Concerning own connections, the algorithm should apply a special treatment of those connections. This is motivated by the type of information, i.e. the type of quality features, that is available for a test subject to give a judgment about the own connection. It can be argued that ratings of the own connection are of a more hypothetical nature, as they rely in most cases only on cue-based quality features, compared to ratings of the connections of the other interlocutors, which rely in most cases on signal-based features, see the discussions in Section 5.3. For that reason, it could be beneficial for modeling

Row definition:		all data of one test call			
per row					
Column definition:					
Column number			Example for #IL = 5	Input variable	Description of input variable
1			1	<i>CallID</i>	Identifier for test call
1 + 1	to	$1 + 1 \cdot \#IL$	2 to 6	$I_{1j}$	Impairment identifiers from the perspective of Interlocutor 1 with regard to Interlocutor $j$ , with $j \in [1 \dots \#IL]$
1 + 1 + 1 · #IL	to	$1 + 2 \cdot \#IL$	7 to 11	$I_{2j}$	Impairment identifiers from the perspective of Interlocutor 2 with regard to Interlocutor $j$ , with $j \in [1 \dots \#IL]$
1 + 1 + 2 · #IL	to	$1 + 3 \cdot \#IL$	12 to 16	$I_{3j}$	Impairment identifiers from the perspective of Interlocutor 3 with regard to Interlocutor $j$ , with $j \in [1 \dots \#IL]$
⋮	⋮	⋮	17 to 21	⋮	Corresponding impairment identifiers from the perspectives of Interlocutors 4...#IL - 1
$1 + 1 + (\#IL - 1) \cdot \#IL$	to	$1 + \#IL \cdot \#IL$	22 to 26	$I_{\#IL}$	Impairment identifiers from the perspective of Interlocutor #IL with regard to Interlocutor $j$ , with $j \in [1 \dots \#IL]$
$1 + \#IL^2 + 1$	to	$1 + \#IL^2 + \#IL$	27 to 31	$R_i$	Ratings of $Q_i$ of Interlocutors $i$ , with $i \in [1 \dots \#IL]$
$1 + \#IL^2 + \#IL + 1 + 0 \cdot \#IL$	to	$1 + \#IL^2 + \#IL + 1 \cdot \#IL$	32 to 36	$I_{1j}$	Ratings of $Q_{ij}$ from the perspective of Interlocutor 1 with regard to Interlocutor $j$ , with $j \in [1 \dots \#IL]$
$1 + \#IL^2 + \#IL + 1 + 1 \cdot \#IL$	to	$1 + \#IL^2 + \#IL + 2 \cdot \#IL$	37 to 41	$I_{2j}$	Ratings of $Q_{ij}$ from the perspective of Interlocutor 2 with regard to Interlocutor $j$ , with $j \in [1 \dots \#IL]$
$1 + \#IL^2 + \#IL + 1 + 2 \cdot \#IL$	to	$1 + \#IL^2 + \#IL + 3 \cdot \#IL$	42 to 46	$I_{3j}$	Ratings of $Q_{ij}$ from the perspective of Interlocutor 3 with regard to Interlocutor $j$ , with $j \in [1 \dots \#IL]$
⋮	⋮	⋮	47 to 51	⋮	Corresponding ratings of $Q_{ij}$ of from the perspectives of Interlocutors 4...#IL - 1
$(1) + 1 + (\#IL - 1) \cdot \#IL$	to	$1 + \#IL^2 + \#IL + \#IL \cdot \#IL$	52 to 56	$I_{\#IL}$	Ratings of $Q_{ij}$ from the perspective of Interlocutor #IL with regard to Interlocutor $j$ , with $j \in [1 \dots \#IL]$

<i>CallID</i>	$I_{11}$	$I_{12}$	$I_{13}$	$I_{21}$	$I_{22}$	$I_{23}$	$I_{31}$	$I_{32}$	$I_{33}$	$R_1R_2R_3$	$R_{11}R_{12}R_{13}$	$R_{21}R_{22}R_{23}$	$R_{31}R_{32}R_{33}$								
1	o	dev2	dev2	o	te1	le1	o	le1	te1	4	2	2	5	3	3	5	1	2	5	2	1
2	te1	o	le1	dev2	o	dev2	le1	o	te1	2	4	2	1	5	5	3	5	3	2	5	1
3	te1	le1	o	le1	te1	o	dev2	dev2	o	2	2	4	1	2	2	2	1	5	3	3	5

Table 6.5: Input specification for the data organization algorithm. Top table: definition of the input format for #IL interlocutors. Bottom table: corresponding data example for #IL = 3.

Output data matrix for $Q_i$ :						Output data matrix for $Q_{ii}$ (own connection):							
$setID$	set definition using $I_{ij}$			$CallID$	$i =$ $SubjID$	$R_i$	$setID$	set definition using $I_{ij}$			$CallID$	$i =$ $SubjID$	$R_{ii}$
1	o	dev2	dev2	1	1	4	1a	o	dev2	dev2	1	1	5
1	o	dev2	dev2	2	2	4	1a	o	dev2	dev2	2	2	5
1	o	dev2	dev2	3	3	4	1a	o	dev2	dev2	3	3	5
2	te1	o	le1	1	2	2	2a	te1	o	le1	1	2	1
2	te1	o	le1	1	3	2	2a	te1	o	le1	1	3	1
2	te1	o	le1	2	1	2	2a	te1	o	le1	2	1	1
2	te1	o	le1	2	3	2	2a	te1	o	le1	2	3	1
2	te1	o	le1	3	1	2	2a	te1	o	le1	3	1	1
2	te1	o	le1	3	2	2	2a	te1	o	le1	3	2	1

Output data matrix for $Q_{ij}$ (excluding own connection):							
$setID$	set definition using $I_{ij}$			$CallID$	$i =$ $SubjID$	$j \neq i$	$R_{ij}$
1b	dev2	o	dev2	1	1	2	3
1b	dev2	o	dev2	1	1	3	3
1b	dev2	o	dev2	2	2	1	3
1b	dev2	o	dev2	2	2	3	3
1b	dev2	o	dev2	3	3	1	3
1b	dev2	o	dev2	3	3	2	3
2b	o	te1	le1	1	2	1	5
2b	o	te1	le1	1	3	1	5
2b	o	te1	le1	2	1	2	5
2b	o	te1	le1	2	3	2	5
2b	o	te1	le1	3	1	3	5
2b	o	te1	le1	3	2	3	5
2c	le1	te1	o	1	2	3	2
2c	le1	te1	o	1	3	2	2
2c	le1	te1	o	2	1	3	2
2c	le1	te1	o	2	3	1	2
2c	le1	te1	o	3	1	2	2
2c	le1	te1	o	3	2	1	2

Output data matrix for modeling, linking $Q_i$ and $Q_{ij}$ :									
$setID$	set definition using $I_{ij}$			$CallID$	$i =$ $SubjID$	$R_i$	$R_{ii}$	$R_{ij}$ with sorting based on the values of $I_{ij}$ with $j \in [1 \dots i - 1, i + 1 \dots \#IL]$	
1	o	dev2	dev2	1	1	4	5	3	3
1	o	dev2	dev2	2	2	4	5	3	3
1	o	dev2	dev2	3	3	4	5	3	3
2	te1	o	le1	1	2	2	1	5	2
2	te1	o	le1	1	3	2	1	5	2
2	te1	o	le1	2	1	2	1	5	2
2	te1	o	le1	2	3	2	1	5	2
2	te1	o	le1	3	1	2	1	5	2
2	te1	o	le1	3	2	2	1	5	2

Table 6.6: Output specification of the data organization algorithm. Each table represents one of four data matrices that are generated; the first three matrices ( $Q_i$ ,  $Q_{ii}$ ,  $Q_{ij}$ ) facilitate statistical analysis; the fourth matrix is used for modeling. Table headings: definition of the output format for  $\#IL$  interlocutors. Table bodies: corresponding data examples for  $\#IL = 3$ .

purposes, if the ratings of the own connections are treated differently, meaning that the data format should support an easy addressing of those ratings. This is realized here by fixed positions of the own connections in the definitions of the impairment sets: In the definition of impairment sets of  $Q_i$  (ratings of whole telemeeting) and  $Q_{ii}$  (ratings of own individual connection into the telemeeting), the own connection is always put at the first position; in the definition of impairment sets of  $Q_{ij}$  (ratings of individual connection of the other interlocutors into the telemeeting), the own connection is always put at the second position. This can be found in the selection criteria used in Table 6.4 as well as in the definition of the output data format in Table 6.6.

Concerning the connections to the interlocutors, the definition of impairment sets is ill-defined, if no countermeasures are taken by the algorithm. A telemeeting with  $\#IL$  interlocutors means that each interlocutor  $i$  is rating the own connection ( $I_{ii}$ ) plus  $\#IL - 1$  connections of the other interlocutors ( $I_{ij}$  with  $j \in [1 \dots i - 1, i + 1 \dots \#IL]$ ). Thus there are  $(\#IL - 1)! = 1 \cdot 2 \cdot \dots \cdot (\#IL - 1)$  possibilities to place those connections in the definition of such an impairment set, given that the own connection  $I_{ii}$  shall have a fixed position. In order to obtain just one unique definition for each set, the algorithm needs to choose one such combination by means of sorting the impairments. For that purpose, the sorting could follow the alphabetical order if the individual impairments are defined by strings, or the sorting could be based on ranking numerical values if the impairments are defined by such values. While this is straight-forward to implement for the output data sets for the statistical analysis, the generation of the output data for the modeling, i.e. estimating  $Q_i$  based on  $Q_{ij}$ , needs to specifically address the cases when multiple connections have the same impairment. In such a case, there are multiple possibilities for placing those individual connections into the columns of the output matrix, while they still fulfill the alphabetical or numerical ordering of the impairment identifiers. This results in an ill-defined assignment of the corresponding ratings, as those ratings would be arbitrarily assigned to the different possible columns in the output data matrix, which would create noise in the modeling results. Table 6.7 clarifies this with a specific example. To avoid such problems, the data organization algorithm needs to be extended to sort the ratings based on additional criteria. Examples proposed by Skowronek et al.<sup>8</sup> are to sort those connections according to the actual rating values or according to auxiliary information such as amount of contribution of the interlocutors to the conversation or their importance (social status, role) for the telemeeting.

Concerning permutations, the algorithm needs to be able to identify in the input data permutations of the same impairment sets. As an example, Table 6.8 shows that combinations of impairments in different calls for different interlocutors are actually permutations of the same impairment set.

<sup>8</sup> Janto Skowronek et al. "Method and Apparatus for Computing the Perceived Quality of a Multiparty Audio or Audiovisual Telecommunication Service or System". Patent application WO/2016/041593 (EP). Deutsche Telekom AG. Mar. 2016

1. Consider the following input data set of Interlocutor 1 for a four-interlocutors telemeeting, in which Interlocutor 1 perceives the same impairment dev2 on the connections of all participants, but assigns different ratings to each connection.

CallID	$I_{1,1}$	$I_{1,2}$	$I_{1,3}$	$I_{1,4}$	$R_1$	$R_{1,1}$	$R_{1,2}$	$R_{1,3}$	$R_{1,4}$
1	o	dev2	dev2	dev2	4	5	3	4	2

2. Concerning the connections to the other interlocutors, a sorting in terms of the impairment identifiers  $I_{ij}$  alone still allows six possibilities to assign the ratings  $R_{1,2}$  to  $R_{1,4}$  to the columns in the output data matrix.

Output data matrix for modeling, linking  $Q_i$  and  $Q_{ij}$ :

possibility	set definition using $I_{ij}$	CallID	$i =$ SubjectID	$R_i$	$R_{ii}$	$R_{ij}$ with $j \neq i$ , different sortings based on the values of $I_{ij}$			
a	o dev2 dev2	1	1	4	5	2	3	4	
b	o dev2 dev2	1	1	4	5	2	4	3	
c	o dev2 dev2	1	1	4	5	3	2	4	
d	o dev2 dev2	1	1	4	5	3	4	2	
e	o dev2 dev2	1	1	4	5	4	2	3	
f	o dev2 dev2	1	1	4	5	4	3	2	

Table 6.7: Example of a data set as input to the data organization algorithm for three interlocutors.

Call ID	Subject ID	Impairment Combination
1	1	$[I_{1,1} I_{1,2} I_{1,3}] = [o \text{ dev2 } \text{dev2}]$
2	2	$[I_{2,1} I_{2,2} I_{2,3}] = [\text{dev2 } o \text{ dev2}]$
3	3	$[I_{3,1} I_{3,2} I_{3,3}] = [\text{dev2 } \text{dev2 } o]$

$\Rightarrow$  are permutations of

SetID for $Q_i$	Set definition
1	$[o \text{ dev2 } \text{dev2}]$

Table 6.8: Example of permutations of impairment combinations in the input data belonging to the same impairment set defined for  $Q_i$ .

*Outline of Processing Steps and Detailed Realization* With the specification of tasks, input, and output, as well as the identification of challenges, the individual processing steps can be defined as follows:

1. Generate *templates*, i.e. set definitions based on combinations of impairments
  - (a) Go through input data
    - i. Take every combination of impairments per call and interlocutors, i.e. tuples of impairment identifiers  $I_{i,1} \dots I_{i,\#IL}$
    - ii. Sort tuples to account for permutations, including treatment of own connections and sorting of connections to other interlocutors
    - iii. Add to lists with templates
  - (b) Remove double entries from lists
2. Assign ratings to sets
  - (a) Go through input data and do per call and subject
    - i. Take tuple of impairment identifiers  $I_{ij}$
    - ii. Sort tuple to account for permutations
    - iii. Get tuple of corresponding ratings  $R_{ij}$ , in same order of impairment identifiers
    - iv. Get corresponding rating of  $R_i$
    - v. Compare sorted tuple with templates
    - vi. If a tuple matches a template  
Add to four output matrices: call and interlocutor identifier, tuple of  $I_{ij}$ ,  $R_i$ , tuple of  $R_{ij}$

A more detailed realization of the algorithm is described by means of pseudo code: Table 6.9 shows the first part on template generation, Table 6.10 the second part on data assignment.

---

```

01 /* automatic generation of templates defining all possible sets */
02 begin
03   for all callID do
04     for all subjectID do
05       /* get tuple of impairments  $I_{ij}$  */
06        $curTuple = data(callID, \text{impairments belonging to } subjectID)$ 
07       /* do reordering to get templates  $T_i$  &  $T_{ii}$  for  $Q_i$  and  $Q_{ii}$  (own connections) */
08        $I_{ii} = curTuple(subjectID)$ 
09        $I_{ij} \dots I_{jN} = curTuple(\text{all except } subjectID)$ 
10        $T_i = [I_{ii} \text{ sort } (I_{ij} I_{ik} \dots I_{iN}) \text{ with } i \neq j \neq k \in [1 \dots N]]$ 
11        $T_{ii} = T_i$ 
12       /* add to output: list of templates  $T_i$  */
13       add  $T_i$  to  $templateList\_T_i$ 
14       /* add to output: list of templates  $T_{ii}$  */
15       add  $T_{ii}$  to  $templateList\_T_{ii}$ 
16       /* do reordering to get templates  $T_{ij}$  for  $Q_{ij}$  by running through all except own connection */
17       for all connectionID  $\neq$  subjectID do
18          $I_{ii} = curTuple(subjectID)$ 
19          $I_{ij} = curTuple(connectionID)$ 
20          $I_{ik} \dots I_{jN} = curTuple(\text{all except } subjectID \ \& \ connectionID)$ 
21          $T_{ij} = [I_{ij} \ I_{ii} \ \text{sort}(I_{ik} \dots I_{iN})] \text{ with } i \neq j \neq k \in [1 \dots N]$ 
22         /* add to output: list of  $T_{ij}$  templates */
23         add  $T_{ij}$  to  $templateList\_T_{ij}$ 
24       remove double entries from  $templateList\_T_i$ 
25       remove double entries from  $templateList\_T_{ii}$ 
26       remove double entries from  $templateList\_T_{ij}$ 
27 end

```

---

Table 6.9: Pseudocode of the data organization algorithm: First algorithm part to automatically generate from the perceptual data the definition of sets of same impairments.

---

```

01 /* automatic assignment of data into sets by matching with templates */
02 begin
03   for all callID do
04     for all subjectID do
05       /* get tuple of impairments Iij */
06       curTuple = data(callID, impairments belonging to subjectID)
07       /* do reordering to match data with template formats for Qi and Qii (own connections) */
08       Iii = curTuple(subjectID)
09       Iij ... IjN = curTuple(all except subjectID)
10       Ti = [Iii sort(Iij Iik ... IiN)] with  $i \neq j \neq k \in [1..N]$ 
11       Tii = Ti
12       /* get indices pointing to reordered impairment tuples */
13       Ti_idx = [idxTo_Iii in_order_of_sorting_above(idxTo_Iij ... idxTo_IiN)]
14       Tii_idx = Ti_idx
15       /* get index pointing to corresponding Ri rating of current subjectID */
16       Ri_idx = idxTo_Ri
17       /* get indices pointing to reordered Rij rating tuples */
18       Rii_idx = [idxTo_Rij idxTo_Rii in_order_of_sorting_above(idxTo_Rik ... idxTo_RiN)]
19       /* match current tuple with templates for Qi and Qii and save corresponding indices */
20       for all entries in templateList_Ti do
21         get currentTemplate from templateList_Ti
22         if currentTemplate == Ti then
23           /* add to output: list of pointers to all data entries needed for the statistical analysis of Ri
24           ratings */
24           add [callID Ri_idx Ii_idx] to pointerList_Ri
25           /* add to output: list of pointers to all data entries needed for the modelling of Ri ratings
26           based on Rij ratings */
26           add [callID Ri_idx Rii_idx Ii_idx Iii_idx] to pointerList_Ri_Rij
27         for all entries in templateList_Tii do
28           get currentTemplate from templateList_Tii
29           if currentTemplate == Tii then
30             /* add to output: list of pointers to all data entries needed for the statistical analysis of Rii
31             ratings */
31             add [callID Rii_idx Iii_idx] to pointerList_Rii
32           /* do reordering to match data with template format for Qij by running through all except own
33           connection */
33           for all connectionID ≠ subjectID do
34             Iii = curTuple(subjectID)
35             Iij = curTuple(connectionID)
36             Iik ... IjN = curTuple(all except subjectID & connectionID)
37             Tij = [Iij Iii sort(Iik ... IiN)] with  $i \neq j \neq k \in [1..N]$ 
38             /* get indices pointing to reordered Rij rating tuples
39             Rij_idx = [idxTo_Rij idxTo_Rii in_order_of_sorting_above(idxTo_Rik ... idxTo_RiN)]
40             /* match current tuple with templates for Qi and Qii and save corresponding indices
41             for all entries in templateList_Tii do
42               get currentTemplate from templateList_Tii
43               if currentTemplate == Tii then
44                 /* add to output: list of pointers to all data entries needed for the statistical analysis of Rij
45                 ratings */
45                 add [callID Rij_idx Iij_idx] to pointerList_Rij
46 end

```

---

Table 6.10: Pseudocode of the data organization algorithm: Second algorithm part to automatically assign the perceptual data to the defined sets of same impairments.

## Summary

Since experimental details can influence results, a perceptual quality assessment method for multiparty telemeetings needs to consist of a stringent and ideally standardized test protocol. A first standardized document for this purpose is ITU-T Recommendation P.1301. On the one hand, P.1301 is based on a set of existing test protocols for non-multiparty situations and provides guidance to those methods; on the other hand, P.1301 provides advices how to apply and modify such protocols for the multiparty case.

Since a number of aspects of multiparty quality assessment are still under study, P.1301 needs to be augmented with additional currently non-standardized methods. One such aspect relevant for the present research is the multiparty quality assessment of asymmetric conditions, in which interlocutors participate in a telemeeting with different connection characteristics or equipment. In such asymmetric cases, interlocutors can have different perceptual experiences of the same telemeeting.

For that reason, this chapter presented a methodology that allows to translate the technical condition of a telemeeting system into such individual perceptual perspectives. The approach is a four-step procedure: (1) description of the multiparty situation, using a structured graphical representation of the end-to-end signal paths from all senders to all receivers; (2) identification of *Degradation-Types* and *Degradation-Points*, defining which types of degradations are under test and where they technically affect the signals; (3) analysis of signal paths and deduction of possible perceptual impairments; and (4) generation of a comprehensive representation for all interlocutors, presented in a convenient format such as a table. This method is a systematic approach to deduct the potential perceptual impairments from the technical system conditions, which provides all information to properly design, conduct and analyze a multiparty quality assessment test. However, it needs to be clarified that this method does not allow to verify if such impairments are actually perceived by participants.

Furthermore, the focus on perceptual impairments instead of technical degradations allows to optimize the experimental design in terms of the number of necessary test calls. Here the idea is to smartly combine technical conditions for the individual interlocutors into the same call in order to reduce the necessary number of test calls. Although the optimization potential depends on the actual set of considered degradations and impairments, and on the technical possibilities of the actual test system, this approach is a useful tool to optimize the amount of experimental effort.

Finally, when considering asymmetric conditions, the data sets collected in perceptual assessment tests can become quite complex given that individual interlocutors have different perspectives of the same telemeeting call. Here, a sorting of the data is necessary, which consists of two principle steps: (1) definition of data subsets sharing the

same perceptual impairments, and (2) grouping of data observations to those sets by applying selection criteria that match the definition of the sets. The actual definitions of the data subsets depend on the quality aggregation levels *Single-Perspective Telemeeting Quality*  $Q_i$  and *Individual Connection Quality*  $Q_{ij}$ . For ratings of  $Q_i$ , each set of ratings is defined by the specific combination of impairments on the individual connections. For ratings of  $Q_{ij}$ , each set of ratings is defined by the specific impairment of the corresponding individual connection under the constraint of a specific combination of impairments on the other individual connections. Since this sorting can be quite laborious and error-prone when performed manually, an algorithm is proposed and described to automate this sorting.

## 7

# Perception of Telemeeting Quality with focus on Group-Communication Aspects

---

---

This chapter contains text passages that either stem from previous texts of the author or that are based on such source texts. For readability purposes those passages are not marked as quotations. The main source documents used for this chapter are:

- Janto Skowronek et al. "Speech recordings for systematic assessment of multi-party conferencing". In: *Proceedings of Forum Acusticum 2011*. Aalborg, Denmark: European Acoustical Society, June 2011
- Janto Skowronek and Alexander Raake. "Einfluss von Bandbreite und räumlicher Sprachwiedergabe auf die kognitive Anstrengung bei Telefonkonferenzen in Abhängigkeit von der Teilnehmeranzahl". In: *Fortschritte der Akustik (DAGA2011) - 37. Deutsche Jahrestagung für Akustik*. Deutsche Gesellschaft für Akustik. Düsseldorf, Germany, Mar. 2011, pp. 873–874
- Janto Skowronek and Alexander Raake. "Investigating the effect of number of interlocutors on the quality of experience for multi-party audio conferencing". In: *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech2011)*. International Speech Communication Association. Florence, Italy, Aug. 2011, pp. 829–832
- Janto Skowronek, Falk Schiffner, and Alexander Raake. "On the influence of involvement on the quality of multiparty conferencing". In: *4th International Workshop on Perceptual Quality of Systems (PQS 2013)*. International Speech Communication Association. Vienna, Austria, Sept. 2013, pp. 141–146
- Janto Skowronek and Alexander Raake. "Assessment of Cognitive Load, Speech Communication Quality and Quality of Experience for spatial and non-spatial audio conferencing calls". In: *Speech Communication* (2015), pp. 154–175. DOI: 10.1016
- Janto Skowronek and Alexander Raake. *Development of scalable conversation test scenarios for multi-party telephone conferences*. ITU-T Contribution COM 12 - C187 - E. Geneva, Switzerland: International Telecommunication Union, Jan. 2011
- Janto Skowronek and Alexander Raake. *Listening test stimuli using scalable multiparty conversation test scenarios*. ITU-T Contribution COM 12 - C 286 - E. Geneva, Switzerland: International Telecommunication Union, Oct. 2011
- Janto Skowronek and Alexander Raake. *Number of interlocutors and QoE in multiparty telephone conferences*. ITU-T Contribution COM 12 - C 287 - E. Geneva, Switzerland: International Telecommunication Union, Oct. 2011
- Janto Skowronek and Alexander Raake. *Update on number of Interlocutors and QoE in multiparty telephone conferences*. ITU-T Contribution COM 12 - C035 - E. Geneva, Switzerland: International Telecommunication Union, Mar. 2013

Since additional documents served as further background material, see Appendix A for the full list of source documents and their contributions to this chapter.

---

---

### What this chapter is about

The present text is built around two aspects that differentiate a multi-party telemeeting from a conventional two-party (video) telephony call. One aspect is addressed in this chapter: the special communicative situation, that is, having a group conversation over a telecommunication system.

More specifically, this chapter reports on two studies that investigated the impact of different communicative situations on different aspects of quality perception. Moreover, in order to get a more complete picture, the studies considered also the interaction between such communicative situations and a number of technical system properties.

The first study investigated in two experiments the impact of the

interaction between technical characteristics of the system, expressed as *Technical System Capability*, and the communicative situation, expressed as *Communication Complexity*. The second study investigated in one experiment the impact of the interaction between *Technical System Capability* and the degree to which a single interlocutor takes part in the conversation, expressed as *Involvement*. In both studies, the impact was investigated on three aspects of quality perception: the perceived system performance, expressed as *Speech Communication Quality*; the perceived effort to follow the group communication, expressed as *Cognitive Load*, and the overall experience as such, expressed as *Quality of Experience*.

### 7.1 Introduction

Assuming a process-oriented perspective, which is here a simplified view on perception, quality perception can be regarded as a process in which an input – the telemeeting – is transformed into an output – the quality judgment. Concerning the input, two different principle aspects of telemeetings can influence quality perception: the technical system properties and the special communicative situation. Concerning the output, the conceptual model described in Chapter 5 considers *Telemeeting Quality* as a construct of two components, a *Telecommunication Component* and a *Group-Communication Component*.

Based on these considerations, telemeeting quality perception can be considered as a *two-in three-out process*: it takes two inputs (technology and communicative situation) and it generates three related outputs (*Telemeeting Quality*, *Telecommunication Component*, *Group-Communication Component*). Looking at telemeeting quality perception from this perspective, the research question arises how these multiple inputs and outputs are related to each other and to what extent this complexity can be decomposed. Such knowledge can be beneficial in different aspects: first, it can contribute to a better understanding of the contextual influences on quality perception in general; second, it can allow to draw proper conclusions from the results of a quality assessment test in light of the applied experimental details; and third, it can provide empirical evidence for the conceptual model with its two components, supporting the practical relevance of those concepts.

To investigate the relation between those multiple inputs and outputs, the two studies reported in this chapter considered both the communicative situation as well as technical system properties. Here, the communicative situation was further decomposed into the effect of *Communication Complexity* in one study and the effect of *Involvement* in another study. Furthermore, both studies aimed at measuring the three outputs *Telemeeting Quality*, *Telecommunication Component*, and *Group-Communication Component* separately.

In the following, the present chapter will first describe those conceptual input and output variables in more detail. Then it will report on the conduction and results of the individual studies.

## 7.2 *Experimental Variables on a Conceptual Level*

The present text considers the conduction of an experiment as a process transforming an input to an output. The input is called here the *Manipulated Variables*, the output is called *Measured Variables*. *Manipulated Variables* are those aspects of an experiment that are controlled by the investigator, and *Measured Variables* are those extracted from the experimental outcome, here the test participants' answers. The present section defines the *Manipulated* and *Measured Variables* on a conceptual level, while the detailed descriptions of the experiments later on will define more concrete individual measures for each of those conceptual variables. The motivation for this approach is to be able to group multiple but similar measures according to their common conceptual meaning.

The *Manipulated Variables* are *Communication Complexity*, *Involvement* and *Technical System Capability*. *Communication Complexity* refers to the communicative aspects of a telemeeting, concerning the interlocutors' contributions to a telemeeting, and is controlled here by simultaneously manipulating the number of interlocutors, the amount of information and the conversational structure. *Involvement* also refers to the communicative aspects of a telemeeting, focusing on the test participant's role as an interlocutor in a telemeeting, and is controlled here by different tasks for the test participants. *Technical System Capability* refers to the telemeeting system, and is manipulated here by applying different sound transmission and reproduction methods.

The *Measured Variables* are *Cognitive Load*, *Speech Communication Quality*, and *Quality of Experience*. *Cognitive Load* refers to the difficulty involved in following the conversation and represents the *Group-communication Component* of telemeeting quality. *Speech Communication Quality* refers to the telemeeting system as such and is thus focusing on the *Telecommunication Component* of telemeeting quality. *Quality of Experience* refers to the overall telemeeting quality, i.e. covering both components.

### 7.2.1 *Manipulated Variable: Communication Complexity*

*Communication Complexity* refers to the structure of the conversation in terms of who is contributing to the discussion at which point in time. Aspects that differentiate conversational structures are the number of speaker changes, interruptions, monologues, longer speech pauses etc. How these aspects differ between multiparty conferences depends on how the interlocutors contribute to the conversation. It is assumed in this work, that an interlocutor's contribution to the conversation depends on how he or she is able to perform four mental tasks: (1) understanding the speech signals from the others, (2) identifying speakers and their roles, (3) processing the information shared during the conference and extracting its implications, and (4) formulating adequate responses. Note that this is a simplified model

of the cognitive processes inside an interlocutor and should not be interpreted as a comprehensive cognitive model of human-to-human communication. Nevertheless, one can reasonably assume that those four mental tasks do take place, due to the following reasoning.

It is a straight-forward observation that participants perform the first task, and that its difficulty is determined by the speech transmission chain from mouth to ear (e.g. signal quality), the speaking behavior of the interlocutors (e.g. pronunciation), and language aspects (e.g. language fluency). The observation that participants also perform the other three tasks is based on knowledge from the literature on computer-mediated group-communication, see also Chapter 2. For example, Olson & Olson<sup>1</sup> discuss the importance for the interlocutors to create a common ground, by adapting to what they perceive from the other interlocutors. Other studies (e.g., Sanford et al.<sup>2</sup>, Daly-Jones et al.<sup>3</sup>, Masoodian<sup>4</sup>) also stress the importance of creating such common ground in computer-mediated communication by referring to the work of Clark & Brennan<sup>5</sup>. Similarly, Fussell & Benimoff<sup>6</sup> discuss that interlocutors "... strive for a shared understanding of the situation, the task, and of one another's background knowledge, expectations, beliefs, attitudes, and the like. They also construct a body of shared knowledge and understanding (common ground) which they can draw upon in their subsequent communications ..."

In terms of using *Communication Complexity* as a *Manipulated Variable* in an experiment, a number of aspects can change *Communication Complexity* by increasing or decreasing the difficulty of accomplishing the four mental tasks. A first aspect is the number of interlocutors. As participants will attempt to understand who the other interlocutors are, and try to adapt their responses, the number of interlocutors will influence the difficulty to identify speakers and formulate responses.

A second aspect is the amount of information shared during a telephone conference. Obviously the more information is shared, the higher is the effort to perform the task of information processing.

A third aspect is the conversational structure in terms of who is contributing to a conversation at which point in time. An increasing complexity in the conversational structure will increase the difficulty of the tasks of speaker identification and information processing, as the order in which people contribute and the order in which information is shared becomes less predictable.

There are more aspects that contribute to the complexity of a telephone conference conversation: the behavior of participants, the topic of the conference, different cultural, language or professional backgrounds, et cetera. However, to limit the scope of this work, the present text will focus on the three aspects directly related to *Communication Complexity*: the number of interlocutors, the amount of information to be exchanged and the conversational structure.

<sup>1</sup> Gary M. Olson and Judith S. Olson. "Distance Matters". In: *Human-Computer Interaction* 15 (2000), pp. 139-178

<sup>2</sup> Alison Sanford, Anne H. Anderson, and Jim Mullin. "Audio channel constraints in video-mediated communication". In: *Interacting with Computers* 16 (2004), pp. 1069-1094

<sup>3</sup> Owen Daly-Jones, Andrew Monk, and Leon Watts. "Some advantages of video conferencing over high-quality audio conferencing: fluency and awareness of attentional focus". In: *International Journal of Human-Computer Studies* 49 (1998), pp. 21-58

<sup>4</sup> Masood Masoodian. "Human-to-Human Communication Support for Computer-Based Shared Workspace Collaboration". PhD Thesis. Department of Computer Science, The University of Waikato, Hamilton, New Zealand, 1996

<sup>5</sup> Herbert H. Clark and Susan E. Brennan. "Grounding in Communication". In: *Perspectives on socially shared cognition*. Ed. by Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley. American Psychological Association, 1991, pp. 127-149

<sup>6</sup> Susan R. Fussell and Nicholas I. Benimoff. "Social and Cognitive Processes in Interpersonal Communication: Implications for advanced telecommunication technologies". In: *Human Factors* 37 (1995), pp. 228-250

### 7.2.2 *Manipulated Variable: Involvement*

The term *Involvement* is considered here as the degree to which a single interlocutor takes part in the conversation, that is in terms of Merriam-Webster<sup>7</sup> “to engage as a participant”.

A synonym is *Engagement*, a term which, however, is usually applied in the user experience domain in the context of user-system interaction. According to O’Brien & Toms<sup>8</sup>, *Engagement* is defined as “a category of user experience characterized by attributes of challenge, positive affect, endurance, aesthetic and sensory appeal, attention, feedback, variety/novelty, interactivity, and perceived user control.” Transferring this concept to the case of a group-communication via a telemeeting system, and labeling it here *Involvement*, the present text considers only those aspects that would fit also to a human to human communication: positive affect, sensory appeal, attention, variety/novelty, and interactivity.

In terms of using *Involvement* as a *Manipulated Variable* in an experiment, two aspects of the experimental task can change an interlocutor’s *Involvement*. The first aspect is the degree to which the task requires a test participant to contribute to the conversation; in other words, to which extent the task is *imposing* an *Involvement* on the test participant from a necessity point of view. Obviously, the two fundamentally different cases are to conduct either conversation tests or listening-only tests; the former requiring the test participant to fully engage into the conversation, the latter excluding the test participant from the conversation. Within each of these test paradigms, special test instructions and dedicated side-tasks can – potentially – also change the interlocutor’s *Involvement*. In the conversation test paradigm, the degree of freedom that is given to the test participants to formulate their contribution can play a role: if the task gives hardly any freedom, then the test participants might feel less involved, as they are then just following instructions; if the task gives full freedom, then the test participants might feel fully involved as it is a real natural conversation. In the listening-only test paradigm, side-tasks concerning the content of the material can play a role: if the task is just to listen to the stimuli and assess their quality, then the test participants might feel less involved, as they are not interested in what is being said; if the task has also a side-task that requires to follow closely the conversation, then the participant might feel more involved, as they are more interested in what is being said. In terms of O’Brien & Toms<sup>9</sup>, such tasks would address the aspects of interactivity and attention.

The second aspect is the degree to which the task influences *Involvement* by triggering a positive emotional affect and sensory appeal, or a certain amount of variety/novelty; in other words, to which extent the task is stimulating a participant-internal motivation to engage. Thus, a task that is perceived as a pleasurable, interesting and motivating experience can lead to a higher *Involvement* than a task that is perceived as less pleasant, less interesting and less motivating.

<sup>7</sup> Merriam-Webster. *Dictionary and Thesaurus*. retrieved on August 19, 2015. Merriam-Webster Incorporated, Springfield, MA. 2015. URL: <http://www.merriam-webster.com/>

<sup>8</sup> Heather L. O’Brien and Elaine G. Toms. “What is user engagement? A conceptual framework for defining user engagement with technology”. In: *Journal of the American Society for Information Science & Technology* 59.6 (2008), pp. 938–955

<sup>9</sup> Heather L. O’Brien and Elaine G. Toms. “What is user engagement? A conceptual framework for defining user engagement with technology”. In: *Journal of the American Society for Information Science & Technology* 59.6 (2008), pp. 938–955

### 7.2.3 *Manipulated Variable: Technical System Capability*

The term *Technical System Capability* refers to those technical properties of the considered system that influence its quality. *Technical System Capability* is introduced here as a term to differentiate between the technical quality provided by the system, and the quality judgment given by the user.

Apparently there is a link between the *Technical System Capability* and the perceived quality of a system in the sense that a low *Technical System Capability* would reflect low perceived quality, and a high *Technical System Capability* would reflect high perceived quality. However, it is known in the field that the mapping of technical system properties or system behavior to perceived quality is not straight-forward, which in turn requires to distinguish between the technical and perceptual side of quality.

In terms of using *Technical System Capability* as a *Manipulated Variable* in an experiment, audiovisual telemeetings can differ in their *Technical System Capability* in terms of numerous technical properties of the whole chain, from speech and video capture via data transmission to sound and video reproduction, see also Chapter 3. Different degradation types can lead to different perceptual quality dimensions<sup>10</sup> in the sense of the *Telecommunication Component*. Similarly different degradation types can influence communicative aspects<sup>11</sup> in the sense of the *Group-Communication Component*.

To limit the scope of this work, the present text covers each component with one technical property, but with different instances of that property. More specifically, the study on the impact of *Communication Complexity* addresses both components; thus it uses combinations of two technical properties. The study on the impact of *Involvement*, however, focuses only on the *Telecommunication Component*; thus it uses one technical property.

### 7.2.4 *Measured Variable: Cognitive Load*

*Cognitive Load* is a concept in cognitive psychology in the context of learning, and refers to the human working memory, which has a limited capacity for processing information per instance of time. When performing a task, this capacity is shared by the intrinsic nature of the task (intrinsic load), the instructions of the task (extraneous load) and some amount used for building up cognitive schemata for the task (germane load). More details about the individual aspects of *Cognitive Load* are given for example in Schnotz & Kuerschner<sup>12</sup> and Bruenken et al.<sup>13</sup>

Concerning typical multiparty telemeeting situations, the present text argues that interlocutors experience fatigue due to a high *Cognitive Load*. In the context of learning, the difficulty of a learning task and its instruction material is determining *Cognitive Load*, see e.g. Bruenken et al. Similarly, the difficulty for a participant to follow and contribute to a telemeeting will influence *Cognitive Load*. That means, the difficulty to perform the four mental tasks of speech under-

<sup>10</sup> Marcel Wältermann. *Dimension-based Quality Modelling of Transmitted Speech*. Springer, 2013

<sup>11</sup> Jessica J. Baldis. "Effects of Spatial Audio on Memory, Comprehension, and Preference during Desktop Conferences". In: *Proceedings of the ACM CHI 2001 Human Factors in Computing Systems Conference*. Ed. by Michel Beaudouin-Lafon and Robert J. K. Jacob. Vol. 3. 1. 2001, pp. 166–173

Alexander Raake et al. "Listening and conversational quality of spatial audio conferencing". In: *Proceedings of the AES 40th International Conference*. Tokyo, Oct. 2010

<sup>12</sup> Wolfgang Schnotz and Christian Kuerschner. "A Reconsideration of Cognitive Load Theory". In: *Educational Psychology Review* 19.4 (2007), pp. 469–508

<sup>13</sup> Roland Bruenken, Jan L. Plass, and Detlev Leutner. "Direct Measurement of Cognitive Load in Multimedia Learning". In: *Educational Psychologist* 38.1 (2003), pp. 53–61

standing, speaker identification, information processing, and response formulation will constitute *Cognitive Load*. In this respect, *Cognitive Load* is a *Measured Variable* that refers to the *Group-Communication Component* of Telemeeeting Quality.

### 7.2.5 *Measured Variables: Speech Communication Quality and Quality of Experience*

Various aspects contribute to telemeeeting quality, see Chapters 4 & 5, and Möller and colleagues proposed to organize these aspects along taxonomies for different telecommunication services such as telephone connections<sup>14</sup>, spoken-dialogue systems<sup>15</sup> or multi-modal human-machine interaction systems<sup>16</sup>. Inspired by Möller's taxonomy for telephone connections, this work considers two groups of quality aspects. The first group is directly related to how well the system enables speech communication: (one-way) voice transmission quality, ease of communication, and conversation effectiveness. In terms of Möller, these aspects are subsumed under the term *Speech Communication Quality*. The second group reflects the perceived quality of a service in a more general view: overall impression, pleasantness, satisfaction, and acceptance. As these aspects address the delight of the user of the service, the present text subsumes them here under the term *Quality of Experience*.

Defining two such levels of perceived quality in this work helps to distinguish how participants perceive the performance of a telemeeeting system and how participants experience a telemeeeting as such. In this way a terminology is available that allows to describe the quality with a focus on the system performance perceived by the user (*Speech Communication Quality*), and in a wider sense incorporating also aspects of using the system in the context of the communication situation at large (*Quality of Experience*).

In this respect, *Quality of Experience* is a *Measured Variable* that covers both the *Telecommunication* and the *Group-Communication Component* of telemeeeting quality. Similarly, *Speech Communication Quality* covers both components as well, since it subsumes quality features of perception and action as well as interaction. However, in the present context, *Speech Communication Quality* is intended to have either a strong focus on the *Telecommunication Component* or on the *Group-Communication Component*. This is possible, if the test protocol triggers test participants to interpret the questions on *Speech Communication Quality* accordingly, for instance by the actual wording of the questions (see Section 5.4 for details).

## 7.3 *Quality Impact of Communication Complexity*

This section reports on two experiments investigating the quality impact of *Communication Complexity*, henceforth referred to as Experiments CC1 and CC2 (CC standing for Communication Complexity).

<sup>14</sup> Sebastian Möller. *Assessment and Prediction of Speech Quality in Telecommunications*. Kluwer Academic publishers, 2000

<sup>15</sup> Sebastian Möller. *Quality of Telephone-Based Spoken Dialogue Systems*. New York, USA: Springer, 2005

<sup>16</sup> Sebastian Möller et al. "A Taxonomy of Quality of Service and Quality of Experience of Multimodal Human-Machine-Interaction". In: *Proceedings of the International Workshop on Quality of Multimedia Experience QoMEX*. 2009

### 7.3.1 Goals

*Motivation* One of the two main differentiators between a multiparty telemeeting and a conventional two-party call is the special communicative situation. In this line of thought, the most obvious difference between multiparty and two-party calls in terms of communicative aspects is the *Communication Complexity*. For that reason, the present study investigates the impact of *Communication Complexity* on the perception of telemeeting quality.

*Approach* The present study aims at investigating the impact of *Communication Complexity* in the context of different *Technical System Capabilities*. This leads to a two-factorial study using both *Communication Complexity* and *Technical System Capabilities* as *Manipulated Variables*.

*Communication Complexity* is controlled by simultaneously manipulating the number of interlocutors, the amount of information and the conversational structure. Note that the number of interlocutors – henceforth abbreviated as #IL – is representatively used as an identifier for the different levels of *Communication Complexity*. *Technical System Capability* is manipulated by applying different sound reproduction methods – henceforth abbreviated as *SndRepr*. Concerning the *Measured Variables*, the study used three variables: *Cognitive Load*, *Speech Communication Quality* and *Quality of Experience*.

According to the iterative research method (Section 1.4), the study consisted of two experiments, CC1<sup>17</sup> and CC2<sup>18</sup>. Conducting two experiments allowed to improve the experimental method in the second experiment CC2 based on the insights gained in the first experiment CC1. Concerning the test methodology, both experiments focused on audio-only communication and applied a listening-only test paradigm. Focusing on audio-only was motivated by the assumption that the effect of *Communication Complexity* on *Cognitive Load* might be stronger for an audio-only than an audiovisual telemeeting, especially when it comes to aspects such as speaker identification effort. The decision to conduct a listening-only test was based on the following reasoning: in order to obtain properly interpretable results, the conversation scenarios needed to be highly comparable, and this can be best achieved by generating recordings of highly structured conversations and presenting them to the test participants.

*Hypotheses* The two experiments actually investigated slightly different hypotheses due to differences in the *Measured Variables* and the questionnaires. These variables, however, can be put into the framework of the *Measured Variables* described in Section 7.2, which in turn allows a common definition of hypotheses for both experiments. For that purpose, this paragraph will first describe the mapping of the original *Measured Variables* to the common conceptual framework. Then this paragraph will formulate a corresponding set of hypotheses that are tested in this work.

<sup>17</sup> Janto Skowronek and Alexander Raake. "Investigating the effect of number of interlocutors on the quality of experience for multi-party audio conferencing". In: *Proceedings of the 12th Annual Conference of the International Speech Communication Association (InterSpeech2011)*. International Speech Communication Association. Florence, Italy, Aug. 2011, pp. 829–832

<sup>18</sup> Janto Skowronek and Alexander Raake. "Assessment of Cognitive Load, Speech Communication Quality and Quality of Experience for spatial and non-spatial audio conferencing calls". In: *Speech Communication* (2015), pp. 154–175. DOI: 10.1016

At the time of preparing the first experiment CC<sub>1</sub>, the *Measured Variables* were defined as *Cognitive Load* and *Quality*, whereas *Quality* was not as precisely defined as it is described in Section 7.2. The reason is that the concept of the *Telecommunication* and *Group-Communication Components* were not yet developed at a stage as presented in Section 5.4 and the differentiation between *Speech Communication Quality* and *Quality of Experience* were not yet considered at a stage as presented in Section 7.2. Instead, the different employed measures were chosen to cover different aspects of *Quality* without a strict link to the aspects of *Speech Communication Quality* vs. *Quality of Experience* and *Telecommunication* vs. *Group-Communication Components*. However, by re-considering the actual wordings of the employed questions, it is possible to split *in retrospect* the variable *Quality* into three more specific *Measured Variables*: one variable reflecting *Speech Communication Quality* with a focus on the *Telecommunication Component*, a second variable reflecting *Speech Communication Quality* with a focus on the *Group-Communication Component*, and a third variable reflecting quality in a broader sense, whereas a more precise assignment to either *Speech Communication Quality* or *Quality of Experience* is not possible.

The second experiment CC<sub>2</sub>, however, already introduced the distinction between *Speech Communication Quality* and *Quality of Experience*, but it did not consider the *Telecommunication* and *Group-Communication Components*. Again, by re-considering the actual wordings of the employed questions, a weighting of *Speech Communication Quality* towards a focus on the *Telecommunication Component* can be identified.

Moving forward to formulating a set of hypotheses, conceptually two *Manipulated Variables* (*SndRepr* & *#IL*) and three *Measured Variables* (*Cognitive Load*, *Speech Communication Quality*, & *Quality of Experience*) were used in the two experiments. This allows to define six hypotheses of the form “[*Manipulation Variable*] has an impact on [*Measured Variable*]”. With the more precise definitions of the *Measured Variables* described above, it is possible to refine the set of hypotheses for each experiment as it is shown in Table 7.1. All resulting hypotheses will be tested in this chapter.

### 7.3.2 Experimental Factors

*Test Tasks* In both experiments the task for test participants was a conventional listening-only assessment task (“Listen to the conversation, and give ratings afterwards”), augmented with a memory test as a side task (“Memorize who said what during the conversation, and answer a memory test afterwards”).

*Conversation Scenarios* To achieve different levels of *Communication Complexity*, dedicated conversation scenarios needed to be developed. The main challenge for the scenarios was to be as comparable as possible while at the same time differing in the *Communication Complexity*.

Measured Variables	Manipulated Variables	
	<i>Communication Complexity</i> (i.e. Number of Interlocutors #IL)	<i>Technical System Capability</i> (i.e. Sound Reproduction Method <i>SndRepr</i> )
Experiment CC1		
<i>Cognitive Load</i>	Hypothesis H1	Hypothesis H2
<i>Speech Communication Quality</i> with focus on <i>Telecommunication Component</i>	Hypothesis H3.1	Hypothesis H4.1
<i>Speech Communication Quality</i> with focus on <i>Group-Communication Component</i>	Hypothesis H3.2	Hypothesis H4.2
<i>Speech Communication Quality</i> and/or <i>Quality of Experience</i>	Hypothesis H5	Hypothesis H6
Experiment CC2		
<i>Cognitive Load</i>	Hypothesis H1	Hypothesis H2
<i>Speech Communication Quality</i> with focus on <i>Telecommunication Component</i>	Hypothesis H3	Hypothesis H4
<i>Quality of Experience</i>	Hypothesis H5	Hypothesis H6

Table 7.1: Overview of the *Manipulated* and *Measured Variables* and definition of the resulting hypotheses for the experiments CC1 and CC2. Each hypothesis can be phrased as “[Manipulated Variable] has an impact on [Measured Variable]”.

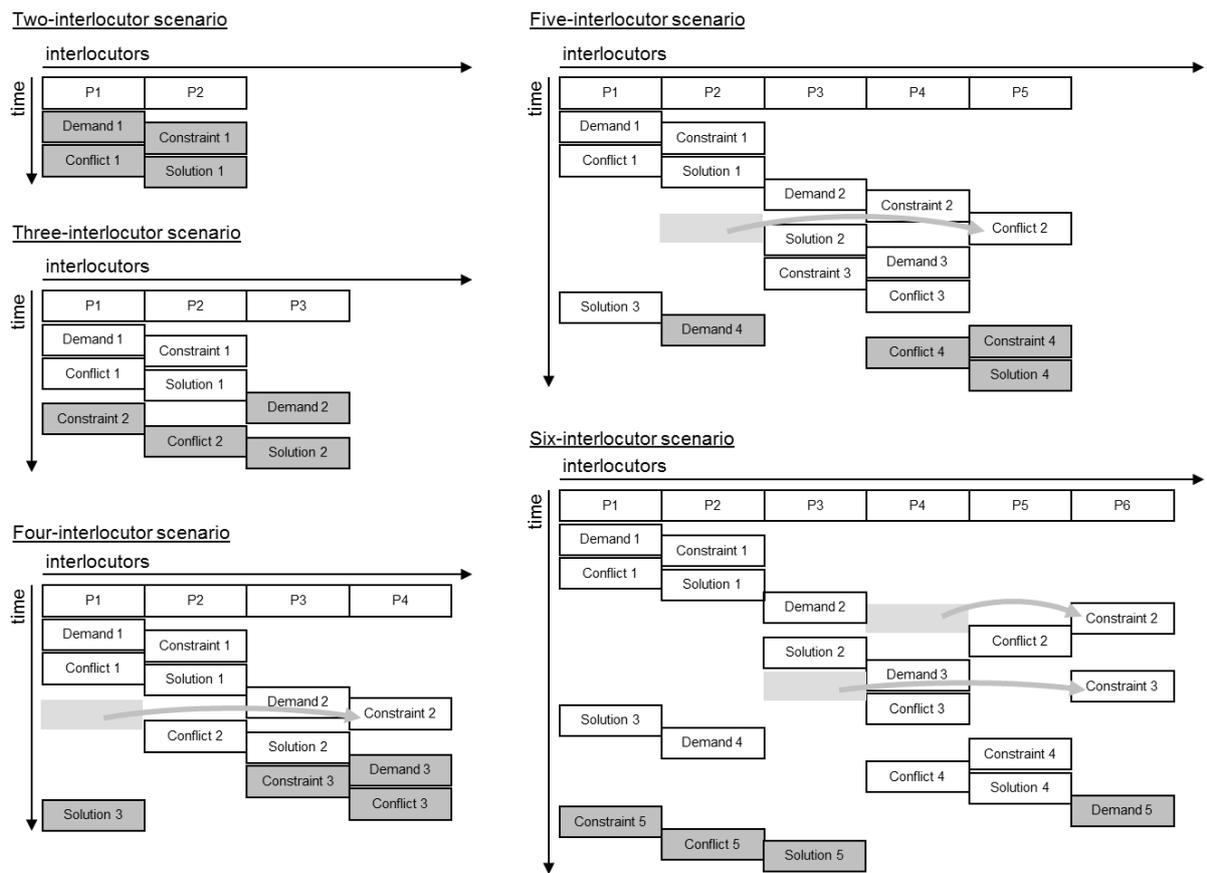
The following descriptions outline, how this was achieved.

The scenario topics were business-related and varied from aligning a meeting date to preparing the steps for publishing a music CD. All scenarios had a common underlying conversational structure, consisting of four conversation phases: a welcome phase; a problem solving phase, in which a number of agenda items are discussed; an information exchange phase, in which information items such as email addresses are shared; and a farewell phase. We realized the increasing *Communication Complexity* in a two-stage approach. First, with each new interlocutor, a new agenda item consisting of four contributions (question, constraint, conflict, solution) was added to the problem solving phase. Second, the previous distribution of conversation contributions across interlocutors was maintained when a new interlocutor and new contributions were added. For example, this means that in a two-interlocutor scenario, all contributions of the first (and only) agenda item are distributed across the two interlocutors. When extending this to a three-interlocutor scenario, the two original interlocutors remained the ones contributing to the first agenda item; the new third interlocutor was involved only in the added second agenda item. This resulted into a scaling scheme as shown in Figure 7.1.

The scenarios were realized in form of scripts, providing the content in bullet points and pictographs to facilitate more natural conversations. In Skowronek et al.<sup>19</sup>, we decided to implement scenarios with two, three, four and six interlocutors. The two-interlocutor scenarios served as anchor points; the three- and four-interlocutor scenarios represented typical real-life telephone conferences; the six-interlocutor scenarios represented more extreme cases in terms of *Communication Complexity*. The total number of scenarios created by the author was 13: three scenarios for each number of interlocutors ( $\#IL \in [2, 3, 4, 6]$ ) and one additional three-interlocutor scenario reserved for training test participants. The full scenario scripts are electronically available<sup>20</sup>.

<sup>19</sup> Janto Skowronek et al. “Speech recordings for systematic assessment of multi-party conferencing”. In: *Proceedings of Forum Acusticum 2011*. Aalborg, Denmark: European Acoustical Society, June 2011

<sup>20</sup> Janto Skowronek. *xCT - Scalable Multiparty Conversation Test scenarios for conferencing assessment (Version 01)*. 2015. DOI: 10.5281/zenodo.16138



With each new interlocutor (P1 to P6), a new agenda item (1 to 5) with four contributions – Demand, Constraint, Conflict, Solution – is added (dark gray). The distribution of previous contributions remains unchanged, except in some cases in which minor deviations were used to better balance the number of contributions per interlocutor (light gray arrows), and respective contributions were shifted to another interlocutor.

Figure 7.1: Scaling scheme to increase the *Communication Complexity* from a two-interlocutor to a six-interlocutor scenario. Previously presented in Janto Skowronek et al. “Speech recordings for systematic assessment of multiparty conferencing”. In: *Proceedings of Forum Acusticum 2011*. Aalborg, Denmark: European Acoustical Society, June 2011

*Recording of Speech Material* To obtain the speech material for the listening-only test, six male German professional speakers played the 13 scenarios via a simulated telemeeting system, consisting of high-quality headsets (Sennheiser HMD-410 & HME-46-3-6), a low-latency fullband audio transmission path (44100 kHz sampling frequency) using the direct monitoring capabilities of the computer audio cards (RME Multiface II). Spatial sound reproduction using a binaural rendering software referred to as SoundScapeRenderer<sup>21</sup> was employed to provide optimal reproduction conditions during the conference calls. In an editing process of the resulting recordings, the author optimized the fit between the recordings and the intended conversation structure defined by the scenario scripts. For example, the author deleted deviating and unnecessary portions of the conference call without affecting the naturalness of the resulting recordings. The naturalness has been checked afterwards by a listener who was not involved in the recording or editing process. Furthermore, the author performed a structural analysis using annotations as well as a conversational analysis based on computed speaker state probabilities (see Hoeldtke and Raake<sup>22</sup> for further details) in order to verify the comparability between the recordings. The outcome of that process was 13 recordings with a total length of 97 minutes and an optimized comparability between individual calls. For more details on the recording session, the editing process and the evaluation, see the report in Skowronek et al<sup>23</sup>.

*Technical Conditions* From the many possible options to vary *Technical System Capability* by means of varying the sound reproduction method (*SndRepr*), the choice fell on two system parameters: bandwidth and spatial sound reproduction. These two parameters appeared to be important aspects with regard to *Cognitive Load* and *Speech Communication Quality*. Some studies<sup>24</sup> showed that spatial sound reproduction can reduce *Cognitive Load* in conferencing scenarios. Bandwidth, especially narrowband vs. wideband speech, is known to have a significant impact on speech quality<sup>25</sup> and speech intelligibility<sup>26</sup>. To achieve clearly perceivable quality differences, the bandwidth was varied from narrowband (NB) to fullband (FB) and spatial sound reproduction varied from non-spatial to spatial (i.e. binaural) including head-tracking.

From the four possible combinations of bandwidth and spatial sound reproduction, the author selected three: NB/non-spatial (*SndRepr* = 1), FB/non-spatial (*SndRepr* = 2), FB/spatial (*SndRepr* = 3). This decision was driven by practical and theoretical limitations. The practical limitation was the fact that only three recordings per number of interlocutors (*#IL*) were available. Hence, for two reasons, only three technical conditions could be tested: first, aiming for a within-subject design, all test participants should judge all presented combinations of *#IL* and *SndRepr*; second, it was necessary to avoid that test participants listen to the same recordings multiple times as the questionnaire also contained a memory test. The theoretical

<sup>21</sup> Matthias Geier, Jens Ahrens, and Sascha Spors. "The soundscape renderer: A unified spatial audio reproduction framework for arbitrary rendering methods". In: *Proceedings of the AES 124th Convention*. Amsterdam, The Netherlands, 2008

<sup>22</sup> Katrin Hoeldtke and Alexander Raake. "Conversation analysis of multi-party conferencing and its relation to perceived quality". In: *Proceedings of the IEEE International Conference on Communications ICC*. Kyoto, Japan, June 2011

<sup>23</sup> Janto Skowronek et al. "Speech recordings for systematic assessment of multi-party conferencing". In: *Proceedings of Forum Acusticum 2011*. Aalborg, Denmark: European Acoustical Society, June 2011

<sup>24</sup> Jessica J. Baldis. "Effects of Spatial Audio on Memory, Comprehension, and Preference during Desktop Conferencing". In: *Proceedings of the ACM CHI 2001 Human Factors in Computing Systems Conference*. Ed. by Michel Beaudouin-Lafon and Robert J. K. Jacob. Vol. 3. 1. 2001, pp. 166-173

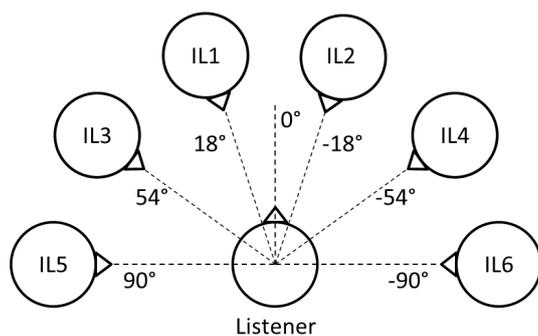
Alexander Raake et al. "Listening and conversational quality of spatial audio conferencing". In: *Proceedings of the AES 40th International Conference*. Tokyo, Oct. 2010

<sup>25</sup> Alexander Raake. *Speech Quality of VoIP – Assessment and Prediction*. Chichester, West Sussex, UK: Wiley, 2006

<sup>26</sup> Hans W. Gierlich. "Wideband Speech Communication – The Quality Parameters as Perceived by the User". In: *Proceedings of Forum Acusticum 2005*. 2005

limitation was to avoid an unclear ranking of different instances of the *Technical System Capability*, which would have been the case for the combinations FB/non-spatial and NB/spatial. The motivation for avoiding such unclear ranking is to be able to draw valid conclusions due to the following reasoning: In order to study the effect of different levels of *Technical System Capability*, it would be necessary to be able to determine upfront the difference in the *Technical System Capability* between FB/non-spatial and NB/spatial. This, however, is not possible without a work-around in form of a pre-study to identify this ranking experimentally. Since such a pre-study would need to be rather comprehensive, as it requires a valid link between *Technical System Capability* and the pre-study results, the author refrained from further following this approach and decided to focus on technical conditions that can be properly interpreted.

To realize the three sound reproduction conditions, test participants were seated in a sound-proof room and listened to the stimuli via AKG K601 headphones equipped with a FastTrak headtracker at a listening level of 79 dB SPL. To this aim, the SoundScapeRenderer software<sup>27</sup> was used for replaying the different stimuli either diotically (non-spatial conditions *SndRepr* = 1 & 2) or dichotically (spatial condition *SndRepr* = 3). The computer running the rendering software was placed outside the listening room. In the spatial condition, the interlocutors No. 1 to 6 were distributed in the virtual acoustic space as shown in Figure 7.2. For the FB-conditions, the recordings were used as is. For the NB-conditions, the author filtered the signals using a telephone bandpass tool (called *c712demo.exe*, bandpass cutoff frequencies at 230 & 3530 Hz) available from the ITU<sup>28</sup>. Table 7.2 gives an overview of the three realized sound reproduction conditions.



<sup>27</sup> Matthias Geier, Jens Ahrens, and Sascha Spors. "The soundscape renderer: A unified spatial audio reproduction framework for arbitrary rendering methods". In: *Proceedings of the AES 124th Convention*. Amsterdam, The Netherlands, 2008

<sup>28</sup> ITU-T. *Recommendation G.191 - Software tools for speech and audio coding standardization*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2010

Figure 7.2: Spatial sound rendering used in the listening test: Positioning of the interlocutors IL1 to IL6 in the virtual acoustic space in front of the listener. Note: As head tracking was deployed, the given angles assume that the listener looks straight ahead.

Reproduction method	<i>SndRepr</i> = 1: NB non-spatial	<i>SndRepr</i> = 2: FB non-spatial	<i>SndRepr</i> = 3: FB spatial
Frequency bandwidth	300-3400 Hz	0-20050 Hz	0-20050 Hz
Audio reproduction	diotic	diotic	dichotic
Speaker location angles	all at 0°	all at 0°	±18°, ±54°, ±90°
Head-tracking	no	no	yes

Table 7.2: Specifications of the sound reproduction conditions used in the listening test.

### 7.3.3 Method

*Design* The design was driven by the following constraints in terms of temporal order of stimuli, scenarios, number of interlocutors *#IL*, and sound reproduction methods *SndRepr*:

1. Three scenario recordings per *#IL* were available.
2. Four different *#IL* and three *SndRepr*, thus 12 possible combinations in total, needed to be tested.
3. All scenarios should be presented with all three *SndRepr*.
4. Each test participant should listen to a scenario recording only once.
5. Overall test time per subject was estimated to be 180 minutes, this needed to be split into smaller sessions to account for fatigue.

For Experiment CC<sub>1</sub>, the choice fell on a test with two 90-minutes sessions per test participant, with at least one day in-between sessions. The design uses a 12<sup>th</sup>-order latin square design to balance the temporal order of the 12 possible combinations of *#IL* and *SndRepr*, and a third-order latin square design repeated every three test participants to balance per *#IL* the assignment of the three scenarios to the three *SndRepr*.

For Experiment CC<sub>2</sub>, the outcome of Experiment 1 suggested to have separate sessions per *#IL*; thus the test comprised four sessions per test participant, with at least one day in-between sessions. The design uses a third-order greco-latin square design to balance for each session the temporal order of scenarios and sound reproduction methods, and a fourth-order latin square design to balance the temporal order of sessions, i.e. the order of number of interlocutors.

*Test participants* The characteristics of the invited test participants are given in Table 7.3.

Participant characteristics	Experiment CC <sub>1</sub>	Experiment CC <sub>2</sub>
Number of participants	13	25
Age: mean / minimum / maximum	31.3 / 26 / 40	28.8 / 23 / 43
Gender: female / male, in percent (absolute numbers)	38% / 62% (5 / 8)	40% / 60% (10 / 15)
Multiparty experience in professional or private life	100%	100%
Recruitment area	from same laboratory	university area, outside laboratory
Background / profession	researchers, not working on telephone quality	researchers, students, staff, alumni
Hearing impairments (self-reported)	none	none

Table 7.3: Test participant statistics for the two listening-only tests on *Communication Complexity*, CC<sub>1</sub> and CC<sub>2</sub>.

*Procedure* In Experiment CC<sub>1</sub>, participants came to two sessions, each lasting around 90 minutes, with a short break after 45 minutes. In Experiment CC<sub>2</sub>, participants came to four sessions; each session lasted between 30 and 60 minutes, depending on the recording length of the test stimuli, which were determined by the number of

interlocutors. In each session, test participants listened to the test stimuli and filled in a questionnaire after each stimulus. The test participants were not allowed to take any notes while listening to the recordings. However, for each stimulus, they got one paper sheet with basic information about the interlocutors (name, affiliation, role). In the first session, the author explained the task to the subjects in an introduction phase, followed by a training phase. During the training, test participants were listening to a three-interlocutor conference that was split up into three equally long parts, each presented with one of the three sound reproduction methods, and the test participants had to fill in a trial questionnaire. At the end of the last session the author conducted a short outtake interview with the test participants.

#### 7.3.4 Data Acquisition

*Questionnaires and Measures* The questionnaires used in the two experiments were designed such that we extracted for each *Measured Variable* multiple measures from the test participants' answers. This was motivated by the fact that there was no agreed-upon questionnaire available for the present specific purpose. The use of such redundancy enables to cover individual facets of the *Measured Variables*, which would in turn enable to identify, which of those measures would be appropriate for a future standardized questionnaire.

The selection of the different measures and the corresponding questionnaire design was inspired by previous similar research<sup>29</sup>, standard quality assessment methods<sup>30</sup>, and Möller's taxonomy for the quality of communication services<sup>31</sup>. Furthermore, the questionnaire of Experiment CC2 was modified based on the insights of Experiment CC1.

The questionnaire of Experiment CC1 consisted of four main parts: quality judgment (question Q1.1), recall in terms of the topics discussed during the conversation (questions Q2.1a – Q2.3a, Q2.1b – Q2.3b), a memory test in terms of "who said what" with ten quotations per conversation (questions Q3.1a – Q3.10a, Q3.1b – Q3.10b), and judgments of speech communication quality and cognitive load (questions Q4.1 – Q4.6). The questionnaire of Experiment CC2 consisted of five main parts: judgment of overall impression (question Q1.1), recall in terms of focal assurance (questions Q2.1a – Q2.4a, Q2.1b – Q2.4b), a memory test in terms of "who said what" with 16 quotations per conversation (questions Q3.1a – Q3.16a, Q3.1b – Q3.16b), judgments of speech communication quality and cognitive load (questions Q4.1 – Q4.5), and judgments of service satisfaction (questions Q5.1 – Q5.3). Table 7.4 provides English translations of the questions as well as the German originals.

The judgments were obtained using a continuous rating scale (developed by Bodden and Jekosch<sup>32</sup>, recommended by the ITU-T as standardized scales for testing spoken dialogue systems<sup>33</sup>). An example extract from the questionnaires is given in Figure 7.3.

Table 7.5 gives an overview of the link between the *Measured Vari-*

<sup>29</sup> Jessica J. Baldis. "Effects of Spatial Audio on Memory, Comprehension, and Preference during Desktop Conferences". In: *Proceedings of the ACM CHI 2001 Human Factors in Computing Systems Conference*. Ed. by Michel Beaudouin-Lafon and Robert J. K. Jacob. Vol. 3. 1. 2001, pp. 166–173

Alexander Raake et al. "Listening and conversational quality of spatial audio conferencing". In: *Proceedings of the AES 40th International Conference*. Tokyo, Oct. 2010

<sup>30</sup> ITU-T. *Recommendation P.800 - Methods for objective and subjective assessment of quality*. International Standard. Geneva, Switzerland: International Telecommunication Union, 1996

ITU-T. *Recommendation P.851 - Subjective quality evaluation of telephone services based on spoken dialogue systems*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2003

<sup>31</sup> Sebastian Möller. *Assessment and Prediction of Speech Quality in Telecommunications*. Kluwer Academic publishers, 2000

<sup>32</sup> Markus Bodden and Ute Jekosch. *Entwicklung und Durchführung von Tests mit Versuchspersonen zur Verifizierung von Modellen zur Berechnung der Sprachübertragungsqualität*. Project Report (unpublished). Institute of Communication Acoustics, Ruhr-University Bochum, Germany, 1996

<sup>33</sup> ITU-T. *Recommendation P.851 - Subjective quality evaluation of telephone services based on spoken dialogue systems*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2003

ables, the actual measures used, and descriptions on how they were extracted from the questionnaires. The following paragraphs provide detailed explanations of the links between measured variable, actual measures and questionnaires.

1.1 Wie war ihr persönlicher **Gesamteindruck** von dieser Telefonkonferenz?



*Cognitive Load* The most direct measure for *Cognitive Load* is *Concentration Effort*, obtained from participants' ratings of questions Q4.3 in Experiment CC1 and Q4.4 in Experiment CC2. Four additional measures addressed the effort to identify speakers and their roles, one of the mental tasks discussed in Section 7.2. The first measure is *Speaker Recognition Effort*, obtained from participants' ratings of questions Q4.1 in Experiment CC1 and Q4.2 in Experiment CC2. The next two measures, *Speaker Recognition Performance* and *Speaker Recognition Confidence*, are extracted from the memory tests. The memory test of Experiment CC1 consisted of ten questions (Q3.1a – Q3.10a) referring to ten transcribed quotes from the conversation. For each quote, test participants could answer who of the interlocutors had made that statement, or they could choose the option "I don't know". Furthermore, ten additional questions (Q3.1b – Q3.10b) were asking for each quote how confident test participants were in their judgment, expressed in percent. *Speaker Recognition Performance* is computed as the percentage of correctly identified speakers per quote (Q3.1a – Q3.10a); *Speaker Recognition Confidence* is computed as the mean of the confidence ratings (Q3.1b – Q3.10b). Similarly, Experiment CC2 applied the same memory test, but now with 16 quotations in order to increase the number of available data points and thus to increase the noise robustness of the two measures.

A crucial note on the memory task should be inserted here: While the memory test followed the method in previous research<sup>34</sup>, the author carefully treated the varying number of interlocutors, ranging from two to six. The motivation was to measure the intrinsic cognitive load, i.e. the cognitive load inherent in the memory task, while keeping the extraneous cognitive load, i.e. the cognitive load inherent in the instructions of the task, constant. To achieve this, test participants always had to choose between only two possible interlocutors per quote, independently from the number of interlocutors in a conversation. In this way the questionnaire kept the task "decide between two interlocutors" (extraneous cognitive load) constant for all stimuli; but measured the cognitive load for the task "remember for all interlocutors who said what" (intrinsic cognitive load), which varied with the different number of interlocutors.

Coming back to the measurement of the effort to identify speakers

Figure 7.3: Example extract from the used questionnaire (in German): question Q1.1 "What was your overall impression of this telephone conference? – extremely bad ... ideal"

<sup>34</sup> Jessica J. Baldis. "Effects of Spatial Audio on Memory, Comprehension, and Preference during Desktop Conferences". In: *Proceedings of the ACM CHI 2001 Human Factors in Computing Systems Conference*. Ed. by Michel Beaudouin-Lafon and Robert J. K. Jacob. Vol. 3. 1. 2001, pp. 166–173

Ryan Kilgore, Mark Chignell, and Paul Smith. "Spatialized audioconferencing: what are the benefits?" In: *Proceedings of the 2003 conference of the Centre for Advanced Studies on Collaborative research CASCON '03*. IBM Press, 2003, pp. 135–144

Alexander Raake et al. "Listening and conversational quality of spatial audio conferencing". In: *Proceedings of the AES 40th International Conference*. Tokyo, Oct. 2010

Experiment CC1		
Part	Question ID	Wording / Description, EN: English translation, DE: German original
1	Q1.1	EN: "What was your <b>overall quality impression</b> of the connection? – <i>extremely bad ... ideal</i> " DE: "Wie war Ihr persönlicher <b>Gesamteindruck von der Qualität</b> der Verbindung? – <i>extrem schlecht ... ideal</i> "
2	Q2.1a ... Q2.2a	EN: "Which opinion or information shared <b>Mr. X</b> during the conference? – <i>Option A, Option B, I don't know</i> " DE: "Welche Meinung vertrat bzw. welche Information gab <b>Herr X</b> in dem Gespräch? – <i>Option A, Option B, Ich weiß nicht</i> "
	Q2.3a	EN: "Which topics, among others, have been addressed during the conference? – <i>Option A, Option B, I don't know</i> " DE: "Welche Themen wurden u.a. in dem Gespräch angesprochen? – <i>Option A, Option B, Ich weiß nicht</i> "
	Q2.1b ... Q2.3b	EN: "I am xxx % sure of my answer." DE: "Ich bin mir meiner Antwort zu xxx % sicher."
3	Q3.1a ... Q3.10a	EN: per Quotation from transcript: "Who said what? – <i>Mr.X, Mr.Y, I don't know</i> " DE: "Wer sagte was?" – <i>Herr X, Herr Y, Ich weiß nicht</i> "
	Q3.1b ... Q3.10b	EN: "I am xxx % sure of my answer." DE: "Ich bin mir meiner Antwort zu xxx % sicher."
4	Q4.1	EN: "During the conference, it was <i>extremely difficult ... extremely easy</i> to recognize <b>who</b> was speaking." DE: "Es viel mir <i>extrem schwer ... extrem leicht</i> um während der Konferenz zu erkennen, <i>welcher Gesprächspartner gerade spricht</i> ."
	Q4.2	EN: "It was <i>extremely difficult ... extremely easy</i> to follow <b>which opinions</b> were exchanged during the conference" DE: "Es viel mir <i>extrem schwer ... extrem leicht</i> um während der Konferenz zu folgen, <i>welche Meinungen</i> vertreten wurden."
	Q4.3	EN: "It required <i>extremely much ... extremely little</i> <b>concentration</b> to follow the conference" DE: "Es erforderte <i>extrem viel ... extrem wenig</i> <b>Konzentration</b> um der Konferenz zu folgen."
	Q4.4	EN: "The <b>speech intelligibility</b> during the conference was <i>extremely bad ... ideal</i> " DE: "Die <b>Sprachverständlichkeit</b> in der Konferenz war <i>extrem schlecht ... ideal</i> "
	Q4.5	EN: "The <b>naturalness</b> of the conversation was <i>extremely bad ... ideal</i> " DE: "Die <b>Natürlichkeit</b> des Gesprächablaufes war <i>extrem schlecht ... ideal</i> "
	Q4.6	EN: "The <b>efficiency</b> of the conversation was <i>extremely bad ... ideal</i> " DE: "Die <b>Effizienz</b> des des Gesprächablaufes war <i>extrem schlecht ... ideal</i> "
Experiment CC2		
Part	Question ID	Wording / Description, EN: English translation, DE: German original
1	Q1.1	EN: "What was your <b>overall impression</b> of the telephone conference? – <i>extremely bad ... ideal</i> " DE: "Wie war ihr persönlicher <b>Gesamteindruck</b> von dieser Telefonkonferenz? – <i>extrem schlecht ... ideal</i> ."
2	Q2.1a	EN: "In your opinion, who spoke more? – <i>Mr.X, Mr.Y, Both equally</i> " DE: "Ihrer Meinung nach, welcher Gesprächspartner sprach mehr? – <i>Herr X, Herr Y, Beide gleich</i> "
	Q2.2a	EN: "In your opinion, who interrupted more? – <i>Mr.X, Mr.Y, Both equally</i> " DE: "Ihrer Meinung nach, welcher Gesprächspartner unterbrach mehr? – <i>Herr X, Herr Y, Beide gleich</i> "
	Q2.3a	EN: "In your opinion, who made more constructive proposals? – <i>Mr.X, Mr.Y, Both equally</i> " DE: "Ihrer Meinung nach, welcher Gesprächspartner machte mehr konstruktive Vorschläge? – <i>Herr X, Herr Y, Beide gleich</i> "
	Q2.4a	EN: "In your opinion, who criticized more? – <i>Mr.X, Mr.Y, Both equally</i> " DE: "Ihrer Meinung nach, welcher Gesprächspartner äußerte mehr Kritikpunkte? – <i>Herr X, Herr Y, Beide gleich</i> "
	Q2.1b ... Q2.3b	EN: "I am xxx % sure of my answer." DE: "Ich bin mir meiner Antwort zu xxx % sicher."
3	Q3.1a ... Q3.16a	EN: per Quotation from transcript: "Who said what? – <i>Mr.X, Mr.Y, I don't know</i> " DE: "Wer sagte was?" – <i>Herr X, Herr Y, Ich weiß nicht</i> "
	Q3.1b ... Q3.16b	EN: "I am xxx % sure of my answer." DE: "Ich bin mir meiner Antwort zu xxx % sicher."
4	Q4.1	EN: "My personal impression of the <b>quality of the connection</b> is <i>extremely bad ... ideal</i> ." DE: "Mein persönlicher Eindruck von der <b>Qualität der Verbindung</b> ist <i>extrem schlecht ... ideal</i> "
	Q4.2	EN: "During the conference, it was <i>extremely difficult ... extremely easy</i> to recognize <b>who</b> was speaking". DE: "Es viel mir <i>extrem schwer ... extrem leicht</i> während der Konferenz zu erkennen, <b>welcher Gesprächspartner gerade spricht</b> ."
	Q4.3	EN: "It was <i>extremely difficult ... extremely easy</i> to follow <b>which opinions</b> were exchanged during the conference". DE: "Es viel mir <i>extrem schwer ... extrem leicht</i> während der Konferenz zu folgen, <b>welche Meinungen</b> vertreten wurden."
	Q4.4	EN: "It required <i>extremely much ... extremely little</i> <b>concentration</b> to follow the conference". DE: "Es erforderte <i>extrem viel ... extrem wenig</i> <b>Konzentration</b> um der Konferenz zu folgen."
	Q4.5	EN: "The <b>speech intelligibility</b> during the conference was <i>extremely bad ... ideal</i> " DE: "Die <b>Sprachverständlichkeit</b> in der Konferenz war <i>extrem schlecht ... ideal</i> "
5	Q5.1	EN: "Attending a telephone conference with such a system would be <i>extremely unsatisfactory ... extremely satisfactory</i> " DE: "Mit einem solchen System einer Telefonkonferenzschaltung beizuwohnen würde mich <i>extrem unzufrieden ... extrem zufrieden</i> stellen."
	Q5.2	EN: "Attending a telephone conference with such a system would be <i>extremely unpleasant ... extremely pleasant</i> " DE: "Mit einem solchen System einer Telefonkonferenz beizuwohnen würde ich als <i>extrem unangenehm ... extrem angenehm</i> empfinden."
	Q5.3	EN: "Attending a telephone conference with such a system would be <i>extremely unacceptable ... extremely unacceptable</i> " DE: "Mit einem solchen System einer Telefonkonferenz beizuwohnen würde ich <i>extrem inakzeptabel ... extrem akzeptabel</i> finden."

Table 7.4: Summary of the questionnaires used in the two listening-only tests Exp1 & Exp2.

Experiment CC1		
Measured Variable	Measure	Extraction from questionnaire
<i>Cognitive Load</i>	<i>Concentration Effort</i>	Directly measured with Q4.3
	<i>Speaker Recognition Effort</i>	Directly measured with Q4.1
	<i>Speaker Recognition Performance</i>	Percentage of correct answers to Questions Q3.1a – Q3.10a
	<i>Speaker Recognition Confidence</i>	Average percentage of answers to Questions Q3.1b – Q3.10b
	<i>Topic Comprehension Effort</i>	Directly measured with Q4.2
	<i>Topic Recognition Performance</i>	Percentage of correct answers to Questions Q2.1a – Q2.3a
	<i>Topic Recognition Confidence</i>	Average percentage of answers to Questions Q2.1b – Q2.3b
<i>Speech Communication Quality with focus on Telecommunication Component</i>	<i>Speech Intelligibility</i>	Directly measured with Q4.4
<i>Speech Communication Quality with focus on Group-Communication Component</i>	<i>Conversation Naturalness</i>	Directly measured with Q4.5
	<i>Conversation Efficiency</i>	Directly measured with Q4.6
<i>Speech Communication Quality and/or Quality of Experience<sup>1</sup></i>	<i>Quality</i>	Directly measured with Q1.1
Remarks: <sup>1</sup> In retrospect, the wording of question Q1.1 allows both interpretations. Hence a unique assignment to either Measured Variable is not possible.		
Experiment CC2		
Measured Variable	Measure	Extraction from questionnaire
<i>Cognitive Load</i>	<i>Concentration Effort</i>	Directly measured with Q4.4
	<i>Speaker Recognition Effort</i>	Directly measured with Q4.2
	<i>Speaker Recognition Performance</i>	Percentage of correct answers to Questions Q3.1a – Q3.16a
	<i>Speaker Recognition Confidence</i>	Average percentage of answers to Questions Q3.1b – Q3.16b
	<i>Focal Assurance</i>	Average percentage of answers to Questions Q2.1b – Q2.4b
<i>Speech Communication Quality with focus on Telecommunication Component Quality of Experience</i>	<i>Topic Comprehension Effort</i>	Directly measured with Q4.3
	<i>Connection Quality</i>	Directly measured with Q4.1
	<i>Speech Intelligibility</i>	Directly measured with Q4.5
	<i>Overall Quality</i>	Directly measured with Q1.1
	<i>Satisfaction</i>	Directly measured with Q5.1
	<i>Pleasantness</i>	Directly measured with Q5.2
	<i>Acceptance</i>	Directly measured with Q5.3

Table 7.5: Relation between the Measured Variables (on their conceptual level), the actual measures used for obtaining respective quantitative data, and extraction of the measures from the questionnaire answers.

and their roles, another measure is *Focal Assurance*, which was applied in Experiment CC2. Four questions (Q2.1a – Q2.4a) referred to who of the interlocutors spoke more, who interrupted more, who made more constructive proposals, and who was more criticizing. Each of these questions was followed by a question asking how confident test participants were in their judgment (Questions Q2.1b – Q2.4b), expressed in percent. These questions about the confidence ask for focal assurance ratings<sup>35</sup>, that is, the strength of cues helping test subjects to assess the involvement of individual interlocutors, and helping to form a sense of who the interlocutors are. *Focal Assurance* was then computed as the mean of the four confidence ratings.

Another facet of cognitive load is the mental task of information processing, discussed in Section 7.2. Experiment CC1 employed three measures. The first measure is *Topic Comprehension Effort*, obtained from test participants' ratings of question Q4.2, and question Q4.3 in Experiment CC2, respectively. The other two measures, *Topic Recognition Performance* and *Topic Recognition Confidence*, are extracted similarly to the speaker recognition performance measures. Three questions (Q2.1a – Q2.3a) referred to three pairs of statements concerning the topics discussed in the conversation. In each pair of statements, one statement concerned a topic that was indeed discussed in the conversation, the other statement concerned a topic that was not discussed at all. For each pair of statements, test participants could answer which statement was discussed in the conversation, or they could choose the option "I don't know". Furthermore, three additional questions (Q2.1b – Q2.3b) asked for each pair of statement, how confident test participants were in their judgment, expressed in percent. *Topic Recognition Performance* is computed as the percentage of correctly identified statements (Q2.1a – Q2.3a); *Speaker Recognition Confidence* is computed as the mean of the confidence ratings (Q2.1b – Q2.3b). For Experiment CC2, the author decided to skip the measures *Topic Recognition Performance* and *Speaker Recognition Confidence*, in order to include the above described measure of *Focal Assurance* without lengthening the questionnaire.

*Speech Communication Quality* As described in the hypotheses definitions, the variable *Speech Communication Quality* has different emphases depending on the actual questions used in the two experiments.

In Experiment CC1, two measures focused on the *Group-Communication Component* of *Speech Communication Quality*. The first measure is *Conversation Naturalness*, obtained from test participants' ratings of question Q4.5; the second measure is *Conversation Efficiency*, obtained from ratings of question Q4.6.

Next, two further measures focused on the *Telecommunication Component* of *Speech Communication Quality*. The first measure is *Connection Quality*, applied in Experiment CC2, and obtained from ratings of question Q4.1. The assignment of this question to the *Telecommunication Component* is motivated by the specific wording of question

<sup>35</sup> Jessica J. Baldis. "Effects of Spatial Audio on Memory, Comprehension, and Preference during Desktop Conferences". In: *Proceedings of the ACM CHI 2001 Human Factors in Computing Systems Conference*. Ed. by Michel Beaudouin-Lafon and Robert J. K. Jacob. Vol. 3. 1. 2001, pp. 166–173

Q4.1, which highlighted the keyword *quality of the connection*. The second measure is *Speech Intelligibility*, obtained from test participants' ratings of question Q4.4 in Experiment CC1 and question Q4.5 in Experiment CC2. The assignment of this question to either the *Telecommunication Component* or the *Group-Communication Component* is not straight-forward. On the one hand speech intelligibility relates to the speech signal's quality in terms of distorted parts of speech, on the other hand it relates to the ability to communicate over the system. To disambiguate this, the author applies the terminology that Raake<sup>36</sup> extracted from related literature<sup>37</sup>, distinguishing between *Comprehensibility*, *Intelligibility*, and *Communicability*. In this terminology, *Intelligibility* "refers to how well the content of an utterance [...] can be identified on the basis of the form", which in turn is separated from *Communicability*, which "means that a speech message is such that it can serve to communicate". Considering this, the author sees *Speech Intelligibility* as a measure focusing on the *Telecommunication Component*, acknowledging that test participants are not aware of such fine distinctions and might also take communicative aspects into account when answering the corresponding question.

*Quality of Experience* As already mentioned, Experiment CC1 did not specifically aim at distinguishing between the two layers *Speech Communication Quality* and *Quality of Experience*. The experiment applied one measure *Quality*, obtained from participants' ratings of question Q1.1. The wording of that question ("overall impression of the connection") actually allows an interpretation in terms of *Speech Communication Quality* due to the keyword *connection* or in terms of *Quality of Experience* due to the keyword *overall impression*. For that reason, the corresponding hypothesis for Experiment CC1 does not distinguish between *Speech Communication Quality* and *Quality of Experience*, see Table 7.1.

Experiment CC2, however, aimed at a more precise distinction between these two layers of quality. For that reason, the question was rephrased such that the emphasis was now on quality in the broader sense of *Quality of Experience*, by using "overall impression of the telephone conference". Thus, Experiment CC2 used a measure *Overall Quality*, obtained from participants' ratings of the updated question Q1.1, which differs from the measure of *Connection Quality* (question Q4.5) described in the previous paragraph.

Given the remaining risk that test subjects interpret this question nevertheless as *Speech Communication Quality*, the author included three more measures addressing service satisfaction: *Satisfaction*, obtained from participants' ratings of question Q5.1; *Pleasantness*, obtained from ratings of question Q5.2; and *Acceptance*, obtained from ratings of question Q5.3.

*Data Processing* All measures were represented in such a way that high values reflect something positive, that is high quality or low cognitive load. This representation is straight-forward for the con-

<sup>36</sup> Alexander Raake. *Speech Quality of VoIP – Assessment and Prediction*. Chichester, West Sussex, UK: Wiley, 2006

<sup>37</sup> Ute Jekosch. "Sprache hören und beurteilen: Ein Ansatz zur Grundlegung der Sprachqualitätsbeurteilung". Habilitation thesis. Universität/Gesamthochschule Essen, Germany, 2003

Ute Jekosch. *Voice and Speech Quality Perception – Assessment and Evaluation*. Berlin, Germany: Springer, 2005

Sebastian Möller. *Assessment and Prediction of Speech Quality in Telecommunications*. Kluwer Academic publishers, 2000

cepts related with most measures (*Overall Quality, Quality, Connection Quality, Speech Intelligibility, Speaker Recognition Performance, Speaker Recognition Confidence, Topic Recognition Performance, Topic Recognition Confidence, Focal Assurance, Conversation Naturalness, Conversation Efficiency, Satisfaction, Pleasantness, Acceptance*). However, it is less intuitive for the effort-related measures (*Speaker Recognition Effort, Topic Comprehension Effort, Concentration Effort*), as the ratings are now inverted: a high effort, meaning a high cognitive load, is coded with a low value.

*Data Validation* A proper choice of quotations used in the memory tests (Q3.1a to Q3.10a in Experiment CC1, Q3.1a to Q3.16a in Experiment CC2) is necessary in order to avoid that any systematic influences skew the results. For that reason, the author inspected the data for quotations with small or zero deviation in the recall performance across subjects. Table 7.6 shows the statistics of that inspection.

Concerning questions with correct answers in almost all cases for Experiment CC1, i.e. 12 or 13 participants were correct, the table shows that most of those questions belong to the two- or three-interlocutor scenarios. Similarly, for Experiment CC2, i.e. 24 or 25 participants were correct, almost all of those questions belong to the two-interlocutor scenarios. Apparently, this can be explained by the expected easiness of memorizing the least complex two-interlocutor scenarios, and – for the participants of Experiment CC1 – also the three-interlocutor scenarios. Concerning the remaining questions that were almost always correctly answered, an analysis in terms of the meaningfulness of these quotations did not reveal any error or insight into why this happened. The same holds for the questions with incorrect answers in almost all cases, i.e. 12 or 13 participants were incorrect in Experiment CC1, 24 or 25 participants were incorrect in Experiment CC2. From these results the author concluded that the used quotations did not introduce any systematic errors.

Furthermore, inspecting the other questions (Q1.1, Q2.x, Q4.x, Q5.x) did not reveal any specific outliers in terms of individual ratings or test participants. Hence none of the data points were excluded from the further analysis.

*Data Analysis and Presentation* Since the experiments had two test factors, number of interlocutors #IL and sound reproduction method *SndRepr*, the analysis comprises two levels. The first analysis level concerns the main effects; that means the effect of #IL if all data over *SndRepr* is compiled and the effect of *SndRepr* if all data over #IL is compiled. The second analysis level concerns the individual effects; that means the effect of #IL per *SndRepr* and the effect of *SndRepr* per #IL. As analysis methods, the author conducted repeated-measures ANOVAs with Greenhouse-Geisser correction when the sphericity condition was violated<sup>38</sup> to check for significant differences of the *Manipulated Variables* #IL and *SndRepr* on the different *Measured Vari-*

<sup>38</sup> Andy Field. *Discovering Statistics using SPSS*. 3rd ed. SAGE publications, 2009

Experiment 1, 120 questions in total (12 scenarios × 10 quotes), 13 participants					
Case	# Questions in Scenarios with #IL equal to				
	2	3	4	6	total
13 participants correct	5	3	1	1	10
12 participants correct and 1 participant wrong	2	4	1	2	9
12 participants correct and 1 participant "I don't know"	0	0	0	0	0
total of almost always correct answers	7	7	2	3	19 (16%)
12 participants wrong or "I don't know" and 1 participant correct	0	0	0	1	1
13 participants wrong or "I don't know"	0	0	0	1	1
total of almost always incorrect answers	0	0	0	1	2 (<1%)
Experiment 2, 192 questions in total (12 scenarios × 16 quotes), 25 participants					
Case	# Questions in Scenarios with #IL equal to				
	2	3	4	6	total
25 participants correct	2	0	0	1	3
24 participants correct and 1 participant wrong	6	1	0	0	7
24 participants correct and 1 participant "I don't know"	4	0	0	2	6
total of almost always correct answers	12	1	0	3	16 (8%)
24 participants wrong or "I don't know" and 1 participant correct	0	0	0	1	1
25 participants wrong or "I don't know"	0	0	0	0	0
total of almost always incorrect answers	0	0	0	1	1 (<1%)

ables. Then the author conducted PostHoc tests (estimated marginal means) with Sidak correction to identify between which pairs of *Manipulated Variables* significant differences can be observed. And the author generated errorbar plots, showing the mean values and 95% confidence intervals, to visualize the effects. Thus, the analysis consisted of nine steps per measure: ANOVA, PostHoc and Errorbar for the main effects, for the individual effect of #IL and for the individual effect of *SndRepr*.

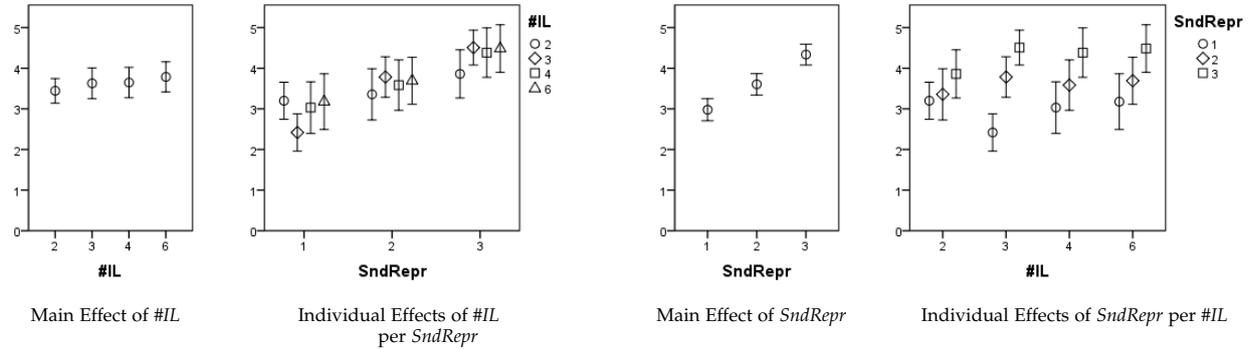
The next two sections present a summary of the results for the two experiments CC1 and CC2, focusing on explaining the main conclusions from the analysis, and visualizing the data with the errorbar plots. The full analysis details with all ANOVA and PostHoc results are presented in Appendix C.1 for Experiment CC1 and Appendix C.2 for Experiment CC2.

Table 7.6: Statistics on quotations with small or zero deviation in performance across subjects.

7.3.5 Results Experiment CC1

Quality

1. Errorbars (mean & 95% confidence interval)



2. Effect of #IL

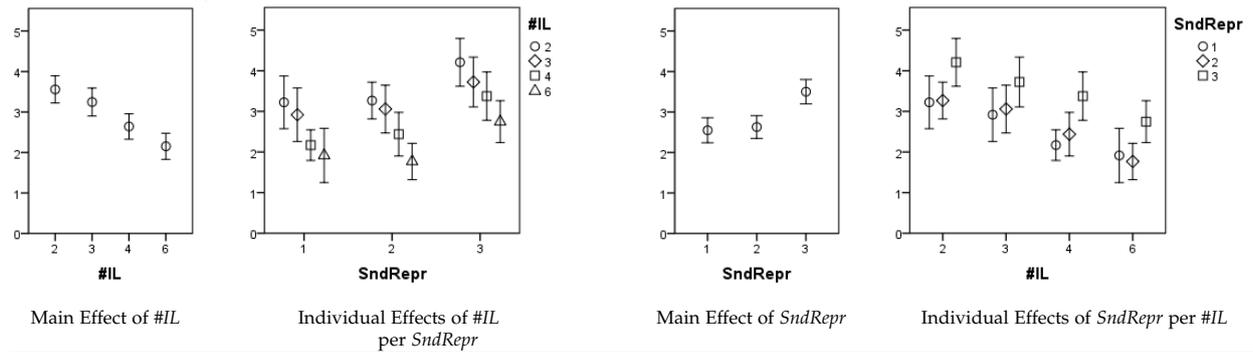
Hypothesis	Main effect	Effect per SndRepr
H5	×	×

3. Effect of SndRepr

Hypothesis	Main effect	Effect per #IL
H6	✓	✓

Concentration Effort

1. Errorbars (mean & 95% confidence interval)



2. Effect of #IL

Hypothesis	Main effect	Effect per SndRepr
H1	✓	✓ (#IL= 2,3 vs. #IL = 4,6)

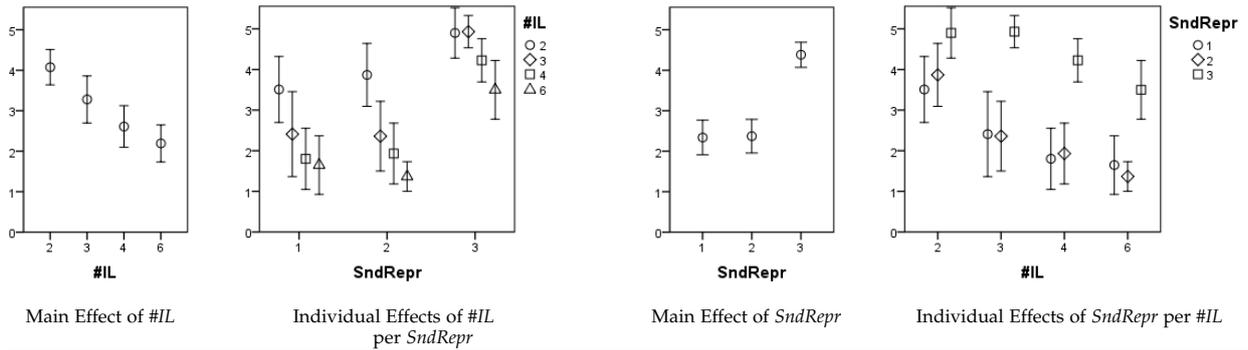
3. Effect of SndRepr

Hypothesis	Main effect	Effect per #IL
H2	✓ (SndRepr = 1,2 vs. = 3)	✓ (SndRepr = 1,2 vs. =3)

Figure 7.4: Analysis summary for Experiment 1 for the measures Quality and Concentration Effort.

**Speaker Recognition Effort**

1. Errorbars (mean & 95% confidence interval)



2. Effect of #IL

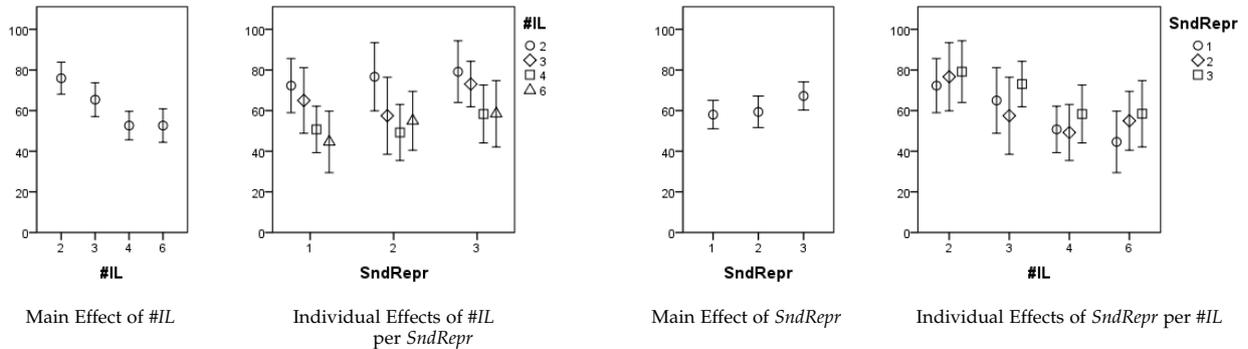
Hypothesis	Main effect	Effect per SndRepr
H1	✓	✓ (#IL = 2,3 vs. #IL = 4,6)

3. Effect of SndRepr

Hypothesis	Main effect	Effect per #IL
H2	✓ (SndRepr = 1,2 vs. 3)	✓ (SndRepr = 1,2 vs. 3)

**Speaker Recognition Performance**

1. Errorbars (mean & 95% confidence interval)



2. Effect of #IL

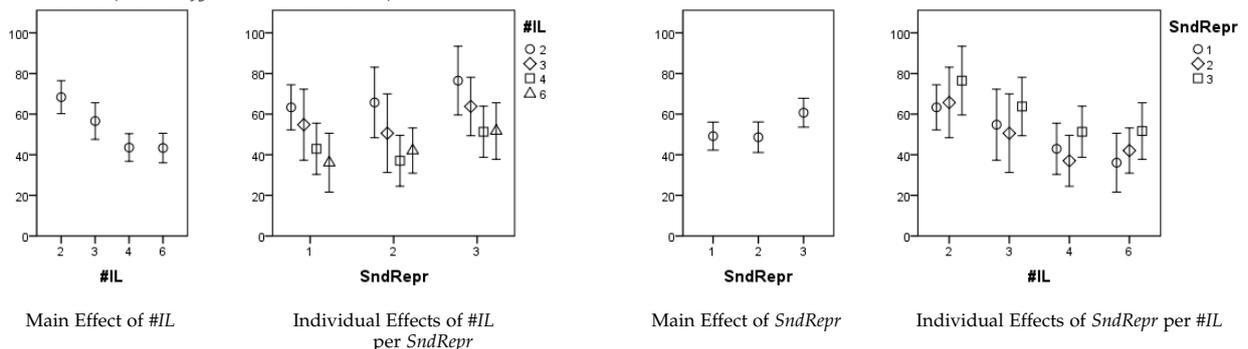
Hypothesis	Main effect	Effect per SndRepr
H1	✓	✓ (#IL = 2 vs. #IL ≥ 3)

3. Effect of SndRepr

Hypothesis	Main effect	Effect per #IL
H2	×	×

**Speaker Recognition Confidence**

1. Errorbars (mean & 95% confidence interval)



2. Effect of #IL

Hypothesis	Main effect	Effect per SndRepr
H1	✓	✓ (#IL = 2 vs. #IL ≥ 3)

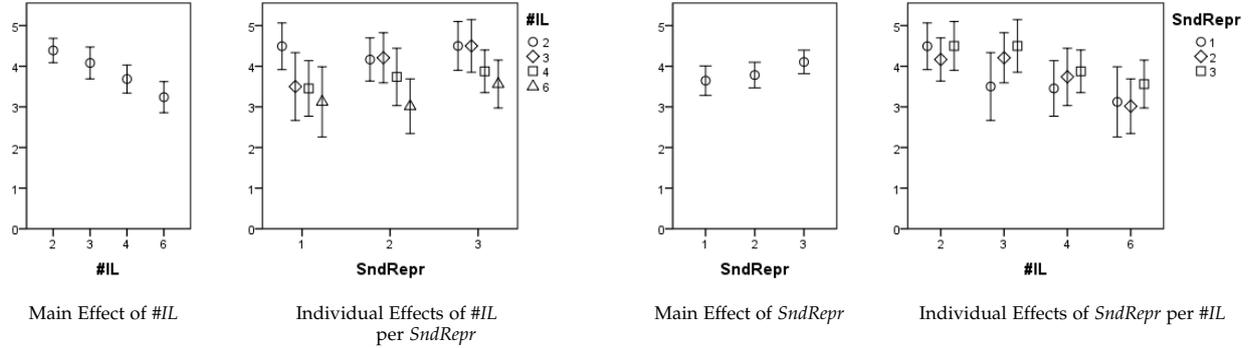
3. Effect of SndRepr

Hypothesis	Main effect	Effect per #IL
H2	✓ (SndRepr = 1,2 vs. 3)	✓ (SndRepr = 2 vs. 3, for #IL=6)

Figure 7.5: Analysis summary for Experiment 1 for the measures Speaker Recognition Effort, Speaker Recognition Performance, and Speaker Recognition Confidence.

**Topic Comprehension Effort**

1. Errorbars (mean & 95% confidence interval)



2. Effect of #IL

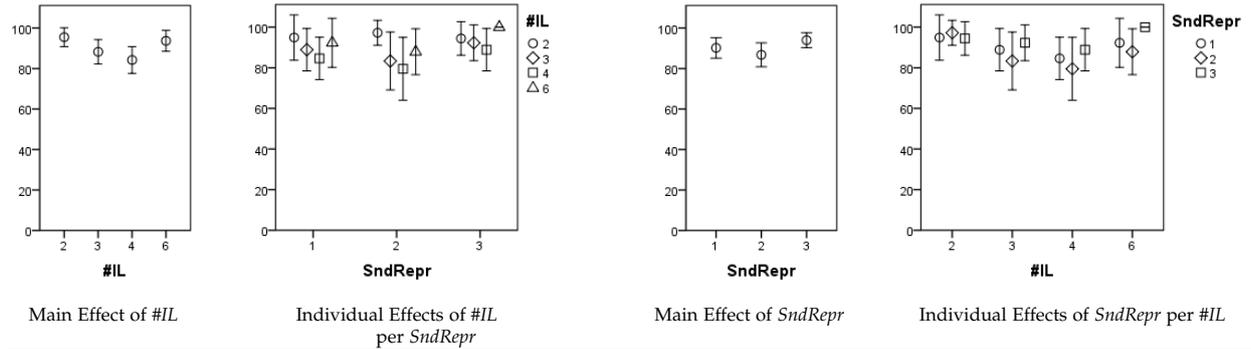
Hypothesis	Main effect	Effect per SndRepr
H1	✓ (#IL = 2 vs. 6)	✓

3. Effect of SndRepr

Hypothesis	Main effect	Effect per #IL
H2	✓	×

**Topic Recognition Performance**

1. Errorbars (mean & 95% confidence interval)



2. Effect of #IL

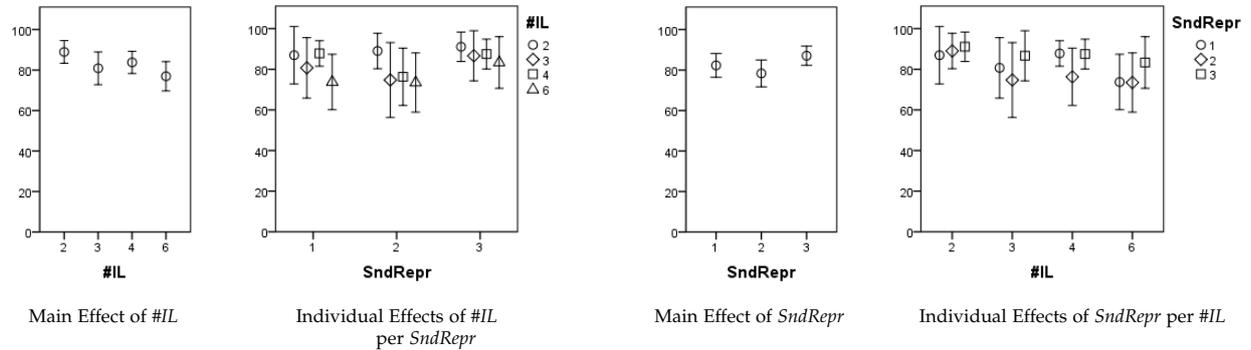
Hypothesis	Main effect	Effect per SndRepr
H1	×	×

3. Effect of SndRepr

Hypothesis	Main effect	Effect per #IL
H2	✓	×

**Topic Recognition Confidence**

1. Errorbars (mean & 95% confidence interval)



2. Effect of #IL

Hypothesis	Main effect	Effect per SndRepr
H1	✓	✓ (for SndRepr = 1)

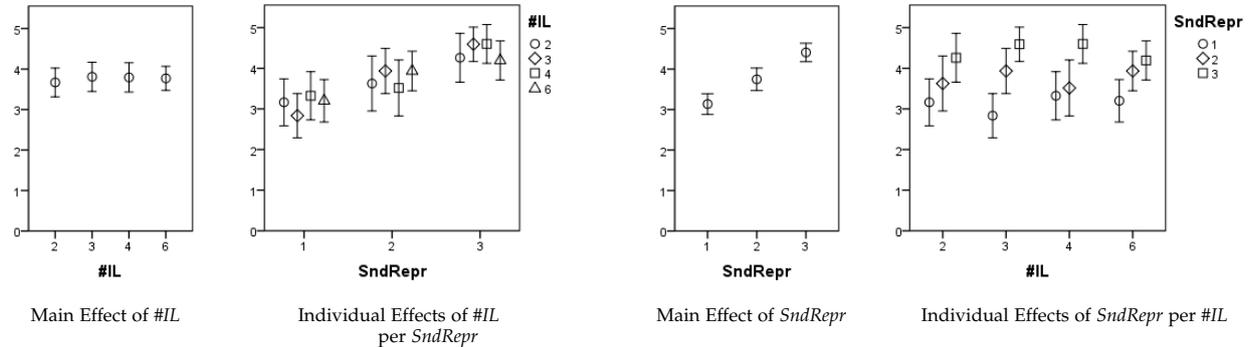
3. Effect of SndRepr

Hypothesis	Main effect	Effect per #IL
H2	✓ (SndRepr = 2 vs. 3)	×

Figure 7.6: Analysis summary for Experiment 1 for the measures Topic Comprehension Effort, Topic Recognition Performance, and Topic Recognition Confidence.

Speech Intelligibility

1. Errorbars (mean & 95% confidence interval)



2. Effect of #IL

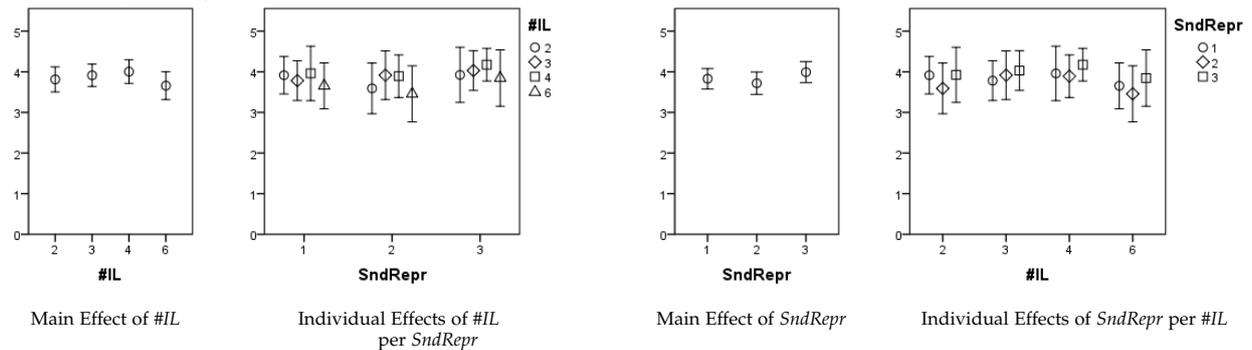
Hypothesis	Main effect	Effect per SndRepr
H3.1	×	×

3. Effect of SndRepr

Hypothesis	Main effect	Effect per #IL
H4.1	✓	✓

Conversation Naturalness

1. Errorbars (mean & 95% confidence interval)



2. Effect of #IL

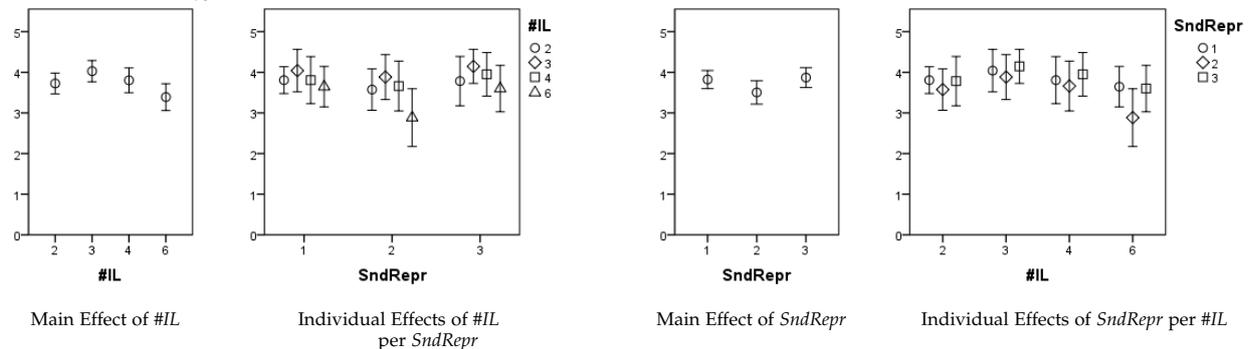
Hypothesis	Main effect	Effect per SndRepr
H3.2	×	×

3. Effect of SndRepr

Hypothesis	Main effect	Effect per #IL
H4.2	×	×

Conversation Efficiency

1. Errorbars (mean & 95% confidence interval)



2. Effect of #IL

Hypothesis	Main effect	Effect per SndRepr
H3.2	✓ (not PostHoc)	✓ (for SndRepr = 2, not PostHoc)

3. Effect of SndRepr

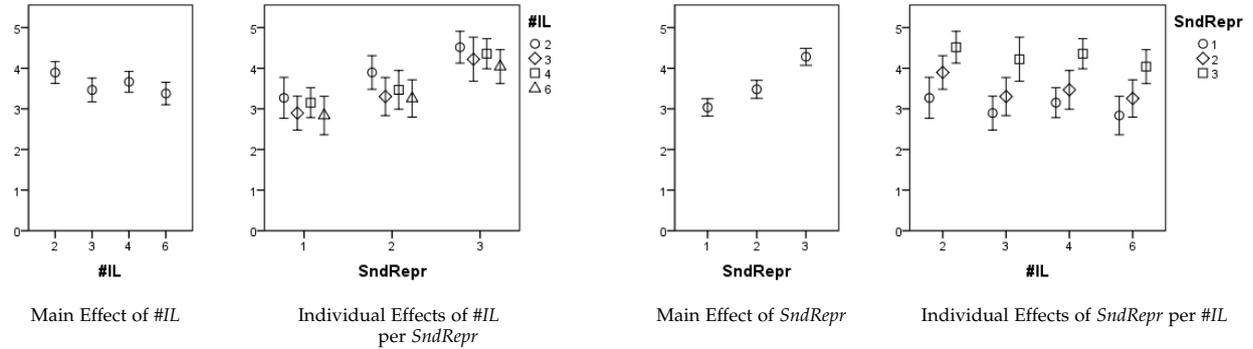
Hypothesis	Main effect	Effect per #IL
H4.2	✓ (SndRepr = 2 vs. 3, not PostHoc)	×

Figure 7.7: Analysis summary for Experiment 1 for the measures Speech Intelligibility, Conversation Naturalness, and Conversation Efficiency.

7.3.6 Results Experiment CC2

Overall Quality

1. Errorbars (mean & 95% confidence interval)



2. Effect of #IL

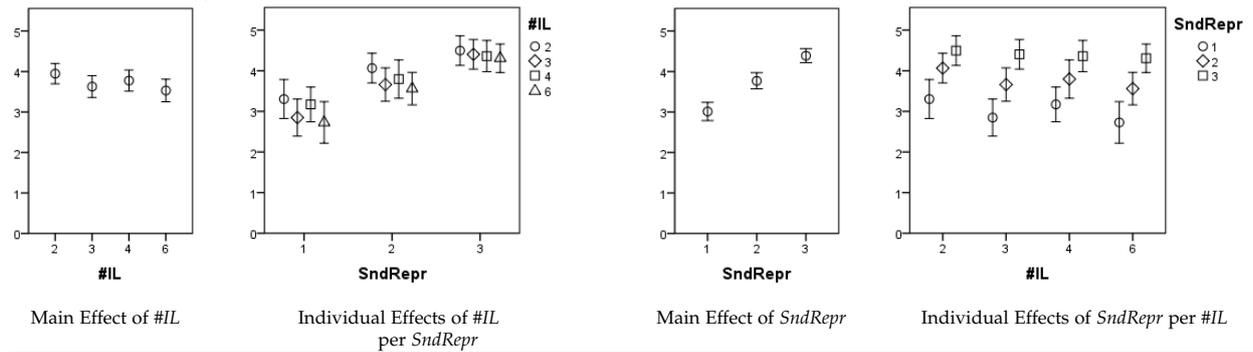
Hypothesis	Main effect	Effect per SndRepr
H5	✓ (#IL = 2 vs. 6)	✓ (for SndRepr = 2)

3. Effect of SndRepr

Hypothesis	Main effect	Effect per #IL
H6	✓ (SndRepr = 1,2 vs. 3)	✓ (SndRepr = 1,2 vs. 3)

Connection Quality

1. Errorbars (mean & 95% confidence interval)



2. Effect of #IL

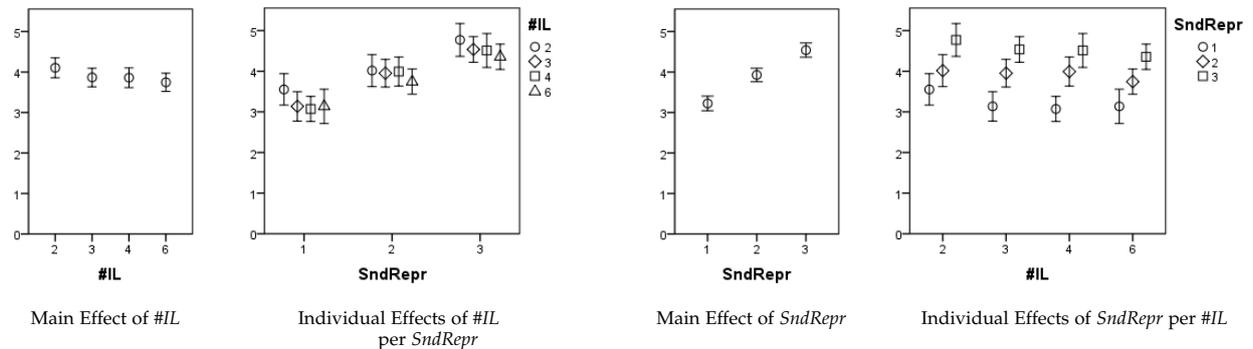
Hypothesis	Main effect	Effect per SndRepr
H3	×	×

3. Effect of SndRepr

Hypothesis	Main effect	Effect per #IL
H4	✓	✓

Intelligibility

1. Errorbars (mean & 95% confidence interval)



2. Effect of #IL

Hypothesis	Main effect	Effect per SndRepr
H3	×	×

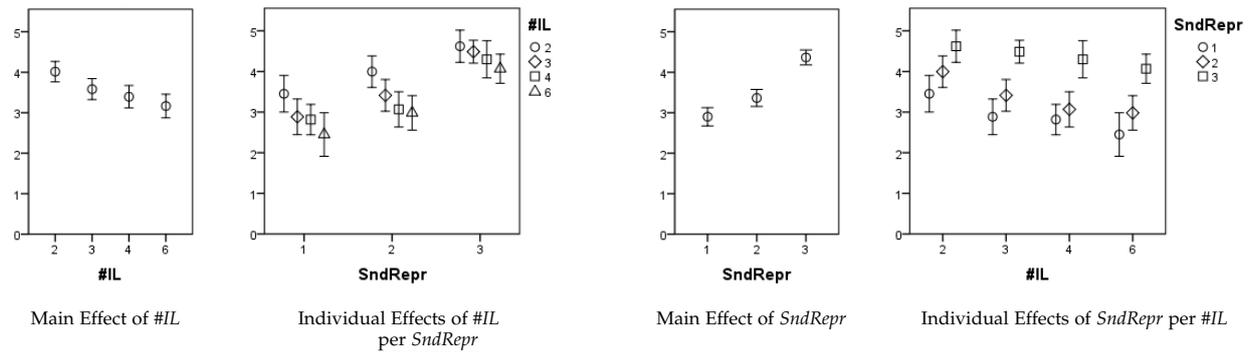
3. Effect of SndRepr

Hypothesis	Main effect	Effect per #IL
H4	✓	✓

Figure 7.8: Analysis summary for Experiment 2 for the measures Overall Quality, Connection Quality, and Intelligibility.

**Satisfaction**

1. Errorbars (mean & 95% confidence interval)



2. Effect of #IL

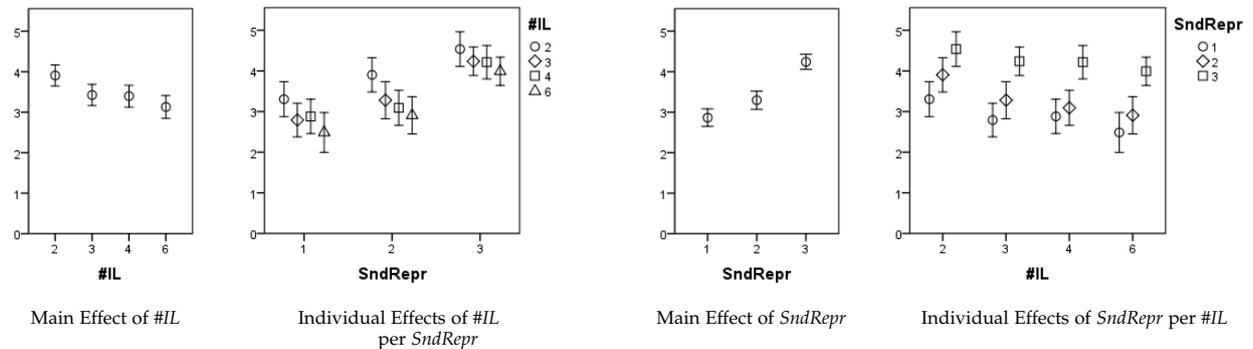
Hypothesis	Main effect	Effect per SndRepr
H5	✓ (#IL = 2 vs. #IL ≥ 3)	✓ (#IL = 2 vs. #IL ≥ 3, for SndRepr = 1,2)

3. Effect of SndRepr

Hypothesis	Main effect	Effect per #IL
H6	✓	✓

**Pleasantness**

1. Errorbars (mean & 95% confidence interval)



2. Effect of #IL

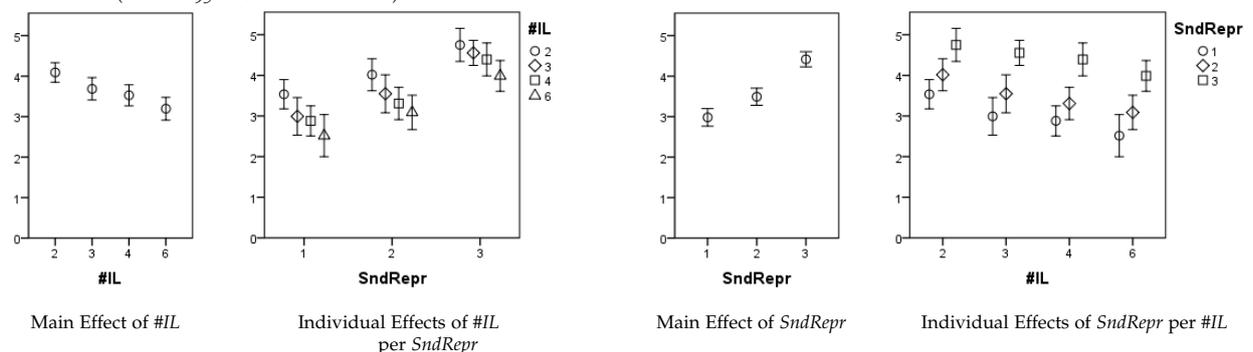
Hypothesis	Main effect	Effect per SndRepr
H5	✓ (#IL = 2 vs. #IL ≥ 3)	✓ (#IL = 2 vs. #IL ≥ 3, for SndRepr = 2)

3. Effect of SndRepr

Hypothesis	Main effect	Effect per #IL
H6	✓	✓

**Acceptance**

1. Errorbars (mean & 95% confidence interval)



2. Effect of #IL

Hypothesis	Main effect	Effect per SndRepr
H5	✓ (#IL = 2 vs. #IL ≥ 3)	✓ (#IL = 2 vs. #IL ≥ 3, for SndRepr = 1,2)

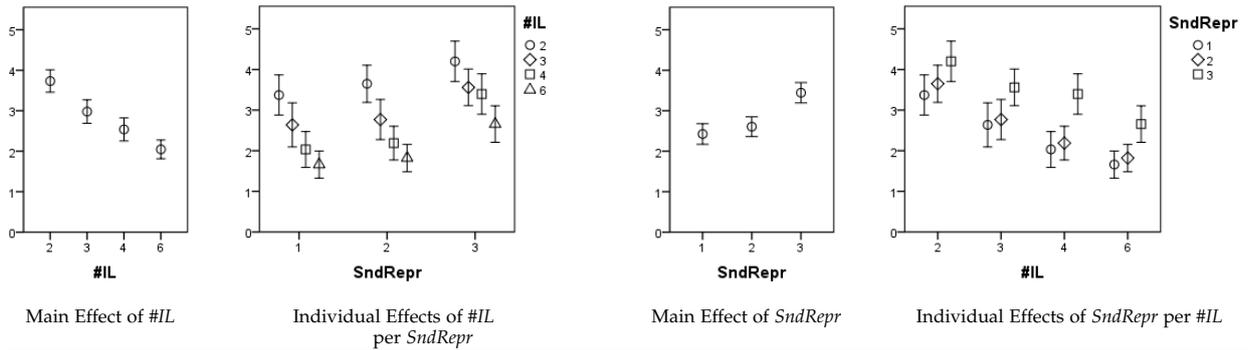
3. Effect of SndRepr

Hypothesis	Main effect	Effect per #IL
H6	✓	✓

Figure 7.9: Analysis summary for Experiment 2 for the measures Satisfaction, Pleasantness, and Acceptance.

**Concentration Effort**

1. Errorbars (mean & 95% confidence interval)



2. Effect of #IL

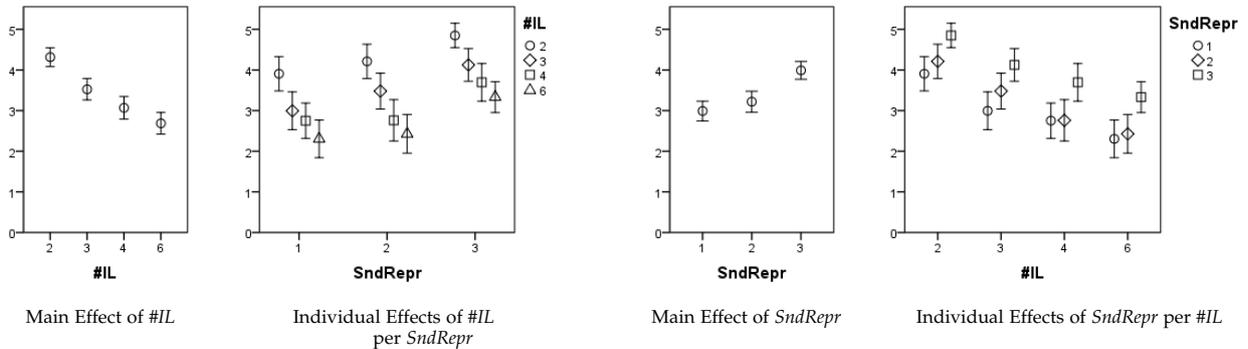
Hypothesis	Main effect	Effect per SndRepr
H1	✓	✓

3. Effect of SndRepr

Hypothesis	Main effect	Effect per #IL
H2	✓ (SndRepr = 1,2 vs. 3)	✓ (SndRepr = 1,2 vs. 3)

**Topic Comprehension Effort**

1. Errorbars (mean & 95% confidence interval)



2. Effect of #IL

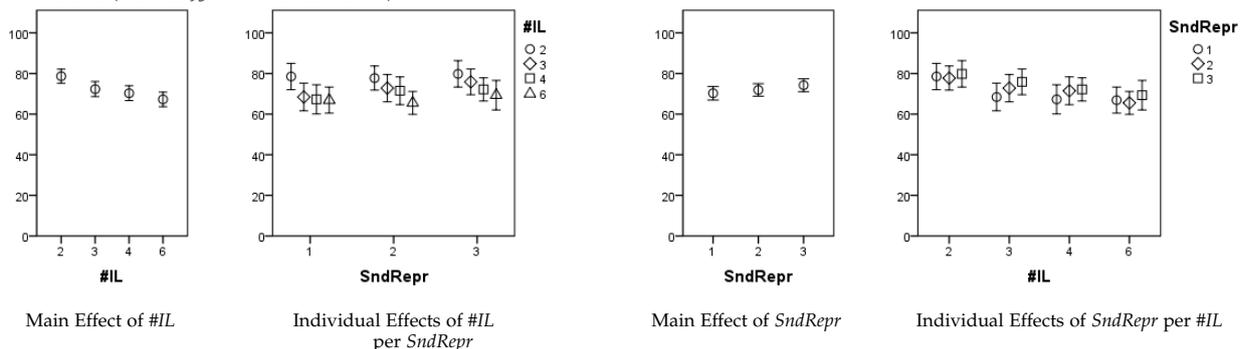
Hypothesis	Main effect	Effect per SndRepr
H1	✓	✓

3. Effect of SndRepr

Hypothesis	Main effect	Effect per #IL
H2	✓ (SndRepr = 1,2 vs. 3)	✓ (SndRepr = 1,2 vs. 3)

**Focal Assurance**

1. Errorbars (mean & 95% confidence interval)



2. Effect of #IL

Hypothesis	Main effect	Effect per SndRepr
H1	✓ (#IL = 2 vs. #IL = 4,6)	✓

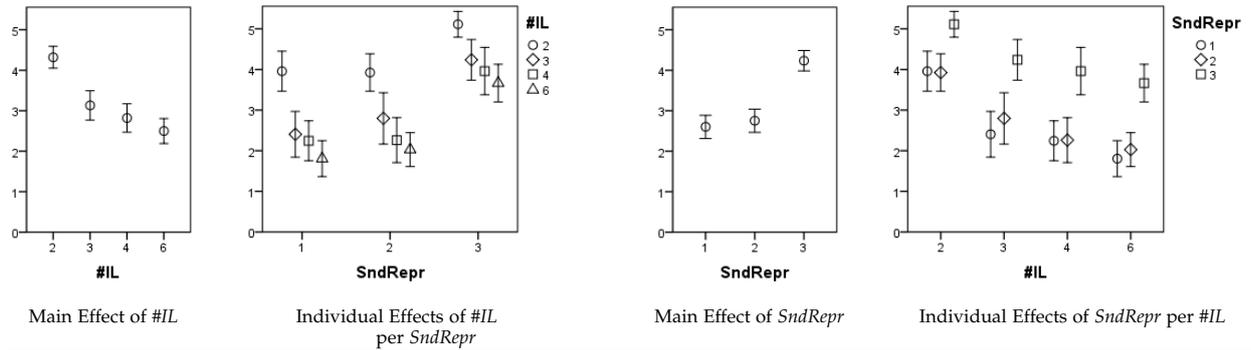
3. Effect of SndRepr

Hypothesis	Main effect	Effect per #IL
H2	×	×

Figure 7.10: Analysis summary for Experiment 2 for the measures Concentration Effort, Topic Comprehension Effort, and Focal Assurance.

**Speaker Recognition Effort**

1. Errorbars (mean & 95% confidence interval)



2. Effect of #IL

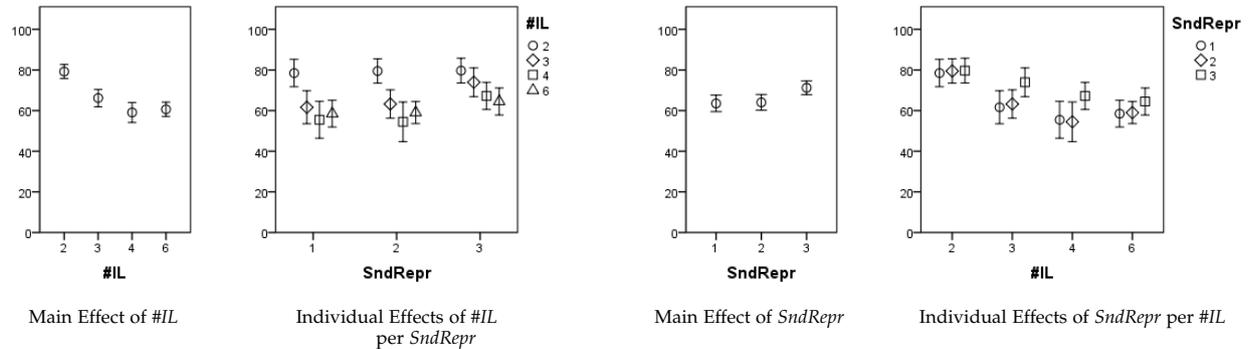
Hypothesis	Main effect	Effect per SndRepr
H1	✓ (#IL = 2 vs. #IL ≥ 3)	✓ (#IL = 2 vs. #IL = 4,6)

3. Effect of SndRepr

Hypothesis	Main effect	Effect per #IL
H2	✓ (SndRepr = 1,2 vs. 3)	✓ (SndRepr = 1,2 vs. 3)

**Speaker Recognition Performance**

1. Errorbars (mean & 95% confidence interval)



2. Effect of #IL

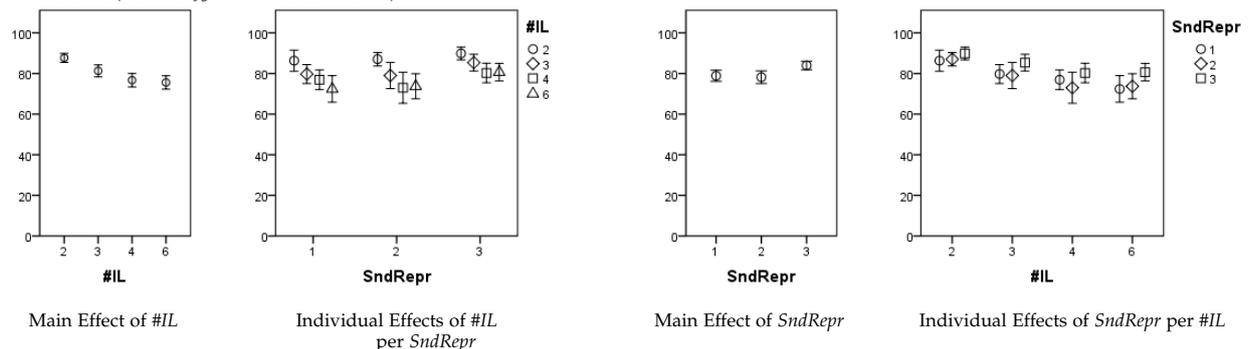
Hypothesis	Main effect	Effect per SndRepr
H1	✓ (#IL = 2 vs. #IL ≥ 3)	✓ (#IL = 2 vs. #IL ≥ 3)

3. Effect of SndRepr

Hypothesis	Main effect	Effect per #IL
H2	✓ (SndRepr = 1,2 vs. 3)	✓ (SndRepr = 1,2 vs. 3)

**Speaker Recognition Confidence**

1. Errorbars (mean & 95% confidence interval)



2. Effect of #IL

Hypothesis	Main effect	Effect per SndRepr
H1	✓ (#IL = 2 vs. #IL ≥ 3)	✓ (#IL = 2 vs. #IL = 4,6)

3. Effect of SndRepr

Hypothesis	Main effect	Effect per #IL
H2	✓ (SndRepr = 1,2 vs. 3)	✓ (SndRepr = 2 vs. 3, for #IL=6)

Figure 7.11: Analysis summary for Experiment 2 for the measures *Speaker Recognition Effort*, *Speaker Recognition Performance*, and *Speaker Recognition Confidence*.

## 7.3.7 Discussion

*Review of Results* Table 7.7 summarizes the main findings, which will be discussed now.

Exp.	Measured Variables	Manipulated Variables	
		Communication Complexity (i.e. Number of Interlocutors #IL)	Technical System Capability (i.e. Sound Reproduction Method <i>SndRepr</i> )
CC1	<i>Cognitive Load</i>	Hypothesis H1 ✓	Hypothesis H2 ✓ <sup>1</sup>
	<i>Speech Communication Quality</i> with focus on <i>Telecommunication Component</i>	Hypothesis H3.1 ×	Hypothesis H4.1 ×
	<i>Speech Communication Quality</i> with focus on <i>Group-Communication Component</i>	Hypothesis H3.2 × <sup>3</sup>	Hypothesis H4.2 × <sup>3</sup>
	<i>Speech Communication Quality</i> and/or <i>Quality of Experience</i>	Hypothesis H5 ×	Hypothesis H6 ✓
CC2	<i>Cognitive Load</i>	Hypothesis H1 ✓	Hypothesis H2 ✓ <sup>1</sup>
	<i>Speech Communication Quality</i>	Hypothesis H3 ×	Hypothesis H4 ✓
	<i>Quality of Experience</i>	Hypothesis H5 ✓ <sup>2</sup>	Hypothesis H6 ✓

Remarks:

✓<sup>1</sup> hypothesis confirmed for comparison spatial audio vs. non-spatial audio condition, not for narrowband vs. fullband.

✓<sup>2</sup> confirmed for comparison pairs #IL =2 vs. #IL ≥ 3 in case of non-spatial sound reproduction.

×<sup>3</sup> inconsistent results, as only a part of the measures showed effects

Hypotheses H1 is confirmed in both experiments, which means that the experimental manipulation concerning the *Communication Complexity* was successful, as the recordings with different degrees of *Communication Complexity* changed the *Cognitive Load* as intended. In other words, H1 confirmed an expected effect.

Hypothesis H2 is confirmed in both experiments due to the effect of the spatial vs. non-spatial sound reproduction method, but there is no effect due to the signal bandwidth (narrow-band vs. full-band conditions). Two conclusions can be drawn from this finding. First, this study contributes to the discussion in the literature on the potential benefit of spatial sound reproduction for multiparty teleconferencing in terms of cognitive load, that is beyond quality. This study supports findings in related work<sup>39</sup> that there is such a benefit, although sometimes this could be confirmed only for self-reported measures, but not for the memory test results<sup>40</sup>. Second, this study shows that only certain system characteristics (here spatial sound reproduction) can lead to a significant change of *Cognitive Load*. In fact the author expected that also bandwidth would have an influence, arguing that fullband signals should – with their larger frequency range – provide more cues to identify speakers and to better understand speech than narrowband signals, which in turn would have an effect on the *Cognitive Load*. If such an effect of audio bandwidth exists, it was not significant in these experiments.

Hypothesis H3.1 is rejected in Experiment CC1, H3.2 could not be consistently verified in Experiment CC1 but it is in parts rejected, and H3 is rejected in Experiment CC2. This means, people were

Table 7.7: Overview of the *Manipulated and Measured Variables* and the results for the six hypotheses H1 to H6. Confirmed hypotheses are denoted with a ✓, rejected hypotheses with a ×.

<sup>39</sup> Jessica J. Baldis. "Effects of Spatial Audio on Memory, Comprehension, and Preference during Desktop Conferencing". In: *Proceedings of the ACM CHI 2001 Human Factors in Computing Systems Conference*. Ed. by Michel Beaudouin-Lafon and Robert J. K. Jacob. Vol. 3. 1. 2001, pp. 166–173

Alexander Raake et al. "Listening and conversational quality of spatial audio conferencing". In: *Proceedings of the AES 40th International Conference*. Tokyo, Oct. 2010

<sup>40</sup> Ryan Kilgore, Mark Chignell, and Paul Smith. "Spatialized audioconferencing: what are the benefits?" In: *Proceedings of the 2003 conference of the Centre for Advanced Studies on Collaborative research CASCON '03*. IBM Press, 2003, pp. 135–144

able to form a judgment about the system's technical quality (given the emphasis on *Speech Communication Quality*) independently from *Communication Complexity*, and – given the confirmation of H4 (see below) – rather independently from the *Cognitive Load* that the people experienced.

Hypothesis H4.1 and in parts H4.2 are rejected in Experiment CC1, but H4 is confirmed in Experiment CC2. The confirmation in Experiment CC2 means that the experimental manipulation concerning the *Technical System Capability* was successful, as the applied sound reproduction methods changed the perceived system quality as intended. Correspondingly, the rejection in Experiment CC1 means at first glance that the experimental manipulation was not successful. However, a more likely reason for rejecting H4.1 and in parts H4.2 is not the experimental manipulation but the non-optimal measurement: the fact that the hypotheses H4.1 and H4.2 were constructed *in-retrospect* for the present analysis suggests that the chosen measures might not be sufficient to confirm H4.1 and H4.2.

Hypothesis H5 is rejected in Experiment CC1 and confirmed in Experiment CC2. Confirmation of H5 in Experiment CC2 can be considered as the major finding of this study, because it says that *Communication Complexity* has an influence on *Quality of Experience*, at least in case of comparing two-interlocutor scenarios against scenarios with three or more interlocutors, and if no spatial audio is available. In light of a rejected H3, this has two implications. First, the questions in CC2 addressing *Speech Communication Quality* and *Quality of Experience* appear to be understood by the subjects as intended: the *Speech Communication Quality* questions focus on the technical quality that people experienced, while the *Quality of Experience* questions have a broader view addressing the conference as such. Second, this result provides one possible explanation for a typical observation from real-life, in which participants are often dissatisfied with telephone conferences, even if they used ordinary equipment: in such situations the communicative situation as such could also contribute to dissatisfaction, independently of the technical quality, as confirmed by the present study.

The rejection of H5 in Experiment CC1 is likely caused by the lack of distinction between *Speech Communication Quality* and *Quality of Experience* due to the actual wording of the question. Apparently, rejecting H5 in Experiment CC1 suggests that test participants interpreted the corresponding question with a focus on the technical quality and not quality in a broader sense. This means that an interpretation in terms of *Speech Communication Quality* and not in terms of *Quality of Experience* was likely used by the test participants, which in turn would then mean again a rejection of Hypothesis H3.

Hypothesis H6 is confirmed in both experiments, which clarifies the discussions concerning H5 and also H4. In Experiment CC1, confirmation of H6 allows an interpretation of the corresponding question as *Speech Communication Quality* instead of *Quality of Experience*. This, however, actually means a confirmation of H4 for Experiment CC1,

which in turn suggests that the experimental manipulation of CC<sub>1</sub> was successful and that the measures for H<sub>4.1</sub> and H<sub>4.2</sub> were not optimal.

In Experiment CC<sub>2</sub>, confirmation of H<sub>6</sub> is in line with an interpretation of *Quality of Experience* as it was intended: it is not only insufficient *Technical System Capability* that can contribute to dissatisfaction concerning *Quality of Experience* (H<sub>5</sub> in Experiment CC<sub>2</sub>); *Communication Complexity* can also contribute to *Quality of Experience*.

In summary, these findings show that – given the optimized test protocol of Experiment CC<sub>2</sub> – not only *Communication Complexity* but also *Technical System Capability* can influence *Cognitive Load*, and that not only *Technical System Capability* but also *Communication Complexity* can influence *Quality of Experience*.

*Review of the Approach and Open Questions for Future Work* Concerning the approach in Experiment CC<sub>1</sub>, the earlier analysis conducted in Skowronek & Raake<sup>41</sup>, revealed possibilities for improving the experimental methodology, which were then implemented in the test protocol of Experiment CC<sub>2</sub>. For the purpose of sharing the lessons learned, those aspects are briefly reviewed here: First, the sensitivity of the memory test turned out to be limited. While the self-reported questions regarding *Cognitive Load* showed a significant influence of both the number of interlocutors and the system conditions, the memory test revealed only a statistically significant effect for the number of interlocutors. Second, due to the experimental design, the variables *number of interlocutors* and *system condition* were mixed in one session. The analysis in Skowronek & Raake<sup>42</sup>, however, showed a dominant effect of the number of interlocutors. Hence, it can be concluded that the sensitivity of test subjects regarding the system quality seems to be reduced by the strong effect due to the number of interlocutors. Third, the number of participants was 13 and thus quite low. Usually the ITU recommends 24 or 32 subjects as an adequate number of test subjects in quality assessment experiments according to protocols as applied here (see for example ITU-T Rec. P.800<sup>43</sup>).

With respect to the lessons learned from the re-analysis for the present text, another weakness of the test protocol of Experiment CC<sub>1</sub> is that the *Measured Variables* were not yet as precisely defined as introduced in Section 7.2. This can be seen in the partly inconsistent and mutually overlapping results for hypotheses H<sub>3</sub>, H<sub>4</sub>, H<sub>5</sub> and H<sub>6</sub>.

Concerning the approach in Experiment CC<sub>2</sub>, we discussed in Skowronek and Raake<sup>44</sup> that most decisions concerning the applied experimental method were driven by the uncertainty whether any effects could be observed at all. Hence, the paradigm was to aim for rather strong differences between experimental conditions. As a consequence, the results of this study should be treated with care when adopting them to real-life scenarios, as there are a number of open questions that will be addressed now.

First, one choice fell on conducting a listening-only test paradigm, as those tests are known to be more sensitive than conversational

<sup>41</sup> Janto Skowronek and Alexander Raake. "Investigating the effect of number of interlocutors on the quality of experience for multi-party audio conferencing". In: *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Inter-speech2011)*. International Speech Communication Association. Florence, Italy, Aug. 2011, pp. 829–832

<sup>42</sup> Janto Skowronek and Alexander Raake. "Investigating the effect of number of interlocutors on the quality of experience for multi-party audio conferencing". In: *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Inter-speech2011)*. International Speech Communication Association. Florence, Italy, Aug. 2011, pp. 829–832

<sup>43</sup> ITU-T. *Recommendation P.800 - Methods for objective and subjective assessment of quality*. International Standard. Geneva, Switzerland: International Telecommunication Union, 1996

<sup>44</sup> Janto Skowronek and Alexander Raake. "Assessment of Cognitive Load, Speech Communication Quality and Quality of Experience for spatial and non-spatial audio conferencing calls". In: *Speech Communication* (2015), pp. 154–175. DOI: 10.1016

tests. However, conversational tests are known to be more realistic than listening-only tests. Hence, future work could investigate if the observed effects can be found in conversation tests as well. In particular, such conversation tests should study if there are scaling effects with regard to *Communication Complexity*.

Second, to maximize comparability of the recordings, the corresponding scenario instructions used here were rather strict and required quite some concentration from the speakers. Concerning future conversation tests, a revision of those scenarios might be necessary to simplify the task for test subjects. Although this could reduce the comparability between scenarios, it might help test subjects to focus not only on the scenario but also on the system's quality.

Third, to avoid that test subjects were overwhelmed by listening to conversations with up to six interlocutors, the author deliberately searched for test participants with experience in multiparty conversations via telecommunication services. As this profile already limited the choice of potential subjects, no further selection criteria were applied. As a consequence, test participants had no experience with spatial sound reproduction. The post-interviews in Experiment CC2 revealed that all 25 test participants noticed the differences between spatial and non-spatial sound reproduction, but six test participants did not notice any difference in bandwidth. Furthermore, even when test participants noticed the different bandwidths, it did not help them much in their memory task, as opposed to the spatial audio reproduction. This strong effect of spatial sound reproduction compared to bandwidth might explain, why no significant difference in *Cognitive Load* between narrowband and fullband conditions could be found. Another possible explanation is that audio bandwidth and spatial-audio representation are likely affecting two different processes in the human working memory: the visuo-spatial sketch pad, processing visual and spatial information, and the phonological loop, processing verbal material and semantic meaning. Baldis<sup>45</sup> discussed that spatial audio sound reproduction reduces *Cognitive Load* as spatial audio is actually using both memory mechanisms while non-spatial audio would be solely processed in the phonological loop. According to this memory-related considerations, non-spatial audio signals will be solely processed in the phonological loop, irrespectively of the bandwidth, and might therefore have hardly any effect on *Cognitive Load*. Future experiments could check for effects of bandwidth or other technical parameters on the *Cognitive Load*, either by inviting subjects with experience in listening to spatial audio if spatial audio conditions will be used as well or by avoiding spatial audio in the test conditions.

Fourth, the questionnaire was on purpose rather long in order to sufficiently cover the variables *Cognitive Load*, *Speech Communication Quality*, and *Quality of Experience*. Concerning *Cognitive Load*, both the method of obtaining a performance-based measure by a memory test and the method of obtaining self-reported measures by assessment questions have some disadvantages. The memory test requires quite

<sup>45</sup> Jessica J. Baldis. "Effects of Spatial Audio on Memory, Comprehension, and Preference during Desktop Conferences". In: *Proceedings of the ACM CHI 2001 Human Factors in Computing Systems Conference*. Ed. by Michel Beaudouin-Lafon and Robert J. K. Jacob. Vol. 3. 1. 2001, pp. 166-173

some data points in order to provide significant results<sup>46</sup>. The disadvantage of the self-reported measures is that it is unclear if such questions are actually reflecting the *Cognitive Load*, even if they are accepted to measure the subjective perception of the mental effort<sup>47</sup>. For future experiments, it may be preferable to increase the practical feasibility by limiting the experimental effort for test participants. Furthermore, given the rather consistent results obtained in this study, the self-report questions, or even a subset, might be a preferable alternative for the memory test. However, more experience should be gained before a definite set of questions can be recommended.

Fifth and finally, the tested technical conditions are quite idealized compared to real life. Spatial sound reproduction and full-band channels are not yet fully established in the market and existing solutions face various impairments not tested here (coding artifacts, network conditions and packet loss, echo and delay, background noise etc.) In this context, also the combination of spatial sound reproduction and narrow-band, which was not tested here, could be a very relevant test scenario, given that spatial audio systems still need to cope with band-limited audio signals, e.g. if participants are connected with traditional equipment. Hence, another direction of future work can be to verify the found effects in test contexts that are closer to real-life scenarios.

## 7.4 Quality Impact of Involvement

This section reports on an experiment investigating the quality impact of *Involvement*, henceforth referred to as Experiment INV (INV standing for involvement).

### 7.4.1 Goals

*Motivation* One of the two main differentiators between a multiparty telemeeting and a conventional two-party audiovisual call is the special communicative situation. Next to the *Communication Complexity* investigated in the previous section, *Involvement* is a second aspect that potentially influences the perception of telemeeting quality. Especially in the context of conducting quality assessment tests it can be reasonably assumed that the more a test participant is feeling to be involved in a conversation, the more the test participant's quality perception will be based on the conversational situation, and thus the better the quality rating will reflect the multiparty experience. For that reason the present study will investigate the impact of *Involvement* on the perception of telemeeting quality.

*Approach* As for the study on *Communication Complexity*, the present study aims at investigating the impact of *Involvement* in the context of different *Technical System Capabilities*. This again leads to a two-factorial experiment, this time using *Involvement* and *Technical System Capabilities* as *Manipulated Variables*.

<sup>46</sup> The data set in Experiment CC<sub>1</sub> (13 subjects, 10 quotations per questionnaire) was too small to provide significant results; the data set in Experiment CC<sub>2</sub> (25 subjects, 16 quotations) was apparently sufficient.

<sup>47</sup> Roland Bruenken, Jan L. Plass, and Detlev Leutner. "Direct Measurement of Cognitive Load in Multimedia Learning". In: *Educational Psychologist* 38.1 (2003), pp. 53–61

*Involvement* is manipulated by giving different tasks to the test participants, whereas the test tasks comprise listening-only and conversation tasks. *Technical System Capability* is manipulated by applying different technical conditions of the sound transmission chain – henceforth abbreviated as *SndTrans* – used for both listening-only and conversation tasks.

Concerning the *Measured Variables*, the study used three variables: *Cognitive Load*, *Speech Communication Quality* and *Quality of Experience*. Furthermore, the study used an additional *Measured Variable*, called *Measured Involvement*, which primarily served as a means to be able to check for the success of the experimental manipulation.

*Hypotheses* At the time of preparing the experiment<sup>48</sup>, the three *Measured Variables* were defined as *Measured Involvement*, *Cognitive Load* and *Quality*, whereas *Quality* was not as precisely defined as it is described in Section 7.2. The reason is that the concept of the *Telecommunication* and *Group-Communication Components* were not yet developed to a degree as presented in Section 5.4 and the differentiation between *Speech Communication Quality* and *Quality of Experience* were not yet considered at a stage as presented in Section 7.2. Instead, the different employed measures were chosen to cover different aspects of *Quality* without a strict link to the aspects of *Speech Communication Quality* vs. *Quality of Experience* and *Telecommunication* vs. *Group-Communication Components*. Thus, two *Manipulated* and three *Measured Variables* were available, allowing to define six hypotheses of the form “[*Manipulation Variable*] has an impact on [*Measured Variable*]”.

However, by re-considering the actual wordings of the employed questions, it is possible to split *in retrospect* the variable *Quality* into two *Measured Variables*: one variable targeting *Speech Communication Quality* with a focus on the *Group-Communication Component*, the other variable targeting quality in a broader sense, whereas a more precise assignment to either *Speech Communication Quality* or *Quality of Experience* is not possible. Accordingly, the hypothesis concerning *Quality* can be split into two parts.

Thus, with two *Manipulated* and four *Measured Variables*, it is possible to define eight hypotheses; Table 7.12 gives the respective overview. All eight hypotheses will be tested in this work.

Measured Variables	Manipulated Variables	
	<i>Involvement</i> (i.e. <i>Task</i> )	<i>Technical System Capability</i> (i.e. Sound Transmission Method <i>SndTrans</i> )
<i>Measured Involvement</i>	Hypothesis H1	Hypothesis H2
<i>Cognitive Load</i>	Hypothesis H3	Hypothesis H4
<i>Speech Communication Quality</i> with focus on <i>Group-Communication Component</i>	Hypothesis H5	Hypothesis H6
<i>Speech Communication Quality</i> and/or <i>Quality of Experience</i>	Hypothesis H7	Hypothesis H8

<sup>48</sup> Janto Skowronek, Falk Schiffner, and Alexander Raake. “On the influence of involvement on the quality of multiparty conferencing”. In: *4th International Workshop on Perceptual Quality of Systems (PQS 2013)*. International Speech Communication Association. Vienna, Austria, Sept. 2013, pp. 141–146

Table 7.8: Overview of the *Manipulated* and *Measured Variables* and definition of the resulting hypotheses for Experiment INV. Each hypothesis can be phrased as “[*Manipulated Variable*] has an impact on [*Measured Variable*]”.

### 7.4.2 Experimental Factors

*Tasks and Conversation Scenarios* Each task aimed at a different level of involvement, while the test conduction was specifically designed to maintain an optimal comparability between those tasks. The decision was to have two tasks per test paradigm (listening-only test and conversation test). The four resulting tasks can be characterized as follows:

Task 1 (T1) “passive listening”: The test participant is just listening to recorded telephone conferences. After listening, the test participant fills in a quality assessment questionnaire.

Task 2 (T2) “listening with writing minutes”: While listening, the test participant is asked to write down minutes of the conversation. After listening, the test participant fills in a quality assessment questionnaire and a questionnaire about the minutes-writing task.

Task 3 (T3) “conversation according to sequential script”: The test subject participates in a conversation with two other interlocutors using a script. The script triggers a fixed sequence of contributions from the interlocutors by stating who will add what type of information in which order. However, the script gives freedom concerning the exact wording, as it does not contain fully formulated sentences. The full scenario scripts are electronically available<sup>49</sup>.

Task 4 (T4) “conversation according to scenario description”: The test subject participates in a conversation with two other interlocutors using a script that provides information in the form of bullet points, tables and pictographs. It allows for more freedom in the order of contribution and the interaction between people as the T3 script does. However, some underlying structure is given by the way information is distributed among the participants (e.g. one has a question and the others have information concerning that question). In fact, these scenarios are shortened versions of the multiparty test scenarios used in Raake et al.<sup>50</sup>. The full scenario scripts are electronically available<sup>51</sup> and have been originally developed for the study in Skowronek et al.<sup>52</sup>.

*Recording of Speech Material* To keep the technical setup essentially the same for both the listening-only and conversation tasks, the recordings for the listening-only tasks were made with the same telephony setup used in the experiment. This test system comprised a central conference bridge using Asterisk and off-the-shelf VoIP telephones (SNOM870), which were connected in a local network (all via one router) and communicated via the G.711 A-law codec. To obtain the recordings, three speakers conversed over this telephony setup and the used conference bridge wrote the speech signals directly to the harddisk. In the listening-only tasks, the conference bridge played back the recordings and the test participants were thus listening to them using the same hardware (SNOM870) as they used for the conversation tasks.

<sup>49</sup> Falk Schiffner and Janto Skowronek. *3seqCT - 3-party Sequential Conversation Test scenarios for conferencing assessment*. 2013. DOI: 10.5281/zenodo.16137

<sup>50</sup> Alexander Raake et al. “Listening and conversational quality of spatial audio conferencing”. In: *Proceedings of the AES 40th International Conference*. Tokyo, Oct. 2010

Alexander Raake and Claudia Schlegel. *3CT - 3-party Conversation Test scenarios for conferencing assessment*. 2010. DOI: 10.5281/zenodo.16129

<sup>51</sup> Janto Skowronek. *3SCT - 3-party Short Conversation Test scenarios for conferencing assessment (Version 01)*. 2013. DOI: 10.5281/zenodo.16134

<sup>52</sup> Janto Skowronek, Julian Herlinghaus, and Alexander Raake. “Quality Assessment of Asymmetric Multiparty Telephone Conferences: A Systematic Method from Technical Degradations to Perceived Impairments”. In: *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*. International Speech Communication Association. Lyon, France, Aug. 2013, pp. 2604–2608

*Optimal Comparability of Tasks, Scenarios and Speech Material* While the tasks should invoke different levels of involvement, they also needed to be as comparable as possible. For that purpose, a number of measures minimized the variation between the different tasks: First, the T4 scenarios had the same underlying conversational structure<sup>53</sup> while the actual content differs between those scenarios<sup>54</sup>. By removing the open discussion part from the original scenarios, the comparability was further increased. The T3 scripts were based on those same scenarios, and the recordings used for T1 and T2 were made according to the script of T3. Second, in all stimuli, two of the three interlocutors were always the same speakers: Two confederates (the first two authors of Skowronek et al.<sup>55</sup>) acted as interlocutors both in the conversation tasks (T3 & T4) and in the recordings for the listening tasks (T1 & T2). Third, the roles of the confederates in the scenario was always the same (one having a question, one a solution); the third role (having a constraint) was either assumed by the third speaker in case of the listening tasks or by the test subject in case of the conversation tasks. Fourth, by means of exercising beforehand, the confederates aimed for the same interaction behavior throughout all scenarios and test sessions. Fifth, the confederates also developed beforehand strategies (e.g. waiting, directly addressing, interrupting) to get more *passive, hesitant* test subjects involved in the conversation in the same way than more *active, talkative* test subjects.

*Technical Conditions* In Skowronek et al.<sup>56</sup>, we decided to limit the technical conditions to one type of degradation, but having different levels of that degradation. The disadvantage of this decision is that any results of this study cannot easily be generalized for other degradations. One advantage, however, is to not introduce additional experimental variables in terms of the perceptual and conversational nature of different degradation types, i.e. different degradation types can lead to different perceptual dimensions<sup>57</sup> and they can lead to different conversational behavior (e.g. effect of speech signal distortions vs. effect of echo or delay). A second advantage is that one can better quantify any observed effects, as multiple data points for that degradation are available, without the need to extend the experimental effort for test participants.

The choice fell on packet loss, as it is one of the most prominent degradations in today's telecommunication systems. By testing different packet loss rates with the test system, the aim was to cover a rather broad quality range from *imperceptible* to *perceptible and very annoying*, while limiting the upper loss rate in order to avoid negative effects on the conversation flow. Eventually, in Skowronek et al., we opted for random packet loss at loss rates of 0, 5, 10 and 15%, which introduced some temporally relative constant noise-like distortions of speech signals. The telephones used apply as packet loss concealment the G.711 Appendix I standard, which is known to show a high quality robustness against packet loss<sup>58</sup>, explaining why it was possible to go up to 15% loss rate without having extreme distortions. Furthermore,

<sup>53</sup> Alexander Raake et al. "Listening and conversational quality of spatial audio conferencing". In: *Proceedings of the AES 40th International Conference*. Tokyo, Oct. 2010

<sup>54</sup> Subjects should not have exactly the same content in all tasks in order to avoid any effects due to such repetitions (e.g. learning, annoyance).

<sup>55</sup> Janto Skowronek, Falk Schiffner, and Alexander Raake. "On the influence of involvement on the quality of multiparty conferencing". In: *4th International Workshop on Perceptual Quality of Systems (PQS 2013)*. International Speech Communication Association. Vienna, Austria, Sept. 2013, pp. 141–146

<sup>56</sup> Janto Skowronek, Falk Schiffner, and Alexander Raake. "On the influence of involvement on the quality of multiparty conferencing". In: *4th International Workshop on Perceptual Quality of Systems (PQS 2013)*. International Speech Communication Association. Vienna, Austria, Sept. 2013, pp. 141–146

<sup>57</sup> Marcel Wältermann. *Dimension-based Quality Modelling of Transmitted Speech*. Springer, 2013

<sup>58</sup> ITU-T. *Recommendation G.113 - Transmission impairments due to speech processing*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2007

the decision not to use more realistic burst-like packet loss behavior was deliberately made in order to avoid any temporal interaction between bursts of lost packets and possible important parts of any information, e.g. to avoid that whole digits of a telephone number are disturbed in one session but not in another. Finally, packet loss simulation was realized with the TC filter and Netem software packages, and the packet loss was introduced in receiving direction of test participants; thus test participants heard all other interlocutors with the same degradation.

### 7.4.3 Method

*Design* The design was driven by the following constraints in terms of temporal order of stimuli, scenarios, tasks, and technical conditions:

1. Number of available scenarios and corresponding recordings: 4
2. Number of tasks: 4
3. Number of technical conditions: 4
4. All combinations of scenarios, tasks and technical conditions should be tested at least once.
5. Each test participant should listen to or play a scenario recording only once.

The choice fell on a fourth-order Hyper-Greco-Latin square, repeated multiple times to cover the targeted number of test participants.

*Test Participants* The characteristics of the invited test participants are given in Table 7.9.

Participant characteristics	Statistics
Number of test participants	20
Age: mean / minimum / maximum	29.4 / 19 / 45
Gender: female / male, in percent (absolute numbers)	55% / 45% (11 / 20)
Multiparty experience in professional or private life	60% (12 of 20)
Recruitment area	university area, outside laboratory
Background / profession	researchers, students, staff, alumni
Hearing impairments (self-reported)	none

Table 7.9: Test participant statistics for the study on involvement.

*Procedure* The test participants came to the laboratory for a single one-hour session. After a general introduction, the test participants went through the four different tasks in an order according to the experimental design. Just before each task, the test participants got the corresponding individual test instructions in written and oral form, and had also a short training call before each of the conversation tasks. The test participants filled in a questionnaire after each call and at the end of the experiment.

#### 7.4.4 Data Acquisition

*Questionnaires and Measures* The questionnaires were designed to extract for most *Measured Variables* one measure, for some however, multiple measures from the test participants' answers. In Skowronek et al.<sup>59</sup> our motivation for this was two-fold: on the one hand to limit the number of questions; on the other hand to cover multiple facets of a *Measured Variable* where it deemed appropriate. The selection of the different measures and the corresponding questionnaire design was inspired by the first experiment on *Communication Complexity* CC1, standard quality assessment methods<sup>60</sup>, and Möller's taxonomy for the quality of communication services<sup>61</sup>.

There were in total four questionnaires A to D. Questionnaire A was given to subjects after each call and consisted of four main parts: self-reported involvement (question QA.1), quality judgment (question QA.2), effort ratings (questions QA3.1, QA3.2a, QA3.2b, QA3.3), and judgments of task impact on quality ratings (questions QA4.1, QA4.2).

Questionnaire B was given to subjects only after each T2 call ("listening with writing minutes") and after they answered Questionnaire A. This questionnaire was used to get more insights on the potential difference between T1 ("passive listening") and T2. Since the answers were in retrospect not very conclusive, this questionnaire will not be further presented and analyzed in this work.

Questionnaire C essentially resembled the writing minutes task and subjects answered it during every T2 call: it comprised a number of specific questions about the actual content of the call, guiding subjects in their minutes writing task. In Skowronek et al.<sup>62</sup>, we did not further analyze the answers of this questionnaire, e.g. if they found all wanted information; but we checked, if they have done the task at all.

Questionnaire D was given to subjects at the end of the experiment. It repeated question QA.1 for all four tasks, i.e. it asked for the degree of involvement for the individual tasks in retrospect, after the subjects had experienced all four tasks. By assigning these questions to the individual experimental stimuli (*Task & SndTrans*), we generated one question QD that can be analyzed in the same way as its counterpart QA.1.

Most questions had to be answered on the same 7-point continuous scale used in the previous experiments CC1 and CC2 (see also Figure 7.3), some questions had to be answered with yes or no, and some by free text. Table 7.10 provides English translations of the questions as well as the German originals.

Table 7.11 gives an overview of the link between the *Measured Variables*, the actual measures used, and descriptions on how they were extracted from the questionnaires, and the following paragraphs provide detailed explanations of those links.

<sup>59</sup> Janto Skowronek, Falk Schiffner, and Alexander Raake. "On the influence of involvement on the quality of multiparty conferencing". In: *4th International Workshop on Perceptual Quality of Systems (PQS 2013)*. International Speech Communication Association. Vienna, Austria, Sept. 2013, pp. 141–146

<sup>60</sup> ITU-T. *Recommendation P.800 - Methods for objective and subjective assessment of quality*. International Standard. Geneva, Switzerland: International Telecommunication Union, 1996

ITU-T. *Recommendation P.805 - Subjective evaluation of conversational quality*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2007

ITU-T. *Recommendation P.851 - Subjective quality evaluation of telephone services based on spoken dialogue systems*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2003

<sup>61</sup> Sebastian Möller. *Assessment and Prediction of Speech Quality in Telecommunications*. Kluwer Academic publishers, 2000

<sup>62</sup> Janto Skowronek, Falk Schiffner, and Alexander Raake. "On the influence of involvement on the quality of multiparty conferencing". In: *4th International Workshop on Perceptual Quality of Systems (PQS 2013)*. International Speech Communication Association. Vienna, Austria, Sept. 2013, pp. 141–146

Questionnaire A		
Part	Question ID	Wording / Description, EN: English translations, DE: German original
1	QA1.1	EN: "If this was a real call, in how far would you have felt as part of that telephone conference? – <i>extremely much ... extremely little</i> " DE: "Wenn es sich um ein echtes Telefonat gehandelt hätte, in wie weit fühlten Sie sich als Teil der Konferenz? – <i>extrem viel ... extrem wenig</i> "
2	QA2.1	EN: "What was your overall quality impression of the connection? – <i>extremely bad ... ideal</i> " DE: "Wie ist Ihrer Meinung nach die Gesamtqualität der Verbindung, die Sie gerade benutzt haben? – <i>extrem schlecht ... ideal</i> "
3	QA3.1	EN: "Did you have any difficulty in talking or hearing over the connection? – <i>yes, no</i> " DE: "Hatten Sie Schwierigkeiten beim Sprechen oder Hören während dieser Verbindung? – <i>ja, nein</i> "
	QA3.2a	EN: "Did your partners have any difficulty in talking or hearing over the connection? – <i>yes, no</i> " DE: "Hatten Ihre Partner Schwierigkeiten beim Sprechen oder Hören während dieser Verbindung? – <i>ja, nein</i> "
	QA3.2b	EN: "I am <i>xxx</i> % sure of my answer." DE: "Ich bin mir meiner Antwort zu <i>xxx</i> % sicher."
	QA3.3	EN: "What was your level of effort to follow the conversation? – <i>extremely much ... extremely little</i> " DE: "Bitte bewerten Sie die Anstrengung um der Konversation zu folgen? – <i>extrem viel ... extrem wenig</i> "
4	QA4.1	EN: "How difficult was it for you to give a rating for the overall quality impression and the effort? – <i>extremely difficult ... extremely easy</i> " DE: "Wie war es für Sie eine Bewertung zur Gesamtqualität und zur Anstrengung abzugeben? – <i>extrem schwer ... extrem leicht</i> "
	QA4.2	EN: "In how far did the conversation (in terms of content and flow) influence your answers? – <i>extremely much ... extremely little</i> " DE: "In wie weit spielte das Gespräch (Inhalt und Ablauf) bei Ihren Urteilen eine Rolle. – <i>extrem viel ... extrem wenig</i> "
Questionnaire D		
Part	Question ID	Wording / Description, EN: English translations, DE: German original
1	QD1.1 ... QD1.4	EN: "For task XXX: If this was a real call, in how far would you have felt as part of that telephone conference? – <i>extremely much ... extremely little</i> " DE: "Für Aufgabe XXX: Wenn es sich um ein echtes Telefonat gehandelt hätte, in wie weit fühlten Sie sich als Teil der Konferenz? – <i>extrem viel ... extrem wenig</i> "

Table 7.10: Summary of the questionnaires used in the study on involvement.

Measured Variable	Measure	Extraction from questionnaire
<i>Measured Involvement</i>	<i>Perceived Involvement</i>	Directly measured with QA1.1
	<i>Perceived Involvement In Retrospect</i>	Measured with QD1.1 ... QD1.4, after assignment to experimental condition
<i>Measured Involvement and/or Cognitive Load</i> <sup>1</sup>	<i>Influence Of Content</i>	Directly measured with QA4.2
	<i>Rating Difficulty</i>	Directly measured with QA4.1
<i>Speech Communication Quality and/or Quality of Experience</i> <sup>2</sup>	<i>Quality</i>	Directly measured with QA2.1
<i>Speech Communication Quality with focus on Group-Communication Component</i>	<i>Own Conversation Effort</i>	Directly measured with QA3.1
	<i>Partners Conversation Effort</i>	Directly measured with QA3.2a
	<i>Partners Conversation Effort Confidence</i>	Directly measured with QA3.2b
<i>Cognitive Load</i>	<i>Perceived Effort</i>	Directly measured with QA3.3
Remarks:		
<sup>1</sup> In retrospect, any difficulty can be caused by a high cognitive load and/or by a strong distraction due to involvement. Hence a unique assignment to either Measured Variable is not possible.		
<sup>2</sup> In retrospect, the wording of question QA2.1 allows both interpretations. Hence a unique assignment to either Measured Variable is not possible.		

Table 7.11: Relation between the Measured Variables (on their conceptual level), the actual measures used for obtaining respective quantitative data, and extraction of the measures from the questionnaire answers.

*Measured Involvement* Two measures were used that directly target *Measured Involvement*. The first measure is *Perceived Involvement*, obtained from question QA.1, which participants rated after each call. The second measure is *Perceived Involvement in Retrospect*, obtained from question QD, which participants answered “in retrospect” at the end of the experiment.

Furthermore, two additional measures were used that concern the task impact on quality ratings, which indirectly indicate the level of *Measured Involvement*, due to the design of the tasks to trigger different levels of involvement. The first measure is *Influence of Content*, obtained from participants’ ratings of question QA4.2. The underlying assumption is that the larger the *Influence of Content*, the more the test participant was concerned with the content and thus the more the participant was involved in the conversation. The second measure is *Rating Difficulty*, obtained from ratings of question QA4.1. While the underlying assumption is the same as for *Influence of Content*, a reconsideration of this question in retrospect revealed that any difficulty can be caused by a strong distraction due to involvement, but it can also be caused by a high cognitive load. Hence a unique assignment to either *Measured Involvement* or *Cognitive Load* is not possible, which is also accounted for in Table 7.11.

*Cognitive Load* Next to above mentioned *Rating Quality*, a second and actually more direct measure for *Cognitive Load* is the *Perceived Effort* (to follow the conference), obtained from test participants’ ratings of question QA3.3.

*Speech Communication Quality and Quality of Experience* As already mentioned, the experiment did not specifically aim at distinguishing between the two layers *Speech Communication Quality* and *Quality of Experience*. The experiment applied one measure *Quality*, obtained from participants’ ratings of question QA2.1. The wording of that question (“overall impression of the connection”) actually allows an interpretation in terms of *Speech Communication Quality* due to the keyword *connection* or in terms of *Quality of Experience* due to the keyword *overall impression*. For that reason, the corresponding hypothesis does not distinguish between *Speech Communication Quality* and *Quality of Experience*, see Table 7.12.

Next to the *Quality* measure, three additional measures specifically addressed the *Group-Communication Component* of *Speech Communication Quality*. Those measures are the *Own Conversation Effort*, obtained from participants’ ratings of question QA3.1; the *Partners Conversation Effort*, obtained from participants’ ratings of question QA3.2a, and the *Partners Conversation Effort Confidence*, obtained from participants’ ratings of question QA3.2b. While it could be argued that those measures also refer to *Cognitive Load*, as they ask for effort, the author assigned those measures to *Speech Communication Quality* due to the fact that the questions are minor modifications of standardized questions of conversational quality, i.e. the conversation effort scale

proposed by the ITU<sup>63</sup>.

*Data Processing* All measures were represented in such a way that high values reflect something positive, that is high quality, high involvement, or low cognitive load. This representation is straightforward for the concepts related with four measures (*Perceived Involvement, Perceived Involvement In Retrospect, Quality, Partners Conversation Effort Confidence*). However, it is less intuitive for the five effort- and influence-related measures (*Influence Of Content, Rating Difficulty, Own Conversation Effort, Partners Conversation Effort, Perceived Effort*), as the ratings are now inverted: a high effort or influence, meaning a high cognitive load, is coded with a low value.

*Data Validation* Inspecting the questions did not reveal any specific outliers in terms of individual ratings or test participants. Hence none of the data points was excluded from the further analysis.

*Data Analysis and Presentation* As the experiments on *Communication Complexity*, also this experiment on *Involvement* had two test factors, *Task* and sound transmission method *SndTrans*, again requiring an analysis on two levels. The first analysis level concerns the main effects; that means the effect of *Task* if all data over *SndTrans* is compiled and the effect of *SndTrans* if all data over *Task* is compiled. The second analysis level concerns the individual effects; that means the effect of *Task* per *SndTrans* and the effect of *SndTrans* per *Task*. As analysis methods, the author conducted one-way ANOVAs<sup>64</sup> to check for significant effects of the *Manipulated Variables*, i.e. *Task* and *SndTrans*, on the different *Measured Variables*. Then the author conducted PostHoc tests with Sidak correction to identify for which pairs of *Manipulated Variables* significant differences can be observed. And the author generated errorbar plots, showing the mean values and 95% confidence intervals, to visualize the effects. Thus, the analysis consisted – as for the experiments on *Communication Complexity* – of nine steps per measure: ANOVA, PostHoc and Errorbar for the Main Effects, for the Individual Effect of *Task* and for the Individual Effect of *SndTrans*.

The next section presents a summary of the results, focusing on the main conclusions from the analysis, and visualizing the data with the errorbar plots. The full analysis details with all ANOVA and PostHoc results are presented in Appendix D.

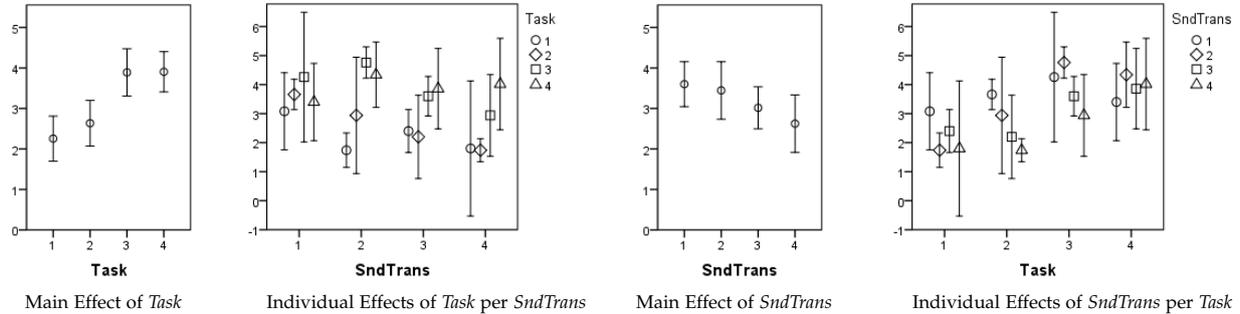
<sup>63</sup> ITU-T. *Recommendation P.805 - Subjective evaluation of conversational quality*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2007

<sup>64</sup> Due to the experimental design (hyper-greco-latin design), repeated-measures ANOVAs as used in the experiments on *Communication Complexity* were only possible here for investigating individual effects, but not for investigating the main effects. For that reason, simple one-way ANOVAs were used instead.

7.4.5 Results Experiment INV

Quality

1. Errorbars (mean & 95% confidence interval)



2. Effect of Task

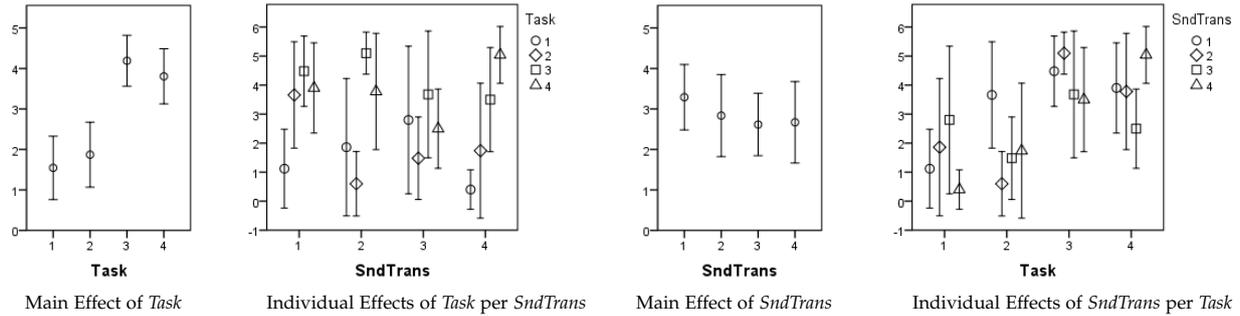
Quality is significantly influenced by the principle test paradigms listening-only tasks (1 & 2) vs. conversation tasks (3 & 4). There are deviations from this observation for individual SndTrans, which are, however, non-systematic.  
 ⇒ Support for Hypotheses H7, but only for comparison listening-only vs. conversation tasks.

3. Effect of SndTrans

There is an effect of SndTrans on Quality: with increasing packet loss rate (from SndTrans = 1 to 4), Quality is decreasing, whereas the differences are only significant between the two extreme cases SndTrans = 1 and 4. There are deviations from this observation for individual Tasks, which are, however, non-systematic.  
 ⇒ Support for Hypotheses H8.

Perceived Involvement

1. Errorbars (mean & 95% confidence interval)



2. Effect of Task

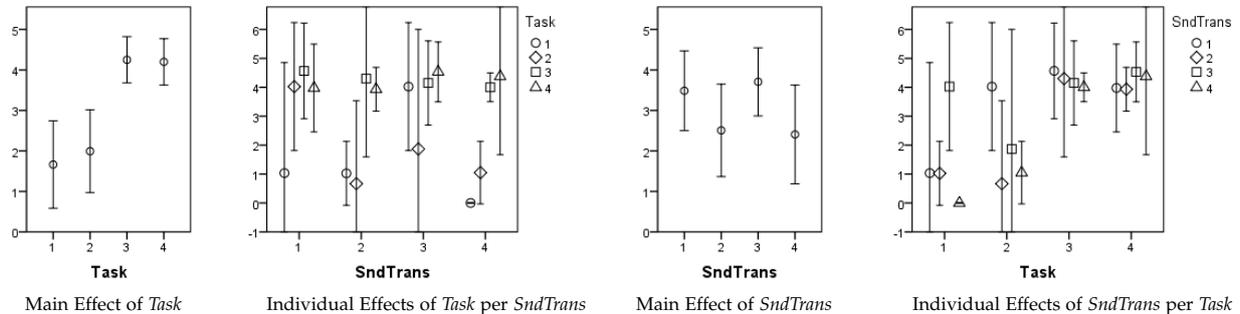
Perceived Involvement is significantly influenced by the principle test paradigms listening-only tasks (1 & 2) vs. conversation tasks (3 & 4). There are deviations from this observation for individual SndTrans, which are, however, non-systematic.  
 ⇒ Support for Hypothesis H1, but only for comparison listening-only vs. conversation tasks.

3. Effect of SndTrans

There is no effect of SndTrans on Perceived Involvement. There are deviations from this observation for individual Tasks, which are, however, non-systematic.  
 ⇒ No support for Hypothesis H2.

Perceived Involvement in Retrospect

1. Errorbars (mean & 95% confidence interval)



2. Effect of Task

Perceived Involvement in Retrospect is significantly influenced by the principle test paradigms listening-only tasks (1 & 2) vs. conversation tasks (3 & 4). There are deviations from this observation for individual SndTrans, which are, however, non-systematic.  
 ⇒ Support for Hypothesis H1, but only for comparison listening-only vs. conversation tasks.

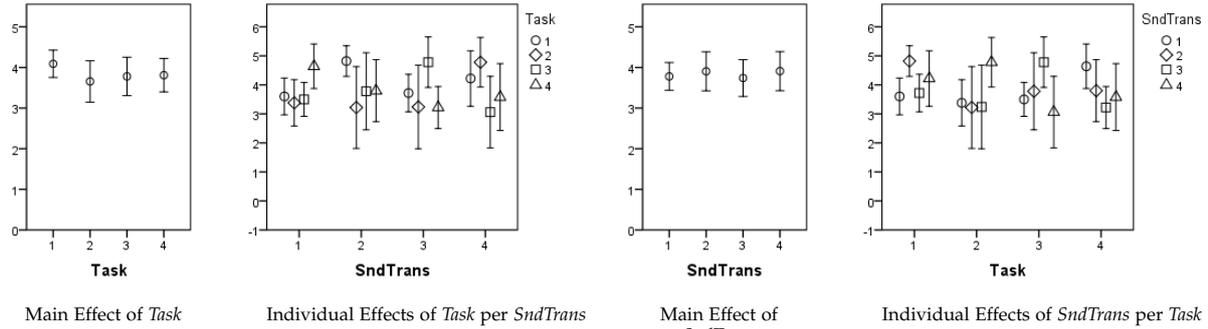
3. Effect of SndTrans

There is an effect of SndTrans on Perceived Involvement in Retrospect, a grouping of SndTrans = 1 & 3, i.e. 0% & 10% packets loss rate, and a grouping of SndTrans = 2 & 4, i.e. 0% & 10% packets loss rate. No reasonable explanation could be found, except that the data appears to be rather noisy, especially when considering the data for individual SndTrans.  
 ⇒ No "interpretable" support for Hypothesis H2.

Figure 7.12: Analysis summary for the measures Quality, Perceived Involvement, and Perceived Involvement in Retrospect.

**Rating Difficulty**

**1. Errorbars (mean & 95% confidence interval)**



**2. Effect of Task**

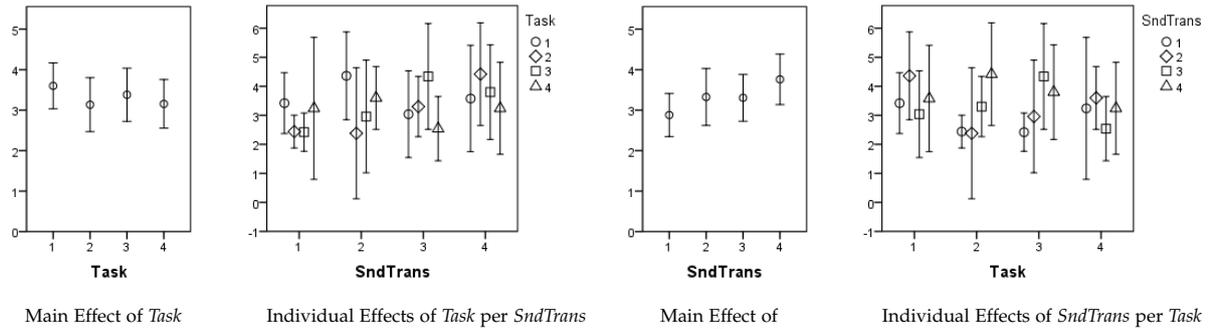
There is no effect of *Task* on *Rating Difficulty*. There are deviations from this observation for individual *SndTrans*, which are, however, non-systematic.  
 ⇒ No support for Hypotheses H1 & H3.

**3. Effect of SndTrans**

There is no effect of *SndTrans* on *Rating Difficulty*. There are deviations from this observation for individual *Tasks*, which are, however, non-systematic.  
 ⇒ No support for Hypotheses H2 & H4.

**Influence of Content**

**1. Errorbars (mean & 95% confidence interval)**



**2. Effect of Task**

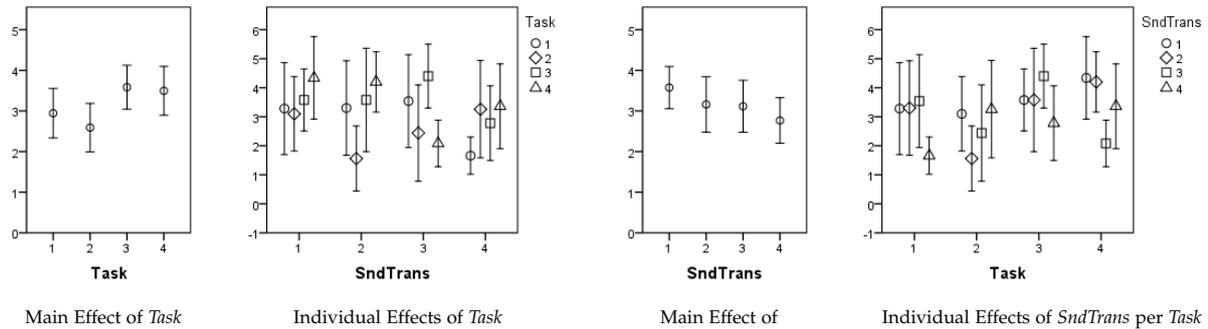
There is no effect of *Task* on *Influence of Content*. There are deviations from this observation for individual *SndTrans*, which are, however, non-systematic.  
 ⇒ No support for Hypothesis H1.

**3. Effect of SndTrans**

There is no effect of *SndTrans* on *Influence of Content*. There are deviations from this observation for individual *Tasks*, which are, however, non-systematic.  
 ⇒ No support for Hypothesis H2.

**Perceived Effort**

**1. Errorbars (mean & 95% confidence interval)**



**2. Effect of Task**

*Perceived Effort* is influenced by the principle test paradigms listening-only tasks (1 & 2) vs. conversation tasks (3 & 4), whereas the difference is only significant between Tasks 2 & 3. There are deviations from this observation for individual *SndTrans*, which are, however, non-systematic.  
 ⇒ Support for Hypothesis H3, but only for comparison of Listening-only Task 2 vs. Conversation Task 3.

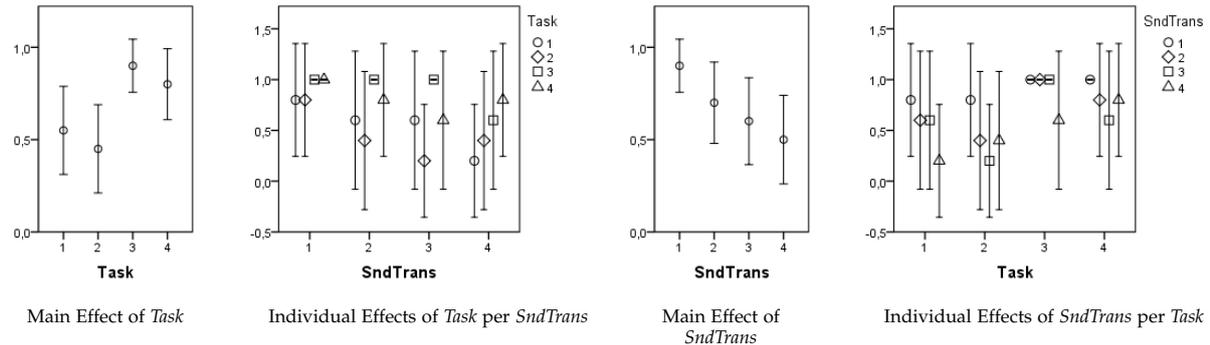
**3. Effect of SndTrans**

There is no effect of *SndTrans* on *Perceived Effort*. There are deviations from this observation for individual *Tasks*, which are, however, non-systematic.  
 ⇒ No support for Hypothesis H4.

Figure 7.13: Analysis summary for the measures *Rating Difficulty*, *Influence of Content*, and *Perceived Effort*.

**Own Conversation Effort**

**1. Errorbars (mean & 95% confidence interval)**



**2. Effect of Task**

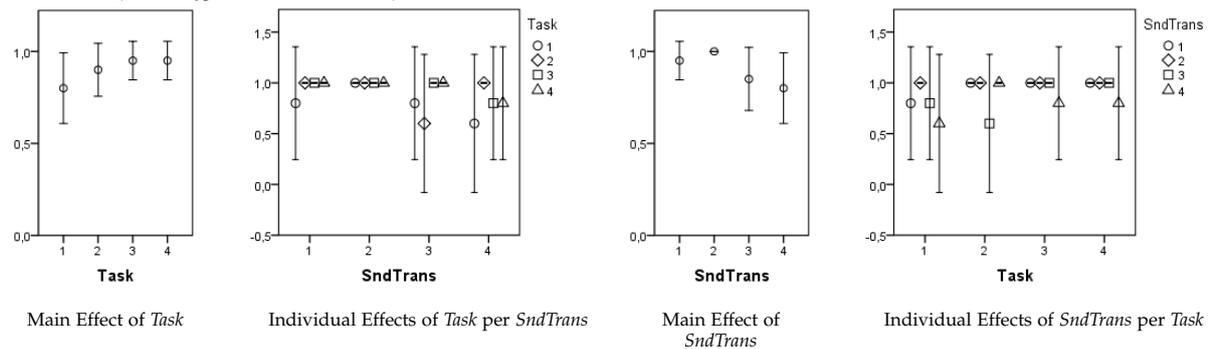
*Own Conversation Effort* is influenced by the principle test paradigms listening-only tasks (1 & 2) vs. conversation tasks (3 & 4), whereas the difference is only significant between Tasks 2 & 3. There are deviations from this observation for individual *SndTrans*, which are, however, non-systematic.  
 ⇒ Support for Hypothesis H5, but only for comparison of Listening-only Task 2 vs. Conversation Task 3.

**3. Effect of SndTrans**

There is an effect of *SndTrans* on *Own Conversation Effort*: with increasing packet loss rate (from *SndTrans* = 1 to 4), *Own Conversation Effort* is increasing (i.e. decreasing values), whereas the differences are only significant between the two extreme cases *SndTrans* = 1 and 4. There are deviations from this observation for individual *Tasks*, which are, however, non-systematic.  
 ⇒ Support for Hypotheses H6.

**Partners Conversation Effort**

**1. Errorbars (mean & 95% confidence interval)**



**2. Effect of Task**

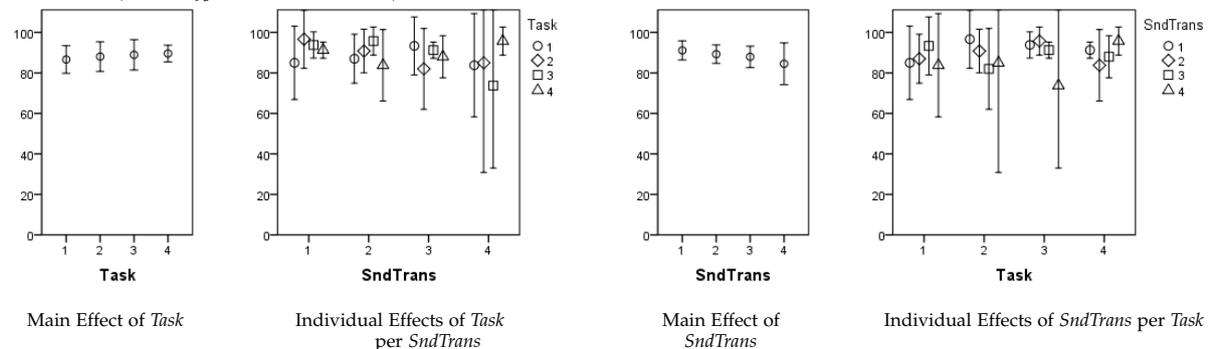
There is no effect of *Task* on *Partners Conversation Effort*. There are deviations from this observation for individual *SndTrans*, which are, however, non-systematic.  
 ⇒ No support for Hypothesis H5.

**3. Effect of SndTrans**

There is no effect of *SndTrans* on *Partners Conversation Effort*. There are deviations from this observation for individual *Tasks*, which are, however, non-systematic.  
 ⇒ No support for Hypothesis H6.

**Partners Conversation Effort Confidence**

**1. Errorbars (mean & 95% confidence interval)**



**2. Effect of Task**

There is no effect of *Task* on *Partners Conversation Effort Confidence*. There are deviations from this observation for individual *SndTrans*, which are, however, non-systematic.  
 ⇒ No support for Hypothesis H5.

**3. Effect of SndTrans**

There is no effect of *SndTrans* on *Partners Conversation Effort Confidence*. There are deviations from this observation for individual *Tasks*, which are, however, non-systematic.  
 ⇒ No support for Hypothesis H6.

Figure 7.14: Analysis summary for the measures *Own Conversation Effort*, *Partners Conversation Effort*, and *Partners Conversation Effort Confidence*.

7.4.6 Discussion

Review of Results Table 7.12 summarizes the main findings, which will be discussed now.

Measured Variables	Manipulated Variables	
	Involvement (i.e. Task)	Technical System Capability (i.e. Sound Transmission Method SndTrans)
Measured Involvement	Hypothesis H1 ✓ <sup>1</sup>	Hypothesis H2 × <sup>2</sup>
Cognitive Load	Hypothesis H3 ✓ <sup>3</sup>	Hypothesis H4 ×
Speech Communication Quality with focus on Group-Communication Component	Hypothesis H5 ✓ <sup>4</sup>	Hypothesis H6 ✓ <sup>5</sup>
Speech Communication Quality and/or Quality of Experience	Hypothesis H7 ✓ <sup>6</sup>	Hypothesis H8 ✓

Remarks:

✓<sup>1</sup> Hypothesis confirmed for measures *Involvement* & *Involvement in Retrospect* for comparison listening-only tasks (T1 & T2) vs. conversation tasks (T3 & T4), but not between the two listening-only tasks (T1 vs. T2) or between the two conversation tasks (T3 vs. T4). Measures *Rating Difficulty* & *Influence of Content* showed no effects.

×<sup>2</sup> Hypothesis rejected for *Perceived Involvement*, though there are some non-reasonable (i.e. differences between 0% and 10% vs. 5% and 15% packet loss rate) for *Perceived Involvement in Retrospect*. Measures *Rating Difficulty* & *Influence of Content* showed no effects.

✓<sup>3</sup> Hypothesis confirmed for measure *Perceived Effort* for comparison T2 vs. T3. Measure *Rating Difficulty* showed no effects.

✓<sup>4</sup> Hypothesis confirmed for measure *Own Conversation Effort* for comparison T2 vs. T3. Measures *Partners Conversation Effort* & *Partners Conversation Effort Confidence* showed no effects.

✓<sup>5</sup> Hypothesis confirmed for measure *Own Conversation Effort* for comparison SndTrans = 1 vs. 4. Measures *Partners Conversation Effort* & *Partners Conversation Effort Confidence* showed no effects.

✓<sup>6</sup> Hypothesis confirmed for comparison listening-only tasks (T1 & T2) vs. conversation tasks (T3 & T4), but not between the two listening-only tasks (T1 vs. T2) or between the two conversation tasks (T3 vs. T4).

Hypothesis H1 is confirmed due to the differences between listening-only and conversation tests. This means, the experimental manipulation in terms of triggering different levels of involvement was only partially successful, since the original intention was to trigger not only differences in involvement between listening-only and conversation tests, but also between the two variations of a listening-only test and conversation test, respectively. Furthermore, this finding holds only for the two measures directly asking about involvement (*Perceived Involvement* and *Perceived Involvement in Retrospect*). The other two measures (*Rating Difficulty* & *Influence of Content*), which are asking – in a manner of speaking – indirectly about involvement, are apparently not suited to reveal significant differences in involvement for different tasks.

Hypothesis H2 is rejected, if one interprets the non-reasonable result for *Perceived Involvement in Retrospect* as a measurement error. This means, people do not experience a change in *Involvement* when the *Technical System Capability* is changing, at least for the present range of degradations. Again, the two measures *Rating Difficulty* and *Influence of Content* are apparently not suited to reveal significant differences in involvement for different degradations.

Hypothesis H3 is confirmed due to the difference between listening-only and conversation tests, whereas this is limited to the two tasks T2 and T3. This means, a change in *Involvement* has an effect on the *Cognitive Load*. Looking at the direction of that effect, i.e. the error

Table 7.12: Overview of the *Manipulated* and *Measured Variables* and definition of the resulting eight hypotheses H1 to H8. Each hypothesis can be phrased as “[Manipulated Variable] has an impact on [Measured Variable]”.

bar plots for *Perceived Effort* in Figure 7.13 and considering the inverse coding of values for this measure, test participants appear to actually perceive a lower *Cognitive Load* when they feel more involved. This finding is contrary to the expectations, as it is more straight-forward to assume in a quality assessment test that a stronger involvement requires more cognitive resources from a test participant, given that the participant needs also cognitive resources for the judgment task. One explanation for this finding could be that feeling more involved makes the assessment task more interesting for test participants. This in turn means that their need to force themselves to stay concentrated throughout the test is reduced, which would then explain the reduction in *Perceived Effort*.

Hypothesis H4 is rejected. This means, the applied degradations do not affect *Cognitive Load*, which is in line with the results of the previous study, in which only spatial audio, but not bandwidth, were found to influence *Cognitive Load*.

Hypothesis H5 is confirmed for one measure (*Own Conversation Effort*) due to the difference between listening-only and conversation tests, whereas this is limited to the two tasks T2 and T3. This suggests that a change in *Involvement* has an effect on the communicative aspects of *Speech Communication Quality*. However, it is very likely that *Own Conversation Effort* as the one measure showing this effect is highly correlated with *Perceived Effort*, which in turn explains the same result as for Hypothesis H3. Concerning the other two related measures (*Partners Conversation Effort*, *Partners Conversation Effort Confidence*), it appears that those are not suitable for revealing significant differences for different test tasks.

Hypothesis H6 is confirmed for the difference between the two most extreme conditions (0% and 15% packet loss rate) for the measure *Own Conversation Effort*. This means, *Technical System Capability* can influence the *Group-Communication Component of Speech Communication Quality*, suggesting that the *Group-Communication Component* of telemeeting quality is relevant also from a technological perspective. Furthermore, this also supports the validity of that measure, which is not surprising, as it is essentially the conversation effort scale proposed in ITU-T Recommendation P.805<sup>65</sup>.

Hypothesis H7 is confirmed due to the difference between listening-only and conversation tests. This is the main finding of the study, and has implications regarding two aspects. First, confirmation of H7 contributes to the discussion about the impact of the two different test paradigms listening-only test and conversation test. The present results obtained for a multiparty setting are in line with existing results for two-party telephony discussed in Möller<sup>66</sup>, who showed that a difference between the two test paradigms exists and that no single generic transformation law between the outcomes of both test paradigms can be found. Second, confirmation of H7 emphasizes that the choice of either listening-only test or conversation test should be taken with great care, as this choice alone is already influencing results. Since the ratings for the corresponding measure *Quality* in

<sup>65</sup> ITU-T. *Recommendation P.805 - Subjective evaluation of conversational quality*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2007

<sup>66</sup> Sebastian Möller. *Assessment and Prediction of Speech Quality in Telecommunications*. Kluwer Academic publishers, 2000

Figure 7.12 are lower for the listening-only tests, test participants were apparently more critical in the listening-only tests. Thus, the widely used approach for two-party settings can also be applied here for multiparty settings: use listening-only tests, if a high sensitivity of test participants is required, use conversational tests, if sensitivity is not an issue but other aspects such as naturalness or interactivity.

Hypothesis H8 is confirmed, which means the experimental manipulation regarding the technical degradation was successful: a stronger degradation (packet loss rate) leads to significantly lower quality scores.

Another important result can be found when comparing the results for H7 and H8, i.e. the impact of tasks and technical degradation on the measure *Quality*. As already discussed in Skowronek et al.<sup>67</sup>, the largest difference of the mean score of *Quality* between two tasks was 1.680 points on the used 7-point continuous scale, while the largest difference between two technical conditions was 0.975. This means, the impact of the test paradigm (listening-only vs. conversation test) can be stronger than the impact of the technical conditions under test, even if those condition differences are strong enough to reach statistical significance, which was the case here.

*Review of the Approach and Open Questions for Future Work* While the present experiment delivered significant results on the impact of *Involvement* on different aspects of telemeeting quality, the results reveal some room for improving the test methodology as well as pose some general questions for future work.

First, the questionnaire contained four questions/measures that were not sensitive to measure significant effects: *Rating Difficulty*, *Influence of Content*, *Partners Conversation Effort*, and *Partners Conversation Effort Confidence*. Hence, the added value of those questions needs to be further validated, e.g. in another study with more data points, to determine whether those questions are useful to keep or whether they can be removed.

Second, the questionnaire was designed before the present framework of *Telecommunication and Group-Communication Components* was developed to a degree as presented in Section 5.4 and the differentiation between Speech Communication Quality and Quality of Experience were not yet considered at a stage as presented in Section 7.2. Despite the approach to re-visit the applied questions and to construct *in retrospect* hypotheses according to that framework, the results of H3 and H5 suggest that the questions are strongly interrelated and thus less suitable to really distinguish between the different aspects of telemeeting quality.

Third, the experimental design was a hyper-greco-latin square balancing two test factors (task & technical condition) with the available scenarios and the stimulus order. The advantage of this design is the limited number of test trials (i.e. total number of stimuli) required to conduct a study with multiple factors. This was an essential advantage with respect to the test protocol, which required introductions

<sup>67</sup> Janto Skowronek, Falk Schiffner, and Alexander Raake. "On the influence of involvement on the quality of multiparty conferencing". In: *4th International Workshop on Perceptual Quality of Systems (PQS 2013)*. International Speech Communication Association. Vienna, Austria, Sept. 2013, pp. 141–146

and training calls per task. One disadvantage of this design, however, is that the number of data points per combination of each factor is very low. For the present data, this means that there were only six ratings per combination of task and technical condition, despite the use of 24 test subjects. Accordingly, the detailed results on the effect of task per condition and on the effect of condition per task is quite noisy (see Figures 7.12 to 7.14). That means to obtain more stable results also on the level of those individual effects, the repetition of the study with another design or with more test sessions is necessary.

Next to these suggestions for improving the methodology, two more general questions for future work emerge from the present results. First, the study focused on one type of degradation, random packet loss, and future work could verify the found effects for other degradations.

Second, despite the effort in the task design, the study did not succeed to trigger different degrees of involvement within the listening-only or within the conversation test paradigms. Future work could attempt to define different tasks that are more likely to succeed. Or future work could further validate the present or new tasks without mixing listening-only and conversation tests, which could increase participants' sensitivity for rating the perceived involvement by removing any dominating effect of the two fundamental test paradigms. Alternatively, future work should investigate other methods for measuring involvement, e.g. physiological, that augment or even replace the employed self-reported measures.

### *Summary*

The goal of the research presented here was to obtain more insights about the influence of communicative aspects on telemeeting quality, whereas the potential interaction of those aspects with the technology was considered as well. For that purpose two studies investigated the influence of *Communication Complexity* and *Involvement* in combination with different *Technical System Capabilities* on telemeeting quality. Here, individual aspects of telemeeting quality are considered by means of five conceptual variables: *Quality of Experience*, referring to the overall telemeeting quality experience in a more general view and including aspects such as overall impression, pleasantness, satisfaction and acceptance; *Speech Communication Quality*, directly related to how well the system enables speech communication and including aspects such as (one-way) voice transmission quality, ease of communication, and conversation effectiveness; *Speech Communication Quality* with a focus on the *Telecommunication Component* by considering questions asking about the (one-way) voice transmission quality; *Speech Communication Quality* with a focus on the *Group-Communication Component* by considering questions asking about ease of communication and conversation effectiveness; and *Cognitive Load*, which represents the *Group-Communication Component*, as it refers to the mental effort to communicate.

Concerning the study on the impact of *Communication Complexity*, two experiments were conducted in which the test methodology of the second experiment was improved based on the first experiment. Focusing on the second experiment, three different aspects of telemeeting quality were considered: *Quality of Experience*, *Speech Communication Quality* with focus on the *Telecommunication Component*, and *Cognitive Load*. Here, the following results were found: As expected, *Communication Complexity* showed a significant impact on *Cognitive Load*; *Technical System Capability* showed a significant impact on *Speech Communication Quality*. Furthermore, an additional cross effect could be observed, as *Technical System Capability* showed also significant impact on *Cognitive Load*. This means, a technically more advanced system (here spatial sound reproduction) can reduce cognitive load under certain conditions. The main finding, however, is the result that both *Communication Complexity* and *Technical System Capability* showed an effect on *Quality of Experience*. This means, not only a technically more advanced system, but also a less complex communicative situation can improve quality experience.

Concerning the study on the impact of *Involvement*, one experiment was conducted, in which three slightly different aspects were considered: *Quality of Experience*, *Speech Communication Quality* with focus on the *Group-Communication Component*, and *Cognitive Load*. Here, the following results were found: As expected, *Involvement* has a significant impact on *Cognitive Load*. However the direction of that effect was unexpected, as test participants appear to perceive actually a lower *Cognitive Load* when they feel more involved. One explanation could be that feeling more involved makes the assessment task more interesting for test participants. This in turn means that their need to force themselves to stay concentrated throughout the test is reduced, which would then explain the reduction in the perceived effort. The main finding, however, is the result that *Involvement* has an impact on *Quality of Experience*, which was visible when comparing the two fundamental test paradigms listening-only and conversation test. This means that the choice of conducting either a listening-only test or a conversation test should be taken with great care, as this choice alone is already influencing results. Another important finding is that the impact of the test paradigm (listening-only vs. conversation test) can be stronger than the impact of the technical conditions under test, even if those condition differences are strong enough to reach statistical significance.

## Perception of Telemeeting Quality with focus on Telecommunication Aspects

---

This chapter contains text passages that either stem from previous texts of the author or that are based on such source texts. For readability purposes those passages are not marked as quotations. The main source documents used for this chapter are:

- Janto Skowronek, Julian Herlinghaus, and Alexander Raake. "Quality Assessment of Asymmetric Multiparty Telephone Conferences: A Systematic Method from Technical Degradations to Perceived Impairments". In: *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*. International Speech Communication Association. Lyon, France, Aug. 2013, pp. 2604–2608
- Janto Skowronek, Anne Weigel, and Alexander Raake. "Quality of Multiparty Telephone Conferences from the Perspective of a Passive Listener". In: *Fortschritte der Akustik - 41. Jahrestagung für Akustik (DAGA)*. Nürnberg, Germany, Mar. 2015
- Janto Skowronek and Alexander Raake. *Report on a study on asymmetric multiparty telephone conferences for a future update of P.1301*. ITU-T Contribution COM 12 - C134 - E. Geneva, Switzerland: International Telecommunication Union, Dec. 2013
- Janto Skowronek. *Documentation on system setup*. Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, June 2013
- Janto Skowronek. *Improved model for audio-only communication*. Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, Dec. 2013
- Janto Skowronek. *Pilot test for audiovisual communication*. Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, Feb. 2014
- Janto Skowronek. *Second test for audiovisual communication*. Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, Oct. 2014
- Janto Skowronek. *Final Project Report – Quality of Multiparty Audio-Video Communication*. Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, Dec. 2014

Since additional documents served as further background material, see Appendix A for the full list of source documents and their contributions to this chapter.

---

### What this chapter is about

The present text is built around two aspects that differentiate a multiparty telemeeting from a conventional two-party (video) telephony call. One aspect is addressed in this chapter: the possibility of encountering asymmetric connections, that is, participants are connected with different equipment or connection properties.

More specifically, the present chapter reports on three studies that investigated the impact of different symmetric and asymmetric connections on quality perception. Considering that quality comprises different aspects, the focus here lies on the perceived system performance, expressed as *Speech Communication Quality* with a focus on the *Telecommunication Component* of telemeeting quality.

## 8.1 Introduction

Assuming a process-oriented perspective, which is here a simplified view on perception, as it was also done in the previous chapter, telemeeting quality perception can be regarded as a process that transforms an input – the telemeeting – into an output – the quality judgment. The previous chapter considered two inputs – the technical system properties and the special communicative situation – and three outputs – *Telemeeting Quality* in a more general view and its two aspects *Telecommunication* and *Group-Communication Component*. The present chapter, however, focuses on one input, the technical system properties and one output, the *Telecommunication Component*. Limiting to one input allows for a deeper investigation of the impact of technical conditions, and especially of asymmetric conditions, on the perceived quality. Limiting to one output allows later, that is in the next chapter, a more precise definition of the target variable when developing multiparty quality models.

Looking at the special aspect of asymmetric conditions, the research question arises how the combination of different technical conditions on the individual connections between interlocutors contribute to the overall telemeeting quality. Such knowledge would be beneficial in different aspects: first, it contributes to a better understanding of the – potentially mutually influencing – impact of different connections on telemeeting quality perception; second, it can provide empirical evidence for the conceptual model in Chapter 5 with its definition of different quality aggregation levels, supporting the practical relevance of those concepts; third, it helps to develop quality prediction models for telemeeting systems that allow to either plan new systems or to monitor systems in operation; and fourth, it would help to develop strategies to improve the quality perception in case of asymmetric conditions with an optimal resource allocation.

To investigate the impact of asymmetric conditions, the studies aimed at measuring the two different aggregation levels of telemeeting quality from the perspective of individual interlocutors: the *Single-Perspective Telemeeting Quality*  $Q_i$  and the *Individual Connection Quality*  $Q_{ij}$ . Furthermore, to obtain a more complete picture due to the possible impact of the communicative situation, the different studies investigated the quality perception using different test paradigms: the first study (consisting of three experiments) applied an audio-only conversation test paradigm, the second study (consisting of two experiments) applied a listening-only test paradigm, and the third study (consisting of two experiments) applied an audiovisual conversation test paradigm.

In the following, the present chapter will first re-capitulate the conceptual input and output variables. Then it will report on the conduction and results of the individual studies.

## 8.2 *Experimental Variables on a Conceptual Level*

To briefly recapitulate the elaborations in Section 7.2, the present text considers *Manipulated Variables* as those aspects of an experiment that are controlled by the investigator, and it considers *Measured Variables* as those aspects that are extracted from the experimental outcome, here the test participants' answers.

In the present chapter, a limited set of the experimental variables from Section 7.2 is used, which are *Technical System Capability* and *Speech Communication Quality*. In addition, the concept of *Quality Aggregation Levels* (see Section 5.3) is included as an additional type of *Measured Variables*. Furthermore, the test paradigm (listening-only vs. conversation test) with its related variable *Involvement* will be considered as a *Manipulated Variable* that is varied across individual experiments.

### 8.2.1 *Focus on Technical System Capability and Speech Communication Quality*

From the different experimental variables that were investigated in Chapter 7, the present chapter limits to the *Technical System Capability* as the *Manipulated Variable* and *Speech Communication Quality* as the *Measured Variables*. This is motivated by the goal to use the collected data also for modeling purposes, as described in Chapter 9. In particular, the focus on one *Manipulated Variable* allows to investigate a broader set of instantiations of this variable, which means here the possibility to test a broader set of *Technical System Capabilities*, which in turn provides more data points for modeling. The focus on one *Measured Variable* allows to confine the set of considered quality aspects, which means to obtain a more precise definition of the target variable to be modeled. In addition, the choice for *Speech Communication Quality* means that the target variable is more focused on the technical aspects of a telemeeting. This in turn confines the modeling goal, as the impact of non-technical communicative aspects are minimized in the target variable, especially when the corresponding questions are designed such that they focus on the *Telecommunication Component*.

### 8.2.2 *Quality Aggregation Levels as Measured Variables*

With the overall intention to investigate the impact of *Technical System Capability* on *Speech Communication Quality*, the present chapter has a strong focus on the impact of technical aspects on the perception of telemeeting quality. Since a telemeeting is essentially providing a set of individual but simultaneous communication channels between multiple interlocutors, the impact of individual connections on quality perception is of primary interest. Especially since the role of individual connections is emphasized when it comes to the perception of asymmetric conditions, which is one of the two main differentiators between a multiparty telemeeting and a conventional two-party call.

In that line of thought, the present chapter describes an investigation of the perception of *Speech Communication Quality* in terms of the two quality aggregation levels *Individual Connection Quality* and *Single-Perspective Telemeeting Quality*. In other words, these two quality aggregation levels define the *Measured Variables* of the experiments described in the next sections.

To briefly restate the concepts of the two quality aggregation levels, *Individual Connection Quality* refers to the quality that a telemeeting participant perceives of an individual connection between two interlocutors, or alternatively between an interlocutor and the telemeeting system. *Single-Perspective Telemeeting Quality* refers to the quality that a telemeeting participant perceives from the whole telemeeting system. For more detailed elaborations on the two quality aggregation levels see Section 5.3.

### 8.2.3 *Involvement as Manipulated Variable across Experiments*

The study on the quality impact of *Involvement* (Section 7.4) showed that the test paradigm, i.e. listening-only test vs. conversation test, can have a significant influence, which may be even larger than the impact of the tested technical conditions. For that reason, the impact of *Technical System Capability* on *Speech Communication Quality* is likely influenced by the test paradigm, and with that also influenced – at least in parts – by the degree of *Involvement*.

Thus, to account for the influence of *Involvement* in terms of the influence of the test paradigm, the present chapter will use both listening-only and conversation tests, in order to obtain a broader picture. In that respect, *Involvement* can be regarded as a *Manipulated Variable* that is varied across individual experiments.

## 8.3 *Quality Impact of an Audio-Only Telemeeting System - Conversation Tests*

This section reports on three experiments investigating the quality impact of *Technical System Capability* by means of audio-only conversation tests, henceforth referred to as Experiments ACT<sub>1</sub>, ACT<sub>2</sub>, and ACT<sub>3</sub>.

### 8.3.1 *Goals*

*Motivation* The choice of audio-only as the communication modality is motivated by the observation that audio-only telemeetings – usually referred to as telephone conferences – are probably the most often used form of telemeetings. The choice of conversation tests as the test paradigm is motivated by the fact that conversation tests put test participants into that special communicative situation that has been observed as one main differentiator from two-party calls.

*Approach* According to the iterative research method (Section 1.4), the study consists of two main experiments ACT<sub>1</sub> & ACT<sub>3</sub>, augmented by one additional side experiment ACT<sub>2</sub>. Conducting two main experiments ACT<sub>1</sub> & ACT<sub>3</sub> allowed to improve the experimental method in the second experiment based on the insights gained in the first experiment. The side experiment ACT<sub>2</sub>, applied the same methodology as experiment ACT<sub>1</sub> and was investigating the specific aspect of loudness<sup>1</sup>.

*Hypotheses* The research goal is to investigate the relation between the *Individual Connection Quality* scores  $Q_{ij}$ , and their relation to the *Single-Perspective Telemeeting Quality*  $Q_i$ . To transform this into a set of concrete hypotheses that can be tested with empirical data, the present text refers to the discussions in Chapter 5 on the potential mutual influence of individual connections and the possible quality aggregation processes.

A first hypothesis concerns the potential mutual influence of the co-present individual connections. Section 5.3 argues that there is such a potential mutual influence both in terms of physical and perceptual impact: a physical mutual influence as the co-presence allows new types of degradations that are not possible in two-party situations (example of echo); a perceptual mutual influence as the co-presence sets a different context than two-party situations and quality perception is known to be context-dependent. While the physical influence can be taken into account when conducting the technical signal path analysis described in Section 6.3, the perceptual mutual influence needs to experimentally be investigated. If such an influence can indeed be observed in empirical data, then it is straight-forward to assume that this influence is not constant across different contexts, i.e. different test conditions. Thus, a first main hypothesis is

H<sub>1</sub>: There is a mutual influence of individual connections on the perception of *Individual Connection Quality*  $Q_{ij}$ , whereas this mutual influence is not constant, i.e. it depends on the actual test condition.

In order to be able to test this hypothesis, it is necessary to define criteria which allow to validate if such a mutual influence exists or not. Those criteria can be found by considering the different possible distributions of impairments across the individual connections. Impairments can be for example on the own connection, on one of the interlocutors' connections, or on both of the interlocutors' connections. With three interlocutors and removing any permutations, eleven combinations are in theory possible. Table 8.1 shows the different combinations and marks those cases with a ✓ that were actually used in the present research. With this set of expected behavior it is possible to test H<sub>1</sub> by verifying if the observed data is significantly different from this behavior (H<sub>1</sub> rejected) or not (H<sub>1</sub> confirmed). In practice that means to check if ratings for not impaired individual connections ( $I_{i,x} = 0$ ) are equal to the reference condition (H<sub>1</sub> rejected) or not (H<sub>1</sub> confirmed) and to check if ratings for connections with

<sup>1</sup> Maxim Spur. "Influences of Loudness Differences in Multiparty Conferencing Quality". Bachelor Thesis. Assessment of IP-based Applications, Technische Universität Berlin, Germany, 2012

Impairment distribution			Expected behavior	Used in present research
$I_{i,i}$	$I_{i,j}$	$I_{i,k}$		
0	0	0	no impairment at all (reference condition)	✓
0	0	1	impairment on one of the interlocutor's connections	✓
0	1	1	same impairments on both of the interlocutor's connections	✓
0	1	2	different impairments on each of the interlocutor's connections	
1	0	1	impairment on one of the interlocutor's connections and same impairment on the own connection	
1	1	1	same impairments on both of the interlocutor's connections and same impairment on the own connection	
1	1	2	different impairments on each of the interlocutor's connections and impairment on the own connection is same as one of the two	
3	0	0	no impairment on any of the interlocutor's connections and impairment on the own connection	
3	0	1	impairment on one of the interlocutor's connections and different impairment on the own connection	✓
3	1	1	same impairments on both of the interlocutor's connections and different impairment on the own connection	
3	1	2	different impairments on each of the interlocutor's connections and different impairment on the own connection	

Table 8.1: Possible distributions of impairments (from the perspective of one interlocutor  $IL_i$ ) across individual connections for the case of three interlocutors.

the same impairments ( $I_{i,x} = I_{i,y}$ ) are equal (H1 rejected) or not (H1 confirmed).

A second hypothesis concerns the possible relation between the two quality aggregation levels. Section 5.3 argues that there are different possibilities to aggregate the perception of the characteristics of individual connections into a quality judgment of the whole telemeeting. One such discussed possibility is that participants form judgments of *Individual Connection Quality*  $Q_{ij}$  and then aggregate them into a judgment of *Single-Perspective Telemeeting Quality*  $Q_i$  ("I perceive one participant distorted – so he/she has a bad connection – and the others without distortions – so they have good connections. Thus, the overall quality is moderate"). However, another possibility is that a person is forming a judgment of  $Q_i$  directly on the basis of the characteristics that the person perceives from the telemeeting, without forming explicit  $Q_{ij}$  judgments of the individual connections ("I perceive one participant distorted and the others without distortions. Thus the overall quality is moderate") or even without decomposing the telemeeting into individual connections at all ("I perceive some distortions. Thus, the overall quality is moderate").

Apparently it is not possible to directly test which of these processes are really taking place in the mind of a person. However, it is possible to test for evidence whether the *Single-Perspective Telemeeting Quality*  $Q_i$  can be expressed as a function of the *Individual Connection Qualities*  $Q_{ij}$  and whether this function shows a consistent behavior across different contexts, i.e. different test conditions. In other words, a second hypothesis tests for evidence that  $Q_i$  can be modeled as a function of  $Q_{ij}$ . While Chapter 9 investigates a number of possible modeling functions, the present chapter focuses on providing first evidence that motivates the modeling attempts later on. For that reason, the present chapter is limited to the simplest possible aggregation

function, a simple average (arithmetic mean). Thus, a second main hypothesis is

H2: An aggregation of *Individual Connection Quality*  $Q_{ij}$  to *Single-Perspective Telemeeting Quality*  $Q_i$  takes place, whereas this aggregation can be modeled as a simple average operation:  $Q_i = 1/N \cdot \sum_{j=1}^N (Q_{ij})$ .

### 8.3.2 Experimental Factors

*Test Tasks* In all three experiments the task for test participants was a conversation test task (“Have a conversation over the test system, and give ratings after each conversation”).

*Conversation Scenarios* The three-party conversation test scenarios proposed by Raake et al.<sup>2</sup> served as basis for the conversation scenarios. Experiments ACT<sub>1</sub> and ACT<sub>2</sub> used shortened versions of those scenarios<sup>3</sup> which essentially removed the open discussion part of the original scenarios. This served the purpose of reducing the time needed to complete a conversation, which means increasing the possible number of conditions per test, and also increasing the comparability of the conversations. Furthermore, based on the practical experience in using those scenarios, ACT<sub>3</sub> used a modified version in which details were optimized to reduce some confusing and unnecessary information for the test participants. Finally, three more scenarios were added to the new scenario set. Both versions of the scenario sets are electronically available<sup>4</sup>.

*Confederate as test participant* In Experiment ACT<sub>1</sub> a female confederate and in Experiment ACT<sub>2</sub> a male confederate complemented the participant groups in case of no-shows of individual test participants; this was necessary four times in ACT<sub>1</sub> and twice in ACT<sub>2</sub>. In Experiment ACT<sub>3</sub>, however, a male confederate was always acting as one of the interlocutors. The intension of this systematic modification of the test protocol was to further increase the comparability of the test calls across test participant groups. This was triggered by the practical experience in running the experiments of ACT<sub>1</sub> and ACT<sub>2</sub>, which showed strong variations in the conversations between different groups, an effect that was also formally measured by Raake et al.<sup>5</sup>. The confederate’s task in ACT<sub>3</sub> was to ensure a smooth conversation and to aim at an equal amount of contributions from all interlocutors. In particular the confederate guided the conversation along the underlying scenario structure, solved misunderstandings, and minimized non-intended discourses. The test participants were informed about the confederate with the information that he was there to help play the scenarios.

*Test Systems* Given the test paradigm of conversation tests, the test systems needed to provide on the one hand a full-working audio-only multiparty communication and on the other hand a full control of that system and easy handling during the experiment sessions. As

<sup>2</sup> Alexander Raake et al. “Listening and conversational quality of spatial audio conferencing”. In: *Proceedings of the AES 40th International Conference*. Tokyo, Oct. 2010

Alexander Raake and Claudia Schlegel. *3CT - 3-party Conversation Test scenarios for conferencing assessment*. 2010. DOI: 10.5281/zenodo.16129

<sup>3</sup> Janto Skowronek, Julian Herlinghaus, and Alexander Raake. “Quality Assessment of Asymmetric Multiparty Telephone Conferences: A Systematic Method from Technical Degradations to Perceived Impairments”. In: *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*. International Speech Communication Association. Lyon, France, Aug. 2013, pp. 2604–2608

Maxim Spur. “Influences of Loudness Differences in Multiparty Conferencing Quality”. Bachelor Thesis. Assessment of IP-based Applications, Technische Universität Berlin, Germany, 2012

<sup>4</sup> Janto Skowronek. *3SCT - 3-party Short Conversation Test scenarios for conferencing assessment (Version 01)*. 2013. DOI: 10.5281/zenodo.16134

Janto Skowronek and Falk Schiffner. *3SCT - 3-party Short Conversation Test scenarios for conferencing assessment (Version 02)*. 2015. DOI: 10.5281/zenodo.16135

<sup>5</sup> Alexander Raake et al. “Listening and conversational quality of spatial audio conferencing”. In: *Proceedings of the AES 40th International Conference*. Tokyo, Oct. 2010

described in Skowronek et al.<sup>6</sup>, Experiment ACT<sub>1</sub> used a telephony setup using off-the-shelf VoIP-telephones (SNOM870) connected via a local router to a conference bridge (Asterisk) running on a Linux laptop. Using an Asterisk plugin, the speech signals were routed via the Jack Audio Connection Kid (JACK) to an audio processing software (PureData). All connections used the G.711 A-law codec and the minimum oneway delay of this test system was 150ms (reference condition). Furthermore, packet loss was realized by applying the TC Filter and Netem softwares to manipulate the packet streams. All signal levels were calibrated with an head and torso simulator (HATS). As described by Spur<sup>7</sup>, Experiment ACT<sub>2</sub> applied in addition an automatic gain control inside the PureData software “in order to create a controlled and level environment for testing loudness differences independently from participant or hardware characteristics.”

In contrast to these two experiments using hardware telephones, ACT<sub>3</sub> applied a proprietary test version of a PC-based VoIP client, provided by Skype (Microsoft), and henceforth referred to as Skype Testing Client. The test client allowed to provide different codecs (SILK, G.711) and audio bandwidths (narrowband, wideband, & super-wideband), and test participants used closed headsets (Beyerdynamic DT790) as end devices, which were connected to the computer via RME sound cards. Again PureData was inserted in the audio path for introducing acoustic distortions, Netem was used for introducing packet loss. The overall oneway delay was 350ms, a value that is above the typically recommended threshold of 150ms according to ITU-T Recommendation G.114<sup>8</sup>, but which is below 400ms, a value that was not critical for the laboratory tests of Schoenenberg<sup>9</sup> using the same type of conversation scenarios. Using head and torso simulator (HATS), the sound reproduction levels were calibrated to 73dB SPL at ear reference point for microphone signals that had a -26dBov digital level for an acoustic sound pressure level of 89dB SPL at the HATS mouth reference point.

*Technical Test Conditions* In light of the fact that a degradation can have different effects for the individual interlocutors (Section 6.3), the technical conditions tested in the experiments are actually determined by two aspects: the technical system characteristics as such and the distribution of those characteristics among the individual connections. Table 8.2 provides a respective overview of all test conditions. Each condition is specified

- (a) in terms of the combination of impairments that participants can perceive from the individual connections (*Impairment Definition & Impairment Description*), using the notation of the method described in Section 6.3, and
- (b) in terms of the technical system characteristics, expressed by means of short descriptions of the technical parameters or, when applicable, by means of the transmission parameters of the E-Model<sup>10</sup>.

<sup>6</sup> Janto Skowronek, Julian Herlinghaus, and Alexander Raake. “Quality Assessment of Asymmetric Multiparty Telephone Conferences: A Systematic Method from Technical Degradations to Perceived Impairments”. In: *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*. International Speech Communication Association. Lyon, France, Aug. 2013, pp. 2604–2608

<sup>7</sup> Maxim Spur. “Influences of Loudness Differences in Multiparty Conferencing Quality”. Bachelor Thesis. Assessment of IP-based Applications, Technische Universität Berlin, Germany, 2012

<sup>8</sup> ITU-T. *Recommendation G.114 - One-way transmission time*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2003

<sup>9</sup> Katrin Schoenenberg. “The Quality of Mediated-Conversations under Transmission Delay”. PhD Thesis. Technische Universität Berlin, Germany, 2016

<sup>10</sup> ITU-T. *Recommendation G.107 - The E-model: a computational model for use in transmission planning*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2014

In addition, the table shows the number of collected observations per condition; a number determined by the applied experimental design corrected by the number of actually observed and valid data points (see later paragraphs for more details).

Motivating the selection of the conditions given in Table 8.2, the common goal for all three experiments was to aim for an optimal balance between number of representative technical characteristics, number of observations per condition, the test effort for participants and the available resources. The motivations for individual choices differed between the three experiments ACT<sub>1</sub> to ACT<sub>3</sub>.

In Experiment ACT<sub>1</sub>, the decision was to test a representative but small set of technical conditions for two reasons. First, the original focus of that study<sup>11</sup> was actually more on the perceptual validation of the technical path analysis described in Section 6.3 than on collecting data for testing the two hypotheses considered here. Second, the boundary constraints in terms of time effort for test participants and available resources was a strong limiting factor, especially since the approach was to let every participant experience every test condition according to ITU-T Recommendation P.1301 Annex F<sup>12</sup>. The optimized experimental design method described in Section 6.4 was not yet developed at that time.

In Experiment ACT<sub>2</sub>, the choice of technical conditions was determined by the specific focus of that study on the impact of loudness differences<sup>13</sup>.

In Experiment ACT<sub>3</sub>, the goal was to test as many technical conditions as possible with the available resources, noting that the author had then worked out the optimized experimental design described in Section 6.4. In addition, the technical characteristics should be representative according to three factors. First, different types of degradations in terms of the end-to-end transmission should be covered, i.e. impairments coming from the acoustic environment (i.e. background noise), from the end device (i.e. signal filtering and echo) and from the network connection (i.e. codecs and packet loss). Second, two instances of each impairment type should be used in order to at least roughly cover the behavior of each impairment type in the quality space. Third, the final choice should be on the one hand relevant for real-life scenarios, i.e. avoiding extremely strong impairments, while on the other hand they should be strong enough to trigger different quality ratings, i.e. allowing for impairments that are slightly stronger than in typical real-life scenarios.

Furthermore, the final choice of conditions, especially in Experiments ACT<sub>1</sub> and ACT<sub>3</sub>, was also dependent on the technical feasibility, i.e. if the test systems actually allowed to realize the different desired conditions. This required some experimentation upfront, given the complexity of the test systems (hardware, software and network), fundamental limits (e.g. computational power, minimum system delay, network bandwidth), and any limitations of the control possibilities (e.g. balancing between the possibility of full-control vs. ensuring a real-time capable communication mode, default behav-

<sup>11</sup> Janto Skowronek, Julian Herlinghaus, and Alexander Raake. "Quality Assessment of Asymmetric Multiparty Telephone Conferences: A Systematic Method from Technical Degradations to Perceived Impairments". In: *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*. International Speech Communication Association. Lyon, France, Aug. 2013, pp. 2604–2608

<sup>12</sup> ITU-T. *Recommendation P.1301 - Subjective quality evaluation of audio and audiovisual telemeetings*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2012

<sup>13</sup> Maxim Spur. "Influences of Loudness Differences in Multiparty Conferencing Quality". Bachelor Thesis. Assessment of IP-based Applications, Technische Universität Berlin, Germany, 2012

Experiment ACT1: Audio-Only Conversation Test, Handsets (monotic), Narrowband codec						
Condition	Impairment combination			Impairment description	Technical parameters	#Obs
<i>Ref1</i>	0	0	0	0: no impairments	Reference: G.711, $T_a = 150ms$ , $OLR = 10dB$	97
<i>L3</i>	0	0	<i>L3</i>	<i>L3</i> : loudness attenuation	$OLR = (10 + 10)dB$	52
<i>EN3</i>	<i>EN3</i>	0	0	<i>EN3</i> : noise in own environment	$P_r = 65dB SPL$	29
<i>TN3</i>	0	0	<i>TN3</i>	<i>TN3</i> : transmitted noise	$P_s = 65dB SPL$	59
<i>Dsym</i>	0	<i>D</i>	<i>D</i>	<i>D</i> : delay	$T_a = (150 + 75)ms$	25
<i>E1</i>	<i>TE1</i>	<i>D</i>	<i>LE1</i>	<i>TE1</i> : talker echo, <i>LE1</i> : listener echo, <i>D</i> : delay	$TELR = 10dB$ , $WEPL' = 10dB$ , $T_a = (150 + 75)ms$	48
<i>P1</i>	0	0	<i>P1</i>	<i>P1</i> : packet loss	$P_{pl} = 10\% \mu_{10} = 1$ (=sparse PL)	54

Experiment ACT2: Audio-Only Conversation Test, Handsets (monotic), Narrowband codec						
Condition	Impairment combination			Impairment description	Technical parameters	#Obs
<i>Ref2</i>	0	0	0	0: no impairments	Reference: G.711, $T_a = 150ms$ , $OLR = 10dB$ , AGC	91
<i>L2</i>	0	0	<i>L2</i>	<i>L2</i> : loudness attenuation	$OLR = (10 + 7)dB$	44
<i>L4</i>	0	0	<i>L4</i>	<i>L4</i> : loudness attenuation	$OLR = (10 + 14)dB$	41
<i>L1</i>	0	0	<i>L1</i>	<i>L1</i> : loudness gain	$OLR = (10 - 7)dB$	48

Experiment ACT3: Audio-Only Conversation Test, Headsets (diotic), Superwideband codec						
Condition	Impairment combination			Impairment description	Technical parameters	#Obs
<i>Ref3</i>	0	0	0	0: no impairments	Reference: $SILK_{SWB}$ , $T_a = 350ms$ , $OLR = 10dB$	58
<i>B1</i>	0	0	<i>B1</i>	<i>B1</i> : wideband codec	$SILK_{WB}$	48
<i>B2</i>	0	0	<i>B2</i>	<i>B2</i> : narrowband codec	G711	46
<i>B2sym</i>	0	<i>B2</i>	<i>B2</i>	<i>B2</i> : narrowband codec	G711	62
<i>FM1</i>	0	0	<i>FM1</i>	<i>FM1</i> : bandpass filtered speech	microphone bandpass 200 Hz - 7000 Hz	45
<i>FL3sym</i>	0	<i>FL3</i>	<i>FL3</i>	<i>FL3</i> : bandpass filtered speech	loudspeaker bandpass 400 Hz - 5000 Hz	47
<i>FM2</i>	0	0	<i>FM2</i>	<i>FM2</i> : bandpass filtered speech	microphone bandpass 400 Hz - 7000 Hz	47
<i>FL4sym</i>	0	<i>FL4</i>	<i>FL4</i>	<i>FL4</i> : bandpass filtered speech	loudspeaker bandpass 800 Hz - 5000 Hz	46
<i>TN1</i>	0	0	<i>TN1</i>	<i>TN1</i> : transmitted noise	$P_s = 60dB(A)$	21
<i>TN2</i>	0	0	<i>TN2</i>	<i>TN2</i> : transmitted noise	$P_s = 65dB(A)$	22
<i>E2</i>	<i>TE2</i>	0	<i>LE2</i>	<i>TE2</i> : talker echo, <i>LE2</i> : listener echo	$TELR = 20dB$ , $WEPL' = 20dB$	48
<i>E3</i>	<i>TE3</i>	0	<i>LE3</i>	<i>TE3</i> : talker echo, <i>LE3</i> : listener echo	$TELR = 30dB$ , $WEPL' = 30dB$	47
<i>P1</i>	0	0	<i>P1</i>	<i>P1</i> : packet loss	$P_{pl} = 10\% \mu_{10} = 1$ (=sparse PL)	62
<i>P2</i>	0	0	<i>P2</i>	<i>P2</i> : packet loss	$P_{pl} = 5\% \mu_{10} = 2$ (=bursty PL)	61
<i>P3</i>	0	0	<i>P3</i>	<i>P3</i> : packet loss	$P_{pl} = 15\% \mu_{10} = 1$ (=sparse PL)	57
<i>P4</i>	0	0	<i>P4</i>	<i>P4</i> : packet loss	$P_{pl} = 10\% \mu_{10} = 2$ (=bursty PL)	57

ior of the test client, fixed settings of VoIP telephones).

### 8.3.3 Method

*Design* The design was driven by a number of constraints in terms of temporal order of stimuli, scenarios, number of technical conditions within a test session, number of observations per condition, overall experimental time per subject and overall test time.

Concerning the scenarios, each scenario should be played only

Table 8.2: Test conditions for the three audio-only conversation tests ACT1 to ACT3. The conditions are defined by the shown combinations of the impairments that each interlocutor perceives from the three individual connections.

once per test group. Eleven scenarios were available for ACT<sub>1</sub> & ACT<sub>2</sub>, 14 scenarios were available for ACT<sub>3</sub>.

Concerning the number of technical conditions within a test session, the design should provide an optimum possible overlap of conditions between subject groups. This is motivated by the following reasoning. It is known in quality research that the experimental context in general has an effect on the test participants' ratings. This is accounted for by various methodological measures such as following standard protocols, having training/familiarization phases, anchor conditions and the like. However, one such context aspect relevant here is the variation of conditions presented during the experiment. If this variation is too low, i.e. not many different conditions are tested, then test participants start to become oversensitive to those few conditions. This could skew results in case of between-subject designs in which subject groups are exposed to different subsets of conditions, an effect that is also referred to as *Range Equalization Bias* Slawomir Zielinski, Francis Rumsey, and Soren Bech. "On Some Biases Encountered in Modern Audio Quality Listening Tests – A Review". In: *Journal of the Audio Engineering Society* 56.6 (2008), pp. 427–451.

Concerning the number of observations per condition, a typical number used in quality assessment tests is to have 24 or 32 observations per condition; in their handbook on telephonometry<sup>14</sup> the ITU-T mentions a number of 30 test participants as rough guidance. However, bearing in mind that one goal of the experiments was to obtain data for modeling purposes, as many data points as feasible would be beneficial. For this purpose, the experimental designs of experiments ACT<sub>1</sub> to ACT<sub>3</sub> should aim for a number of observations in the order of 48 to 64, which is twice the typical range of 24 to 32.

Concerning the overall experimental time per test participant, this time should be limited in order to avoid negative effects due to fatigue, boredom or learning. For that purpose the maximum test time per participant was limited to 70 to 90 minutes in all three experiments. This allowed between nine and 13 test calls given the used conversation scenarios, which also nicely fitted the additional requirement that each scenario is only used once per participant group.

Concerning the overall test time, the time and budget constraints allowed to aim for 20 groups in ACT<sub>1</sub> (maximum 60 observations per condition due to three test participants), nine groups in ACT<sub>2</sub> (maximum 27 observations per condition due to three test participants), and 32 groups in ACT<sub>3</sub> (maximum 64 observations per condition due to two test participants and one confederate).

With these different constraints in mind, ACT<sub>1</sub> used a tenth-order greco-latin design to balance order of conditions, test condition and scenario, repeated twice to cover 20 groups, from which 17 groups were actually conducted. ACT<sub>2</sub> used a ninth-order greco-latin design, repeated once, to cover nine groups, meaning 27 test participants. ACT<sub>3</sub> used a more complex design in which a number of conditions were judged by all test participants (within-subject design),

<sup>14</sup> ITU-T. *Handbook on Telephonometry*. Geneva, Switzerland: International Telecommunication Union, 1992

and other conditions were judged only by subsets of the test participants (between-subject design). The choice of conditions for the within-subject or between-subject design part was based on to their practical relevance from the perspective of the project under which ACT<sub>3</sub> was conducted. The conditions for the within-subject part were the reference condition as well as the different codec and packet loss conditions; the conditions for the between-subject part were the microphone filter, loudspeaker filter, background noise and echo conditions.

*Test participants* The characteristics of the invited test participants are given in Table 8.3.

Participant characteristics	ACT <sub>1</sub>	ACT <sub>2</sub>	ACT <sub>3</sub>
Number of participants (without confederate and any removed subjects)	47	24	63
Age: mean / minimum / maximum	30.5 / 16 / 69	26.6 / 18 / 31	29.7 / 20 / 57
Gender: female / male, in percent (absolute numbers)	72% / 28% (34 / 13)	38% / 62% (9 / 15)	55% / 45% (35 / 28)
Number female / male / mixed gender groups (not considering confederates)	8 / 2 / 7	1 / 2 / 6	8 / 5 / 19
Multiparty experience in professional life (working, studying)	23% (11 of 47)	67% (16 of 24)	56% (35 of 63)
Multiparty experience in private life	19% (9 of 47)	29% (7 of 24)	30% (19 of 63)
Groups with subjects knowing each other (all or some)	8	7	7
Groups with subjects knowing the confederates	0	0	0
Number of subjects with experience with similar listening or telephony experiments	n.a.	50% (12 of 24)	30% (19 of 63)
Hearing impairments (self-reported)	none	none	none

Table 8.3: Test participant statistics for the three audio-only conversation tests ACT<sub>1</sub> to ACT<sub>3</sub>.

*Procedure* In experiment ACT<sub>1</sub> and ACT<sub>3</sub>, participants came for a single 90 minutes session, in experiment ACT<sub>2</sub> for a single 70 minutes session. After a general introduction, the test participants conducted first a “Get to know” call, in which the test supervisor participated as well, and everybody introduced themselves to the others according to a number of questions asked by the test supervisor. In ACT<sub>3</sub> example degradations were presented during that first call to limit the potential noise in the collected data by priming the expectations of the test participants with these example degradations. After the “Get to know” call, the test participants had a training call under reference condition without the test supervisor, in which the participants could get used to playing through the conversation scenarios and after which they answered a trial questionnaire. After that training call, the test participants conducted the test calls according to the conversation scenarios under different technical conditions and answered a questionnaire after each call. According to the different experimental designs, experiment ACT<sub>1</sub> comprised ten test calls, ACT<sub>2</sub> nine test calls, and ACT<sub>3</sub> thirteen test calls. After roughly half the test calls, test participants had a short break of two to five minutes. At the end of the session, the test supervisor conducted a short outtake interview with the test participants.

### 8.3.4 Data Acquisition

*Questionnaires and Measures* According to the iterative research method (Section 1.4), the questionnaire design of ACT<sub>3</sub> was updated compared to ACT<sub>1</sub> and ACT<sub>2</sub>. At the time of conducting the first two experiments<sup>15</sup> a formal categorization of the *Measured Variables* as described in Sections 7.2 and 8.2 in terms of *Speech Communication Quality, Quality of Experience, and Telecommunication and Group-Communication Components* was not yet developed. At the time of conducting Experiment ACT<sub>3</sub>, however, these distinctions were fully considered, which was then used for optimizing the wording of the questions.

In ACT<sub>1</sub> & ACT<sub>2</sub> the questionnaire was inspired by the first experiment on Communication Complexity CC<sub>1</sub> (Section 7.3), as well as standard assessment questionnaires according to the ITU<sup>16</sup>. The questionnaire consisted of four main parts, covering different aspects of quality: telemeeting quality judgment (question Q1.1), judgment of cognitive load (question Q2.1), individual connection quality judgments (questions Q3.1 to Q3.3), and judgments of *Speech Communication Quality* and service satisfaction (questions Q4.1 and Q4.2). In ACT<sub>3</sub>, however, the questionnaire was optimized to focus more specifically on the *Speech Communication Quality with focus on the Telecommunication Component*. For that purpose, all redundant questions and all questions concerning the communicative aspects were removed, and the quality questions were rephrased such that test participants interpret them more as questions concerning the technical quality they perceive. The resulting questionnaire consisted of three main parts: overall telemeeting quality judgment (question Q1.1), individual connection quality judgments (questions Q2.1 to Q2.3), and a free text field for any additional comments (question Q3.1).

Table 8.4 provides English translations of the questions as well as the original German texts.

Table 8.5 gives an overview of the link between the *Measured Variables*, the actual measures used, and descriptions on how they were extracted from the questionnaires, while the following two paragraphs provide detailed explanations of these links.

*Individual Connection Quality* In both questionnaires the wording of the questions about the quality of the individual connections had a focus on the technical aspects of quality due to the keyword “connection”. For that reason, it can be assumed that these questions measure *Speech Communication Quality* with focus on the *Telecommunication Component*.

*Single-Perspective Telemeeting Quality* As mentioned above, the questionnaire of ACT<sub>1</sub> & ACT<sub>2</sub> covered different aspects of quality, whereas the framework of the *Measured Variables* described in Sections 7.2 and 8.2 was not fully developed. Now, with this framework available, the questions can be assigned *in retrospect* to the different *Measured*

<sup>15</sup> Janto Skowronek, Julian Herlinghaus, and Alexander Raake. “Quality Assessment of Asymmetric Multiparty Telephone Conferences: A Systematic Method from Technical Degradations to Perceived Impairments”. In: *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*. International Speech Communication Association. Lyon, France, Aug. 2013, pp. 2604–2608

Maxim Spur. “Influences of Loudness Differences in Multiparty Conferencing Quality”. Bachelor Thesis. Assessment of IP-based Applications, Technische Universität Berlin, Germany, 2012

<sup>16</sup> ITU-T. *Recommendation P.800 - Methods for objective and subjective assessment of quality*. International Standard. Geneva, Switzerland: International Telecommunication Union, 1996

ITU-T. *Recommendation P.805 - Subjective evaluation of conversational quality*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2007

Experiment ACT <sub>1</sub> & ACT <sub>2</sub>		
Part	Question ID	Wording / Description, EN: English translation, DE: German original
1	Q1.1	EN: "What was your personal intuitive <b>overall impression</b> of this telephone conference? – <i>bad ... excellent</i> " DE: "Wie war ihr persönlicher intuitiver <b>Gesamteindruck</b> von dieser Telefonkonferenz? – <i>schlecht ... ausgezeichnet</i> "
2	Q2.1	EN: "It required <i>very much ... very little</i> concentration to follow the conference." DE: "Es erforderte <i>sehr viel ... sehr wenig</i> Konzentration um der Konferenz zu folgen."
3	Q3.1	EN: "The quality of the <b>own</b> connection was <i>bad ... excellent</i> ." DE: "Die Qualität der <b>eigenen</b> Verbindung war <i>schlecht ... ausgezeichnet</i> ."
	Q3.2	EN: "The quality of the connection of <b>Mr./Ms. X</b> was <i>bad ... excellent</i> ." DE: "Die Qualität der Verbindung von <b>Herrn/Frau X</b> war <i>schlecht ... ausgezeichnet</i> ."
	Q3.3	EN: "The quality of the connection of <b>Mr./Ms. Y</b> was <i>bad ... excellent</i> ." DE: "Die Qualität der Verbindung von <b>Herrn/Frau Y</b> war <i>schlecht ... ausgezeichnet</i> "
4	Q4.1	EN: "It required <i>very much ... very little</i> effort to communicate with the other participants." DE: "Es erforderte <i>sehr viel ... sehr wenig</i> Aufwand um mit den Gesprächspartnern kommunizieren zu können."
	Q4.2	EN: "I would find it <i>very unacceptable ... very acceptable</i> to attend a telephone conference with such a system." DE: "Mit einem solchen System einer Telefonkonferenz beizuwohnen würde ich <i>sehr inakzeptabel ... sehr akzeptabel</i> finden."
Experiment ACT <sub>3</sub>		
Part	Question ID	Wording / Description, EN: English translation, DE: German original
1	Q1.1	EN: "What was your personal intuitive <b>overall impression</b> of the system's quality? – <i>extremely bad ... ideal</i> " DE: "Wie war Ihr persönlicher intuitiver <b>Gesamteindruck</b> von der Qualität des Systems? – <i>schlecht ... ausgezeichnet</i> "
2	Q2.1	EN: "Your <b>own connection</b> into the conference was <i>bad ... excellent</i> ." DE: "Ihre <b>eigene Verbindung</b> in die Konferenzschaltung war <i>schlecht ... ausgezeichnet</i> ."
	Q2.2	EN: "The <b>connection of Mr./Ms. X</b> was <i>bad ... excellent</i> ." DE: "Die <b>Verbindung von Herrn/Frau X</b> in die Konferenzschaltung war <i>schlecht ... ausgezeichnet</i> ."
	Q2.3	EN: "The <b>connection of Mr./Ms. Y</b> was <i>bad ... excellent</i> ." DE: "Die <b>Verbindung von Herrn/Frau Y</b> in die Konferenzschaltung war <i>schlecht ... ausgezeichnet</i> ."
3	Q3.1	EN: "Please describe, if and which degradations or special characteristics you perceived. If you have other remarks, please write them down here as well. – <i>Free text</i> " DE: "Bitte beschreiben Sie kurz, ob und welche Störungen oder Besonderheiten Ihnen aufgefallen sind. Wenn Sie sonstige Anmerkungen haben, schreiben Sie diese bitte auch hier auf."

Variables as follows: the question Q1.1 on *Quality* used the phrase "overall impression of this telephone conference" which refers to the quality experience in a wider sense and is therefore assigned to *Quality of Experience*; the question Q2.1 on *Concentration Effort* is according to Chapter 7 a direct self-reported measure of *Cognitive Load*; the question Q4.1 on *Conversation Effort* addresses *Speech Communication Quality* with a focus on the *Group-Communication Component*<sup>17</sup>; and the question Q4.2 on *Acceptance* addresses service satisfaction, which covers quality aspects beyond signal transmission and is therefore assigned to *Quality of Experience*.

As mentioned above, the questionnaire for ACT<sub>3</sub> asked only for quality, whereas the wording was modified to "overall impression of the system's quality" in order to change the focus to *Speech Communication Quality* with an emphasis on the *Telecommunication Component*. The motivation was to increase the comparability of the questions on *Single-Perspective Telemeeting Quality* (Q1.1) and *Individual Connection Quality* (Q3.1 to Q3.3) by ensuring that both types focus on the technical aspects of quality.

*Rating scales* Experiments ACT<sub>1</sub> and ACT<sub>2</sub> used the 5-point Absolute Category Scale (ACR) for all questions. The motivation was

Table 8.4: Summary of the questionnaires used in the three audio-only conversation tests ACT<sub>1</sub>, ACT<sub>2</sub> & ACT<sub>3</sub>.

<sup>17</sup> Similar to the discussion in Section 7.4, *Conversation Effort* is not assigned to *Cognitive Load* but to *Speech Communication Quality* since the origin of this question is a standardized question of conversational quality proposed by the ITU ITU-T. Recommendation P.805 - *Subjective evaluation of conversational quality*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2007

Experiments ACT1 & ACT2		
Measured Variable	Measure	Extraction from questionnaire
Individual Connection Quality		
<i>Speech Communication Quality</i> with focus on <i>Telecommunication Component</i>	<i>Connection Quality</i>	Directly measured with Q3.1, Q3.2, & Q3.3
Single-Perspective Telemeeting Quality		
<i>Cognitive Load</i>	<i>Concentration Effort</i>	Directly measured with Q2.1
<i>Speech Communication Quality</i> with focus on <i>Group-Communication Component</i>	<i>Conversation Effort</i>	Directly measured with Q4.1
<i>Quality of Experience</i>	<i>Quality</i>	Directly measured with Q1.1
	<i>Acceptance</i>	Directly measured with Q4.2
Experiment ACT3		
Measured Variable	Measure	Extraction from questionnaire
Individual Connection Quality		
<i>Speech Communication Quality</i> with focus on <i>Telecommunication Component</i>	<i>Connection Quality</i>	Directly measured with Q2.1, Q2.2, & Q2.3
Single-Perspective Telemeeting Quality		
<i>Speech Communication Quality</i> with focus on <i>Telecommunication Component</i>	<i>System's Quality</i>	Directly measured with Q1.1

to collect the data, especially the quality judgments, according to existing standards<sup>18</sup>. This would facilitate future modeling attempts using existing standardized models, which in turn are also based on the ACR scale. In ACT3, however, test participants rated the question on *Single-Perspective Telemeeting Quality*  $Q_i$  with the 5-point ACR scale, while they rated the questions for the individual connections  $Q_{ij}$  using the extended continuous (EC) rating scale<sup>19</sup>, which was also used for the experiments in Chapter 7. The motivation was to use this change of the visual representation for reducing the chance that test subjects consider the questions concerning the *Single-Perspective Telemeeting Quality* and the *Individual Connection Quality* as essentially the same. The disadvantage of this approach, however, is that a scale transformation is necessary in order to properly compare the ACR ratings of the *Single-Perspective Telemeeting Quality*  $Q_i$  and the EC ratings of the *Individual Connection Quality*  $Q_{ij}$ . Such a scale transformation has been developed by Wältermann et al.<sup>20</sup> and made publicly available by Köster et al.<sup>21</sup>:

$$\widehat{Q}_{ACR} = -0.0262 \cdot Q_{EC}^3 + 0.2368 \cdot Q_{EC}^2 + 0.1907 \cdot Q_{EC} + 1 \quad (8.1)$$

**Data Processing** All measures were represented in such a way that high values reflect something positive, that is high quality or low cognitive load. This representation is straight-forward for the quality and acceptance measures, but it is less intuitive for the effort-related measures, as the ratings are now inverted: a high effort, meaning a high cognitive load, is coded with a low value.

**Data Validation** The data was inspected for any necessary exclusion due to either technical reasons or the participants' rating behavior. Concerning technical reasons, the technical complexity of the test systems was rather high and involved a number of manual operations

Table 8.5: Relation between the Measured Variables (on their conceptual level), the actual measures used for obtaining respective quantitative data, and extraction of the measures from the questionnaire answers.

<sup>18</sup> ITU-T. *Recommendation P.800 - Methods for objective and subjective assessment of quality*. International Standard. Geneva, Switzerland: International Telecommunication Union, 1996

<sup>19</sup> Markus Bodden and Ute Jekosch. *Entwicklung und Durchführung von Tests mit Versuchspersonen zur Verifizierung von Modellen zur Berechnung der Sprachübertragungsqualität*. Project Report (unpublished). Institute of Communication Acoustics, Ruhr-University Bochum, Germany, 1996

ITU-T. *Recommendation P.851 - Subjective quality evaluation of telephone services based on spoken dialogue systems*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2003

<sup>20</sup> Marcel Wältermann, Alexander Raake, and Sebastian Möller. *Comparison between the discrete ACR scale and an extended continuous scale for the quality assessment of transmitted speech*. ITU-T Contribution COM 12 - C 39 - E. Geneva, Switzerland: International Telecommunication Union, Feb. 2009

<sup>21</sup> Friedemann Köster et al. "Comparison between the Discrete ACR Scale and an Extended Continuous Scale for the Quality Assessment of Transmitted Speech". In: *Fortschritte der Akustik (DAGA2015) - 41. Jahrestagung für Akustik*. Nürnberg, Germany, Mar. 2015

of the test supervisor between calls. This means, there was a residual chance for operation errors (< 10% of the cases), despite the efforts made to automate as much as possible by means of scripts and presets. If the test supervisor observed such errors at the very beginning of a call, i.e. before test participants started to speak, then the call was interrupted and re-initiated. In most cases, however, the error could often be detected at earliest when the test participants were already engaged in a call. In such situations interrupting and re-initiating a call is experienced by the test participants as a strong disturbance, which may bias any judgments that the participants would make about the re-initiated call. For that reason, the supervisor let participants continue the call, even though this in turn required to exclude those calls from the data afterwards.

Concerning the rating behavior, inspecting the data for ACT<sub>1</sub> did not reveal any specific outliers in terms of individual ratings or test participants. Hence none of the test participant ratings were excluded from the further analysis. Inspecting the data for ACT<sub>2</sub>, Spur<sup>22</sup> found that one test participant gave all questions in all test calls the same rating, suggesting a strong lack of involvement of that test participant or a strong misunderstanding of the rating task. For that reason, the data of that one test participant was removed from the further analysis. Inspecting the data for ACT<sub>3</sub> did not reveal any specific rating outliers. However, one test participant was excluded since the participants' language skill was not reaching mother-tongue level, and the ratings were nearly constant for all calls. And from another participant, all calls after the break were excluded, since the door of that participant's test room was, as found out after the calls, not properly closed.

Furthermore, sometimes (< 5% of the cases) test participants forgot to answer one individual question from the questionnaires. Such calls need to be completely excluded for the modeling attempts described in Chapter 9. However, to be consistent, those calls were also excluded from the further analysis in the present chapter.

*Data Analysis and Presentation* The first data analysis step was to organize the collected data according to the algorithm described in Section 6.5. To remember, the purpose of that algorithm is to account for the fact that telemeeting participants can have different perceptions of the same call. In other words, the algorithm sorts the collected data such that all calls from all test participants are grouped according to the conditions defined in Table 8.2.

The second and most important analysis step was to determine the expected relations between the *Single-Perspective Telemeeting Quality* score  $Q_i$  and the *Individual Connection Quality* scores  $Q_{ii}$ ,  $Q_{ij}$ , &  $Q_{ik}$ . This is necessary to be able to test the two hypotheses H1 (mutual influence of individual connections exists) and H2 (aggregation is a simple average). The relations between  $Q_i$ ,  $Q_{ii}$ ,  $Q_{ij}$ , and  $Q_{ik}$  can be determined by considering the different possible distributions of impairments across the individual connections shown in Table 8.1,

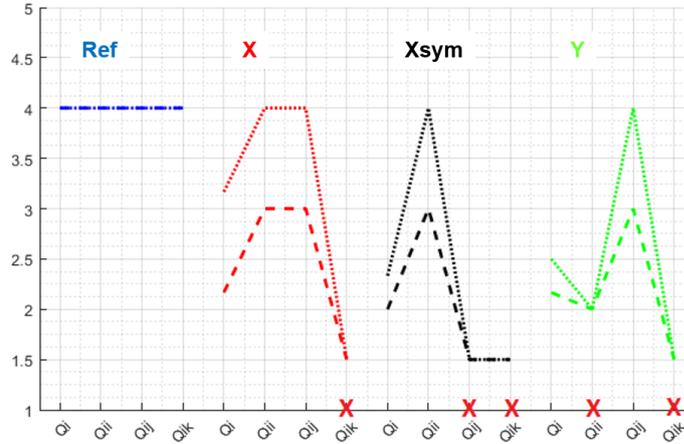
<sup>22</sup> Maxim Spur. "Influences of Loudness Differences in Multiparty Conferencing Quality". Bachelor Thesis. Assessment of IP-based Applications, Technische Universität Berlin, Germany, 2012

and by including the hypothesis H2 into the considerations. Figure 8.1 explains this process by means of representative hypothetic examples and provides a corresponding visualization of those expected relations.

The third analysis step was to apply one-sample t-tests per condition in order to verify whether the observed scores of  $Q_i$ ,  $Q_{ii}$ ,  $Q_{ij}$ , and  $Q_{ik}$  are significantly different from their expected values. As visualization, the author generated errorbar plots showing the mean values and 95% confidence intervals of the observed data, and inserted into the plots the expected values according to the method shown in Figure 8.1. In this visualization, significant differences correspond to cases in which the expected values lie outside the confidence intervals.

Finally, the fourth analysis step was to determine for which conditions the two hypotheses H1 and H2 are confirmed or rejected.

1. Hypothetic examples of the relations between Single-Perspective Telemeeting Quality score  $Q_i$  and the Individual Connection Quality scores  $Q_{ii}$ ,  $Q_{ij}$ , &  $Q_{ik}$ . The examples cover eight possible cases (Ref, X, Xsym, Y, dashed and dotted lines) as outlined below.



The y-axis reflects a typical 5-point quality scale. The x-axis indicates the different ratings, i.e.  $Q_i$ ,  $Q_{ii}$ ,  $Q_{ij}$ , and  $Q_{ik}$ . The red crosses at the x-axis indicate which individual connections are impaired.

2. The dashed and dotted lines refer to different cases concerning the hypotheses H1 and H2.

Dashed lines:

H1 holds (mutual influence exist) & H2 holds (aggregation is simple average:  $Q_i = 1/3 \cdot (Q_{ii} + Q_{ij} + Q_{ik})$ )

Dotted lines:

H1 does not hold (no mutual influence exist) & H2 holds (aggregation is simple average:  $Q_i = 1/3 \cdot (Q_{ii} + Q_{ij} + Q_{ik})$ )

3. The cases **Ref**, **X**, **Xsym**, and **Y** are defined by the different possible distributions of impairments across the individual connections  $ii$ ,  $ij$ , &  $ik$ .

**Ref:** Reference condition  $[I_{i,i} \ I_{i,j} \ I_{i,k}] := [0 \ 0 \ 0]$

All individual connections and the overall telemeeting should be equal and optimal, independently from the two hypotheses H1 and H2.  $\Rightarrow Q_i, Q_{ii}, Q_{ij}$ , and  $Q_{ik}$  are equal and high (here assumed to have a value of 4.0).

**X:** Asymmetric conditions without an impairment on own connection  $[I_{i,i} \ I_{i,j} \ I_{i,k}] := [0 \ 0 \ x_1]$

Impairment  $x_1$  on connection  $ik \Rightarrow Q_{ik}$  lower than the reference value (here assumed to have a value of 1.5)

If H1 does not hold (dotted line), then:

- No impairment on own connection  $ii \Rightarrow Q_{ii}$  equal to the reference value
- No impairment on connection  $ij \Rightarrow Q_{ij}$  equal to the reference value

If H1 holds (dashed line), then:  $Q_{ii}$  &  $Q_{ij}$  are different from reference value (here assumed to have a value of 3.0)

**Xsym:** Symmetric conditions without an impairment on own connection  $[I_{i,i} \ I_{i,j} \ I_{i,k}] = [0 \ x_1 \ x_1]$

Same impairments  $x_1$  on connections  $ij$  &  $ik \Rightarrow Q_{ij}$  &  $Q_{ik}$  lower than the reference value

If H1 does not hold (dotted line), then: No impairment on own connection  $ii \Rightarrow Q_{ii}$  equal to the reference value

If H1 holds (dashed line), then:  $Q_{ii}$  is different from the reference value (here assumed to have a value of 2)

**Y:** Asymmetric conditions with an impairment on own connection  $[I_{i,i} \ I_{i,j} \ I_{i,k}] = [x_2 \ 0 \ x_1]$

Impairment  $x_2$  on connection  $ii \Rightarrow Q_{ii}$  lower than the reference value

Impairment  $x_1$  on connection  $ik \Rightarrow Q_{ik}$  lower than the reference value

If H1 does not hold (dotted line), then: no impairment on connection  $ij \Rightarrow Q_{ij}$  equal to the reference value

If H1 holds (dashed line), then:  $Q_{ij}$  is different from reference value

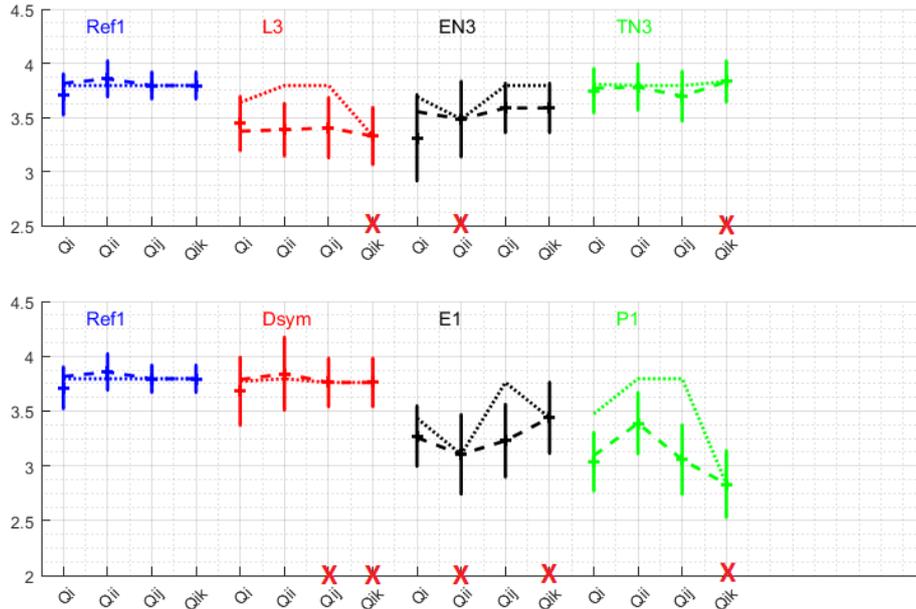
4. In all shown cases, it is assumed that H2 holds. Thus, the shown values for  $Q_i$  are the computed mean values for the hypothetical values of  $Q_{ii}$ ,  $Q_{ij}$ , and  $Q_{ik}$ . If H2 does not hold, then  $Q_i$  has different values than the ones shown.

Figure 8.1: Visualization of expected relations between the Single-Perspective Telemeeting Quality score  $Q_i$  and the Individual Connection Quality scores  $Q_{ii}$ ,  $Q_{ij}$ , and  $Q_{ik}$ .

8.3.5 Results of Experiment ACT1

Figure 8.2 shows the visualization of the results and provides a summary of the analysis; the analysis in full detail is given in Appendix E.1.

1. Visualization of Results and Expectations



The errorbars show the mean values and 95% confidence intervals for the obtained ratings of  $Q_i$ ,  $Q_{ii}$ ,  $Q_{ij}$ , and  $Q_{ik}$ . The red crosses at the x-axis indicate which individual connections are impaired. The dashed lines reflect the expectations when Hypothesis 1 (mutual influence) holds, the dotted lines when Hypothesis 1 does not hold. Both dashed and dotted lines assume that Hypothesis 2 (mean function) holds. The reference condition Ref1 is repeated in each plot for better visual comparison.

Brief explanation of shown condition names (for full details see Table 8.2):

Ref1 = reference condition. L3 = loudness impairment. EN3 = noise in own environment. TN3 = transmitted noise. Dsym = delay (symmetric). E1 = echo. P1 = packet loss.

2. Analysis Summary

Result for reference condition:

$Q_i, Q_{ii}, Q_{ij}, Q_{ik}$  are all essentially equal and optimal.  $\Rightarrow$  Ref1 serves as reference condition for this test.

Results concerning Hypothesis H1 (mutual influence), see dotted lines.

Condition	Result	Rationale
EN3, Dsym	×	$Q_{ij} = Q_{ik}$ are not significantly different from the expected value.
TN3	×	$Q_{ii}$ & $Q_{ij}$ are not significantly different from the expected value.
L3, P1	✓	$Q_{ii}$ & $Q_{ij}$ are significantly different from the expected values.
E1	✓	$Q_{ij}$ is significantly different from the expected value.

Results concerning Hypothesis H2 (mean function), see dashed lines.

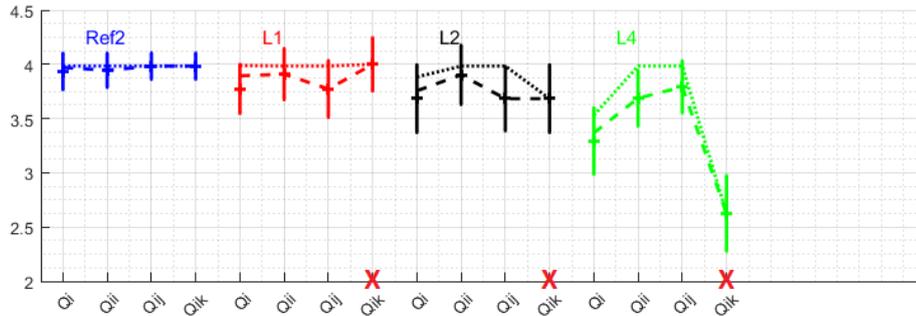
Condition	Result	Rationale
All	✓	$Q_i$ is not significantly different from the expected value.

Figure 8.2: Results for Experiment ACT1.

8.3.6 Results of Experiment ACT2

Figure 8.3 shows the visualization of the results and provides a summary of the analysis; the analysis in full detail is given in Appendix E.2.

1. Visualization of Results and Expectations



The errorbars show the mean values and 95% confidence intervals for the obtained ratings of  $Q_i$ ,  $Q_{ii}$ ,  $Q_{ij}$ , and  $Q_{ik}$ . The red crosses at the x-axis indicate which individual connections are impaired. The dashed lines reflect the expectations when Hypothesis 1 (mutual influence) holds, the dotted lines when Hypothesis 1 does not hold. Both dashed and dotted lines assume that Hypothesis 2 (mean function) holds.

Brief explanation of shown condition names (for full details see Table 8.2):

Ref2 = reference condition. L1, L2, L4 = loudness impairments.

2. Analysis Summary

Result for reference condition:

$Q_i, Q_{ii}, Q_{ij}, Q_{ik}$  are all essentially equal and optimal.  $\Rightarrow$  Ref2 serves as reference condition for this test.

Results concerning Hypothesis H1 (mutual influence), see dotted lines.

Condition	Result	Rationale
L1	×	$Q_{ii}$ & $Q_{ij}$ are not significantly different from the expected values.
L2	✓	$Q_{ij}$ is significantly different from the expected value.
L4	✓	$Q_{ii}$ is significantly different from the expected value.

Results concerning Hypothesis H2 (mean function), see dashed lines.

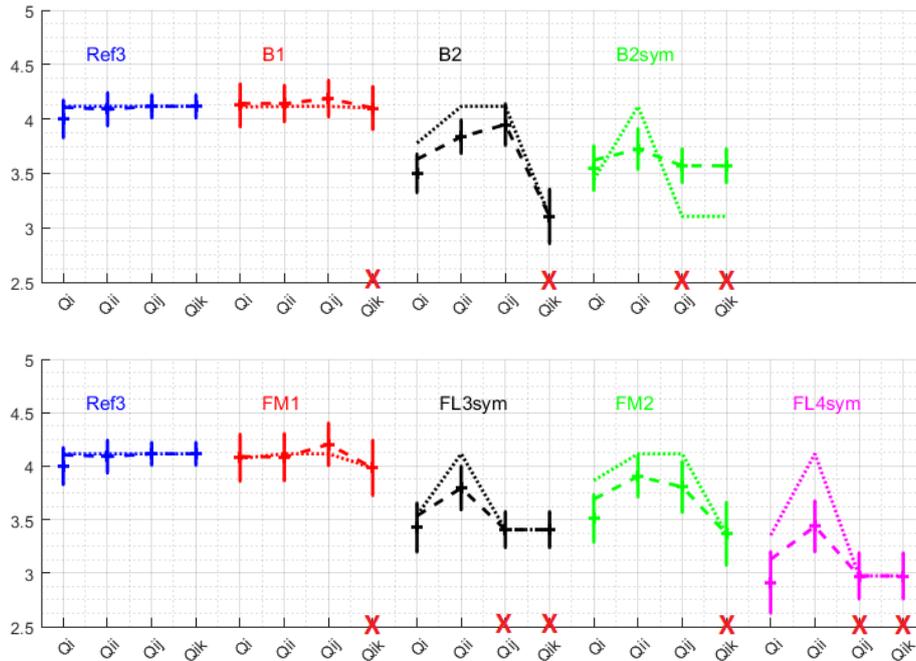
Condition	Result	Rationale
All	✓	$Q_i$ is not significantly different from the expected value.

Figure 8.3: Results for Experiment ACT2.

8.3.7 Results of Experiment ACT3

Figures 8.4 and 8.5 show the visualization of the results and provide summaries of the analysis; the analysis in full detail is given in Appendix E.3.

1. Visualization of Results and Expectations



The errorbars show the mean values and 95% confidence intervals for the obtained ratings of  $Q_i$ ,  $Q_{ii}$ ,  $Q_{ij}$ , and  $Q_{ik}$ . The red crosses at the x-axis indicate which individual connections are impaired. The dashed lines reflect the expectations when Hypothesis 1 (mutual influence) holds, the dotted lines when Hypothesis 1 does not hold. Both dashed and dotted lines assume that Hypothesis 2 (mean function) holds. The reference condition Ref3 is repeated in each plot for better visual comparison.

Brief explanation of shown condition names (for full details see Table 8.2):

Ref3 = reference condition. B1, B2, B2sym = audio bandwidth limitation of codecs ("sym" = symmetric). FM1, FM2 = bandpass filtered microphone signal. FL3sym, FL4sym bandpass filtered loudspeaker signal ("sym" = symmetric).

2. Analysis Summary

Result for reference condition:

$Q_i$ ,  $Q_{ii}$ ,  $Q_{ij}$ ,  $Q_{ik}$  are all essentially equal and optimal.  $\Rightarrow$  Ref3 serves as reference condition for this test.

Results concerning Hypothesis H1 (mutual influence), see dotted lines.

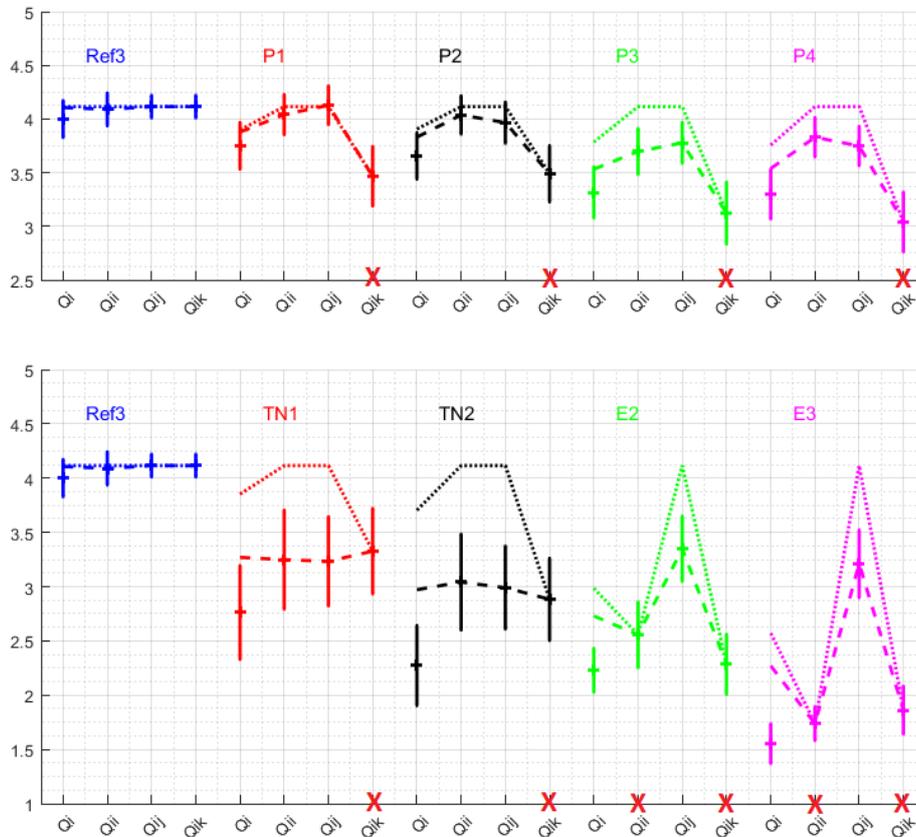
Condition	Result	Rationale
B1, FM1	×	$Q_{ii}$ & $Q_{ij}$ are not significantly different from the expected values.
B2,FL3sym,FL4sym	✓	$Q_{ii}$ is significantly different from the expected value.
B2sym, FM2	✓	$Q_{ii}$ & $Q_{ij}$ are significantly different from the expected values.

Results concerning Hypothesis H2 (mean function), see dashed lines.

Condition	Result	Rationale
All	✓	$Q_i$ is not significantly different from the expected value.

Figure 8.4: Results for Experiment ACT3, Part 1.

1. Visualization of Results and Expectations



The errorbars show the mean values and 95% confidence intervals for the obtained ratings of  $Q_i$ ,  $Q_{ii}$ ,  $Q_{ij}$ , and  $Q_{ik}$ . The red crosses at the x-axis indicate which individual connections are impaired. The dashed lines reflect the expectations when Hypothesis 1 (mutual influence) holds, the dotted lines when Hypothesis 1 does not hold. Both dashed and dotted lines assume that Hypothesis 2 (mean function) holds. The reference condition Ref<sub>3</sub> is repeated in each plot for better visual comparison.

Brief explanation of shown condition names (for full details see Table 8.2):

Ref<sub>3</sub> = reference condition. P<sub>1</sub>, P<sub>2</sub>, P<sub>3</sub>, P<sub>4</sub> = packet loss. TN<sub>1</sub>, TN<sub>2</sub> = transmitted noise. E<sub>2</sub>, E<sub>3</sub> = echo.

2. Analysis Summary

Result for reference condition:

$Q_i$ ,  $Q_{ii}$ ,  $Q_{ij}$ ,  $Q_{ik}$  are all essentially equal and optimal.  $\Rightarrow$  Ref<sub>3</sub> serves as reference condition for this test.

Results concerning Hypothesis H<sub>1</sub> (mutual influence), see dotted lines.

Condition	Result	Rationale
P <sub>1</sub> , P <sub>2</sub>	×	$Q_{ii}$ & $Q_{ij}$ are not significantly different from the expected values.
P <sub>3</sub> , P <sub>4</sub> , TN <sub>1</sub> , TN <sub>2</sub>	✓	$Q_{ii}$ & $Q_{ij}$ are significantly different from the expected values.
E <sub>2</sub> , E <sub>3</sub>	✓	$Q_{ij}$ is significantly different from the expected value.

Results concerning Hypothesis H<sub>2</sub> (mean function), see dashed lines.

Condition	Result	Rationale
P <sub>1</sub> , P <sub>2</sub>	✓	$Q_i$ is not significantly different from the expected value.
P <sub>3</sub> , P <sub>4</sub> , TN <sub>1</sub> , TN <sub>2</sub> , E <sub>2</sub> , E <sub>3</sub>	×	$Q_i$ is significantly different from the expected value.

Figure 8.5: Results for Experiment ACT<sub>3</sub>, Part 2.

### 8.3.8 Discussion

*Review of Results* Table 8.6 summarizes the main findings for the three audio-only conversation tests ACT<sub>1</sub>, ACT<sub>2</sub>, and ACT<sub>3</sub>.

Experiment	Conditions		Hypothesis H1	Hypothesis H2
	Asymmetric	Symmetric		
ACT <sub>1</sub>	TN <sub>3</sub>	EN <sub>3</sub> , Dsym	×	✓
	L <sub>3</sub> , P <sub>1</sub> , E <sub>1</sub>	–	✓	✓
ACT <sub>2</sub>	L <sub>1</sub>	–	×	✓
	L <sub>2</sub> , L <sub>4</sub>	–	✓	✓
ACT <sub>3</sub>	B <sub>1</sub> , FM <sub>1</sub> , P <sub>1</sub> , P <sub>2</sub>	–	×	✓
	B <sub>2</sub> , FM <sub>2</sub>	FL <sub>3sym</sub> , FL <sub>4sym</sub> , B <sub>2sym</sub>	✓	✓
	P <sub>3</sub> , P <sub>4</sub> , TN <sub>1</sub> , TN <sub>2</sub> , E <sub>2</sub> , E <sub>3</sub>	–	✓	×

Hypothesis H<sub>1</sub> is confirmed in 16 out of 24 tested cases, and rejected in eight cases. This means that a clear evidence is found that a mutual influence of the individual connections exists, while such influence depends on the actual technical condition. Searching for a systematic behavior that could explain when a condition shows that influence or not, the Table 8.6 shows that the type of impairment may be a reason, as all noise and echo conditions show a mutual influence. However, the type of impairment is most likely not the only reason, since different conditions of other types, e.g. packet loss P<sub>1</sub> to P<sub>4</sub> or loudness impairments L<sub>1</sub> to L<sub>4</sub>, sometimes show that influence and sometimes not. Apparently, also the strength of the impairments correlates with the existence of a mutual influence of individual connections. This can be seen in the visualizations in Figures 8.2 to 8.5: those conditions with quality scores closer to the reference condition do not show the mutual influence, those conditions with quality scores further away from the reference condition do show the mutual influence. However, it is not clearly visible whether a stronger impairment necessarily means a stronger mutual impact.

Hypothesis H<sub>2</sub> is confirmed in 18 of 24 tested cases, and rejected in six cases. This means that the *Single-Perspective Telemeeting Quality*  $Q_i$  can be expressed in many cases as a simple average of the *Individual Connection Quality* scores  $Q_{ij}$  of the individual connections. However, there are technical conditions in which this simple relation does not hold. As for the mutual influence, next to some specific impairment types (noise and echo), the strength of an impairment appears to correlate with the need to express the relation between  $Q_i$  and  $Q_{ij}$  with a more sophisticated function.

*Review of the Approach and Open Questions for Future Work* There are a number of open questions concerning the experimental method that this section will address now: the limitation in terms of communicative aspects, the potential effects of changing the experimental method, and rating on multiple scales.

Concerning the limitation of communicative aspects, the experiments considered one degree of communication complexity, as the experiments considered only three-party communication using sce-

Table 8.6: Overview of the results for the audio-only conversation tests ACT<sub>1</sub>, ACT<sub>2</sub>, & ACT<sub>3</sub>, concerning the two hypotheses H<sub>1</sub> & H<sub>2</sub>. Confirmed hypotheses are denoted with a ✓, rejected hypotheses with a ×.

narios with the same structure and with similar cognitive demand (except for possible minor differences between the scenarios of ACT1 and ACT2 and the scenarios of ACT3). Hence, future work could repeat these experiments with a different number of interlocutors, different conversation test scenarios or a different test paradigm (conversation test vs. listening-only test). The latter two aspects will be addressed in Sections 8.4 and 8.5.

Concerning the changes of the experimental method, the results of experiments ACT1 and ACT2 revealed only cases in which Hypothesis 2 is confirmed, i.e.  $Q_i$  can be expressed as a simple mean of  $Q_{ij}$ . This triggered a review and modification of the experimental method in experiment ACT3, which eventually revealed a few cases (6 of 24 over all three experiments) in which Hypothesis H2 was rejected. The changes from ACT1 & ACT2 to ACT3 targeted a reduction of any measurement noise (conversation scenarios, confederate as test participant, example impairments in training call) and an increase of the number of test conditions per experiment (design method), and the chosen conditions as such. Since there were multiple changes between the two test methods, it is not possible to identify individual aspects of the method that may have caused the differences in the results. This would require dedicated experiments in future work. Instead, one can conclude that applying the *whole package* of changes was sufficient to *detect* those few conditions in which the *Single-Perspective Telemeeting Quality*  $Q_i$  is more than a simple mean of the *Individual Connection Quality* scores  $Q_{ij}$ .

Concerning the rating on multiple scales, the applied questionnaires asked questions on both aggregation levels  $Q_i$  and  $Q_{ij}$  on the same questionnaire. Here, it might be that test participants do form some average of  $Q_{ij}$  simply due to the presentation of both  $Q_i$  and  $Q_{ij}$  on the same questionnaire, despite the usage of different scale designs and questions. To investigate this aspect, future work could apply a different experimental method in which participants rate per test call or stimuli either  $Q_i$  or  $Q_{ij}$ . This could be achieved by using a between-subject design in which the two questions are split across test participants, or a within-subject design using multiple sessions, one per question.

#### 8.4 Quality Impact of an Audio-Only Telemeeting System - Listening-Only Tests

This section reports on two experiments investigating the quality impact of *Technical System Capability* by means of listening-only tests, henceforth referred to as Experiments LOT1 and LOT2.

##### 8.4.1 Goals

*Motivation* In Skowronek et al.<sup>23</sup> we discussed two motivations for conducting listening-only tests in addition to the audio-only conversation tests described in the previous section. First, in real-life, it

<sup>23</sup> Janto Skowronek, Anne Weigel, and Alexander Raake. "Quality of Multi-party Telephone Conferences from the Perspective of a Passive Listener". In: *Fortschritte der Akustik - 41. Jahrestagung für Akustik (DAGA)*. Nürnberg, Germany, Mar. 2015

is often the case that only a few persons are actively contributing while a number of participants are just listening to the conversation. Conversation tests are less suited for such a use case if they are not specifically designed to put participants into such a passive listener role. Second, it is known in the field that test subjects are less sensitive to quality impairments in conversation tests than in listening-only tests<sup>24</sup>, which was also shown for multiparty scenarios in Section 7.4. A third motivation refers to the modeling described in Chapter 9. Since typical listening-only tests usually apply stimuli that are shorter than full conversations, more technical conditions can be tested in a listening-only test than in a corresponding conversation test. From the modeling perspective this means that listening-only tests have the advantage to obtain more data with the same amount of resources compared to conversation tests.

*Approach* According to the iterative research method (Section 1.4), the study consists of two experiments LOT1 and LOT2, allowing to improve the experimental and analysis method in the second experiment based on the insights gained in the first experiment. The focus of the first experiment was to gain experience in the generation of proper listening test stimuli for multiparty telemeetings<sup>25</sup> and to gain first insights on the quality perception from the perspective of a passive participant<sup>26</sup>. The focus of the second experiment was to test a large number of technical conditions, which partly have also been used in the conversation tests, in order to increase the knowledge on multiparty perception comparing listening-only and conversation tests, and to further facilitate the modeling attempts described in Chapter 9 by including additional technical conditions.

*Hypotheses* The hypotheses are the same two investigated in the previous section:

- H1: There is a mutual influence of individual connections on the perception of *Individual Connection Quality*  $Q_{ij}$ , whereas this mutual influence is not constant, i.e. it depends on the actual test condition.
- H2: An aggregation of *Individual Connection Quality*  $Q_{ij}$  to *Single-Perspective Telemeeting Quality*  $Q_i$  takes place, whereas this aggregation a simple average operation:  $Q_i = 1/N \cdot \sum_{j=1}^N (Q_{ij})$ .

#### 8.4.2 Experimental Factors

*Test Tasks* The task for test participants was in all three experiments a listening-only test task (“Listen to the test stimuli and give ratings after each stimulus”).

*Speech Material* To account for the special communicative situation as one main differentiator between multiparty and two-party calls, ITU Recommendation P.1301 advises to use stimuli that “sufficiently resemble the conversational situation.”<sup>27</sup> Bearing this in

<sup>24</sup> Sebastian Möller. *Assessment and Prediction of Speech Quality in Telecommunications*. Kluwer Academic publishers, 2000

ITU-T. *Recommendation P.805 - Subjective evaluation of conversational quality*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2007

<sup>25</sup> Anne Weigel. “Beurteilung der Qualität von Telefonkonferenzen mit asymmetrischen Verbindungen aus der Perspektive eines passiven Zuhörers”. Bachelor Thesis. Assessment of IP-based Applications, Technische Universität Berlin, Germany, 2014

<sup>26</sup> Janto Skowronek, Anne Weigel, and Alexander Raake. “Quality of Multiparty Telephone Conferences from the Perspective of a Passive Listener”. In: *Fortschritte der Akustik - 41. Jahrestagung für Akustik (DAGA)*. Nürnberg, Germany, Mar. 2015

<sup>27</sup> ITU-T. *Recommendation P.1301 - Subjective quality evaluation of audio and audiovisual telemeetings*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2012

mind, Weigel<sup>28</sup> generated the speech stimuli with a length of about 40s, each containing a meaningful excerpt of a longer three-party conversation, whereas each excerpt featured only two interlocutors. The speech material was taken from recordings from Experiment ACT<sub>3</sub> for which the test participants gave their written consent in reusing those recordings in other tests. Thus, the speech material is based on the same scenarios used in ACT<sub>3</sub>.

The result of the selection and editing process were 43 recordings with relatively equal amount of contributions from both speakers. One speaker in each recording was always the male confederate of ACT<sub>3</sub>; this was motivated by making it easier for test participants to assign the questions on the individual connections to the right speaker. The second speaker was a male test participant from ACT<sub>3</sub> in 80% and a female test participant in 20% of the cases.

Concerning the technical realization of the recordings during ACT<sub>3</sub>, the headset microphone signals were digitally recorded in Puredata before any further processing (degradations, coding, transmission) was applied. Thus, the recordings consisted of individual tracks of each interlocutor at the best possible quality provided by the used test system: high-quality microphone characteristics of the headsets (Beyerdynamic DT790), high-quality AD conversion (RME Multiface) at 44100kHz / 16 bit. The loudspeaker signals, i.e. the speech signals after all processing was applied, were recorded as well. Since the recordings were made on three individual computers – one per interlocutor in ACT<sub>3</sub> – the author synchronized those recordings by means of an algorithm based on the correlation function and exploiting both the microphone and loudspeaker signals.

Concerning the technical realization of the stimuli for LOT<sub>1</sub>, the stimuli were processed by Weigel<sup>29</sup>, who used a professional audio production software (ProTools) for the electro-acoustic degradations and an ITU software toolbox<sup>30</sup> for the codec and packet loss degradations.

Concerning the technical realization of the stimuli for LOT<sub>2</sub>, the stimuli were processed with the same test system as ACT<sub>3</sub> (Beyerdynamic Headsets, RME soundcards, Puredata, Skype Testing Client, Netem, see Section 8.3). However, to reduce the risk of technical problems during the test, the author processed with the help of a student co-worker the recordings offline as follows: The edited recordings were inserted at send-side in Puredata into the digital signal at the same position in the signal path used for recording, then transmitted over the test system, and then recorded in Puredata before sending them to the loudspeaker outputs. During the test, the processed recordings were played back to the test participants via the earphones of the Beyerdynamics headsets, whereas playback signal levels were calibrated to match those applied in the conversation test. That means, the speech material is processed with the exact same technical processing chain from microphones to earphones as it was used in the conversation test.

<sup>28</sup> Anne Weigel. "Beurteilung der Qualität von Telefonkonferenzen mit asymmetrischen Verbindungen aus der Perspektive eines passiven Zuhörers". Bachelor Thesis. Assessment of IP-based Applications, Technische Universität Berlin, Germany, 2014

<sup>29</sup> Anne Weigel. "Beurteilung der Qualität von Telefonkonferenzen mit asymmetrischen Verbindungen aus der Perspektive eines passiven Zuhörers". Bachelor Thesis. Assessment of IP-based Applications, Technische Universität Berlin, Germany, 2014

<sup>30</sup> ITU-T. *Recommendation G.191 - Software tools for speech and audio coding standardization*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2010

*Technical Conditions* As for the audio-only conversation tests, the technical conditions are determined by the two aspects *combination of impairments on the individual connections* and the *technical system characteristics* as such. Table 8.7 provides a respective overview of all test conditions, the technical characteristics expressed as E-model<sup>31</sup> or ProTools parameters.

Motivating the selection of the conditions, the common goal was – as for the conversation tests – to optimally balance the number of technical characteristics and observations per condition and the test effort. The motivations for certain choices though differed between the two experiments LOT<sub>1</sub> to LOT<sub>2</sub>. In Experiment LOT<sub>1</sub>, the decision was to test a representative but small set of technical conditions, given that the experiment's primary goal was to gain first experience in conducting a multiparty-relevant listening-test. In Experiment LOT<sub>2</sub>, the decision was to cover as many conditions of ACT<sub>3</sub> as the experimental design would allow, whereas each tested technical system characteristic was to be tested in an asymmetric configuration, i.e. one connection with impairment and one connection without impairment, and in a symmetric configuration, i.e. both connections with impairment.

#### 8.4.3 Method

*Design* With 43 recordings of about 40s, and a typical listening-only test time of one hour, it was possible to fit all recordings into one test session per participant. This allowed to choose a within-subject design, i.e. all conditions are presented to all test participants.

In LOT<sub>1</sub>, three recordings were used as sources for the training stimuli and 32 recordings were used as sources for the test stimuli. In Skowronek. et al.<sup>32</sup> we described the distribution of the nine technical conditions across the 32 stimuli: "Per asymmetric condition, each technical impairment was applied four times, twice on the confederate (the one speaker present in all recordings), twice on the other speakers. Per symmetric condition, each technical impairment was applied twice, since the confederate and the other speakers are affected simultaneously. The reference condition was applied four times in order to have roughly 10% of the stimuli processed with the reference condition."

In LOT<sub>2</sub>, the author chose a different approach to distribute technical conditions across the speakers. Instead of testing few technical conditions and repeating them in one test session with different speakers, each technical condition was presented only once, which allowed for a set of 37 technical conditions per session, plus five stimuli for training. However, this could lead to interaction effects for particular combinations of technical condition and speakers, if all test participants are exposed to the same set of condition-speaker combinations. To reduce such effects, two such sets of combinations were distributed across the test participants. Those sets were based on a pseudo-random assignment of technical condition to speaker,

<sup>31</sup> ITU-T. *Recommendation G.107 - The E-model: a computational model for use in transmission planning*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2014

<sup>32</sup> Janto Skowronek, Anne Weigel, and Alexander Raake. "Quality of Multiparty Telephone Conferences from the Perspective of a Passive Listener". In: *Fortschritte der Akustik - 41. Jahrestagung für Akustik (DAGA)*. Nürnberg, Germany, Mar. 2015

Experiment LOT1: Listening-Only Test, Headphones (diotic), Wideband codec						
Condition	Impairment combination		Impairment description		Technical parameters	#Obs
<i>Ref5</i>	—	0	0	0: no impairments	Reference: <i>G722</i> , <i>OLR</i> = 10dB	192
<i>B2</i>	—	0	<i>B2</i>	<i>B2</i> : narrowband codec	<i>G711</i>	96
<i>B2sym</i>	—	<i>B2</i>	<i>B2</i>	<i>B2</i> : narrowband codec	<i>G711</i>	48
<i>P6</i>	—	0	<i>P5</i>	<i>P5</i> : packet loss	$P_{pl} = 5\%$ $\mu_{10} = 1$ (=random PL)	95
<i>B2symP5</i>	—	<i>B2</i>	<i>B2P5</i>	<i>B2</i> : narrowband codec, <i>P5</i> : packet loss	<i>G711</i> , $P_{pl} = 5\%$ $\mu_{10} = 1$ (=random PL)	96
<i>E4</i>	—	0	<i>LE4</i>	<i>LE4</i> : listener echo	(ProTools:) Echo Gain = 8%, $T_a = 245ms$	95
<i>E4sym</i>	—	<i>LE4</i>	<i>LE4</i>	<i>LE4</i> : listener echo	(ProTools:) Echo Gain = 8%, $T_a = 245ms$	48
<i>R</i>	—	0	<i>R</i>	<i>R</i> : reverb	(ProTools:) Small room, Pre-delay = 33ms, Decay = 82ms	96
<i>CN</i>	—	0	<i>CN</i>	<i>CN</i> : speech correlated noise	(ProTools:) Pink noise -20dB, gate parameters: threshold = -47.1dB, attack = 33.4ms, hold = 1200ms, release = 799.8ms	95

Experiment LOT2: Listening-Only Test, Headphones (diotic), Superwideband codec						
Condition	Impairment combination		Impairment description		Technical parameters	#Obs
<i>Ref3</i>	—	0	0	0: no impairments	Reference: <i>SILK<sub>SWB</sub></i> , $T_a = 350ms$ , <i>OLR</i> = 10dB	75
<i>TN1</i>	—	0	<i>TN1</i>	<i>TN1</i> : transmitted noise	$P_s = 60dB(A)$	25
<i>TN2</i>	—	0	<i>TN2</i>	<i>TN2</i> : transmitted noise	$P_s = 65dB(A)$	25
<i>B2</i>	—	0	<i>B2</i>	<i>B2</i> : narrowband codec	<i>G711</i>	50
<i>B2sym</i>	—	<i>B2</i>	<i>B2</i>	<i>B2</i> : narrowband codec	<i>G711</i>	25
<i>FL4sym</i>	—	<i>FL4</i>	<i>FL4</i>	<i>FL4</i> : bandpass filtered speech	loudspeaker bandpass 800 – 5000Hz	25
<i>FM1</i>	—	0	<i>FM1</i>	<i>FM1</i> : bandpass filtered speech	microphone bandpass 200 – 7000Hz	50
<i>FM1sym</i>	—	<i>FM1</i>	<i>FM1</i>	<i>FM1</i> : bandpass filtered speech	microphone bandpass 200 – 7000Hz	25
<i>FM2</i>	—	0	<i>FM2</i>	<i>FM2</i> : bandpass filtered speech	microphone bandpass 400 – 7000Hz	50
<i>FM2sym</i>	—	<i>FM2</i>	<i>FM2</i>	<i>FM2</i> : bandpass filtered speech	microphone bandpass 400 – 7000Hz	25
<i>FM4</i>	—	0	<i>FM4</i>	<i>FM4</i> : bandpass filtered speech	microphone bandpass 800 – 5000Hz	50
<i>FM4sym</i>	—	<i>FM4</i>	<i>FM4</i>	<i>FM4</i> : bandpass filtered speech	microphone bandpass 800 – 5000Hz	25
<i>FM4P3</i>	—	0	<i>FM4P3</i>	<i>FM4</i> : bandpass filtered speech, <i>P3</i> : packet loss	microphone bandpass 800 – 5000Hz, $P_{pl} = 15\%$ $\mu_{10} = 1$ (=random PL)	50
<i>FM4P3sym</i>	—	<i>FM4P3</i>	<i>FM4P3</i>	<i>FM4</i> : bandpass filtered speech, <i>P3</i> : packet loss	microphone bandpass 800 – 5000Hz, $P_{pl} = 15\%$ $\mu_{10} = 1$ (=random PL)	50
<i>E2</i>	—	0	<i>LE2</i>	<i>LE2</i> : listener echo	$WEPL' = 20dB$	50
<i>E2sym</i>	—	<i>LE2</i>	<i>LE2</i>	<i>LE2</i> : listener echo	$WEPL' = 20dB$	25
<i>E3</i>	—	0	<i>LE3</i>	<i>LE3</i> : listener echo	$WEPL' = 30dB$	50
<i>E3sym</i>	—	<i>LE3</i>	<i>LE3</i>	<i>LE3</i> : listener echo	$WEPL' = 30dB$	25
<i>P1</i>	—	0	<i>P1</i>	<i>P1</i> : packet loss	$P_{pl} = 10\%$ $\mu_{10} = 1$ (=sparse PL)	50
<i>P1sym</i>	—	<i>P1</i>	<i>P1</i>	<i>P1</i> : packet loss	$P_{pl} = 10\%$ $\mu_{10} = 1$ (=sparse PL)	25
<i>P3</i>	—	0	<i>P3</i>	<i>P3</i> : packet loss	$P_{pl} = 15\%$ $\mu_{10} = 1$ (=sparse PL)	50
<i>P3sym</i>	—	<i>P3</i>	<i>P3</i>	<i>P3</i> : packet loss	$P_{pl} = 15\%$ $\mu_{10} = 1$ (=sparse PL)	25
<i>P4</i>	—	0	<i>P4</i>	<i>P4</i> : packet loss	$P_{pl} = 10\%$ $\mu_{10} = 2$ (=bursty PL)	50
<i>P4sym</i>	—	<i>P4</i>	<i>P4</i>	<i>P4</i> : packet loss	$P_{pl} = 10\%$ $\mu_{10} = 2$ (=bursty PL)	25

whereas it was avoided that the same condition-speaker combinations occur in both sets.

*Test participants* The characteristics of the invited test participants are given in Table 8.8.

Table 8.7: Test conditions for the two listening-only tests LOT1 and LOT2. The conditions are defined by the shown combinations of the impairments that each interlocutor perceives from the three individual connections.

Participant characteristics	LOT <sub>1</sub>	LOT <sub>2</sub>
Number of participants	24	25
Age: mean / minimum / maximum	31 / 19 / 52	33.6 / 22 / 54
Gender: female / male, in percent (absolute numbers)	50% / 50% (12 / 12)	72% / 28% (18 / 7)
Multiparty experience in professional or private life	29% (7 of 24)	n.a.
Number of subjects with experience with similar listening or telephony experiments	70% (17 of 24)	n.a.
Hearing impairments (self-reported)	none	none

*Procedure* In both tests, participants came for a single 60 minutes session. After a general introduction, the test participants had a training of three stimuli in LOT<sub>1</sub> and five stimuli in LOT<sub>2</sub>, in which a subset of the technical conditions were presented, and in which the participants answered trial questionnaires. Then the participants listened to the test stimuli and answered after each stimulus the questionnaire. To avoid fatigue, the test participants had a short break after half of the stimuli: Weigel<sup>33</sup> opted for a fixed duration of 10 minutes in LOT<sub>1</sub>; the author opted for a minimum duration of 2 minutes in LOT<sub>2</sub>. At the end of the session, the test supervisor conducted a short outtake interview with the test participants.

#### 8.4.4 Data Acquisition

*Questionnaires and Rating Scales* Both listening-only tests applied the same questionnaire used in the conversation test ACT<sub>3</sub>. Thus both tests measured *Speech Communication Quality with focus on the Telecommunication Component* for both *Single-Perspective Telemeeting Quality*  $Q_i$  and *Individual Connection Quality*  $Q_{ij}$ ; and both tests applied two different rating scales: Absolute Category Rating (ACR) scale for  $Q_i$  and Extended Continuous (EC) scale for  $Q_{ij}$ . See Section 8.3 for more details, in particular Tables 8.4 and Table 8.5 on the questions and Equation 8.1 on the scale transformation that was applied to map the EC scale to the ACR scale.

However, there is one substantial change compared to the conversation test. The question on the own connection was deleted, since this question is very hard to answer in a listening-only test. Referring back to the discussions in Sections 5.3 and 6.5, even in conversation tests quality judgments on the own connection are in most cases rather hypothetical and rely on indirect cue-based quality features (e.g. conversation flow, remarks of interlocutors), at least if there are no impaired feedback signals (e.g. side tone degradations, talker echo). In case of a listening-only test, this hypothetical character would be even stronger, which makes such a question even more abstract for test participants.

*Data Processing and Validation* All measures were represented in such a way that high values reflect high quality. Furthermore, inspecting the questions did not reveal any specific outliers in terms of unusual rating behavior of test participants. Hence none of the test participants were excluded from the further analysis. Concerning the removal of

Table 8.8: Test participant statistics for the two listening-only tests LOT<sub>1</sub> & LOT<sub>2</sub>.

<sup>33</sup> Anne Weigel. "Beurteilung der Qualität von Telefonkonferenzen mit asymmetrischen Verbindungen aus der Perspektive eines passiven Zuhörers". Bachelor Thesis. Assessment of IP-based Applications, Technische Universität Berlin, Germany, 2014

data points in case a test participant forgot to answer all questions – an effect that happened in <5% of the conversation test questionnaires – this was only necessary in two cases for LOT1. In LOT2, the paper questionnaire was replaced by an electronic questionnaire<sup>34</sup> running on a tablet, which had nearly the same visual design, but forced test participants to answer all questions before continuing with the next stimulus.

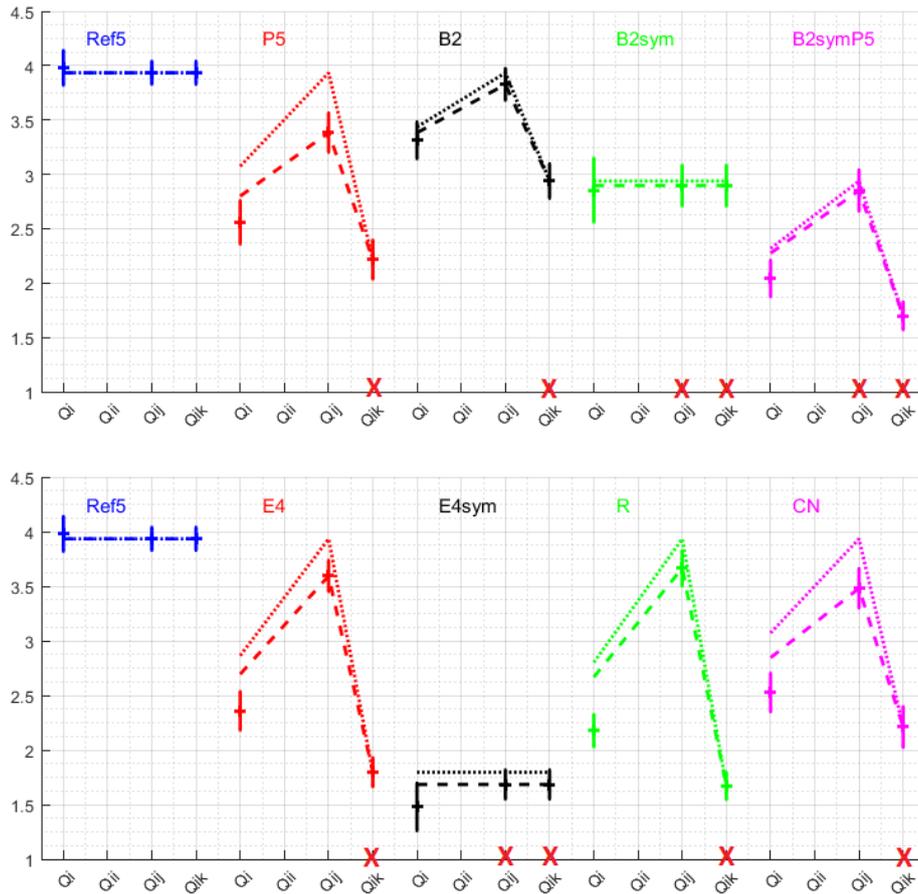
*Data Analysis and Presentation* The data analysis comprised the same four steps conducted for the audio-only conversation tests: (a) organize the collected data according to the algorithm described in Section 6.5; (b) determine the expected relations between the *Single-Perspective Telemeeting Quality* score  $Q_i$  and the *Individual Connection Quality* scores  $Q_{ij}$  and  $Q_{ik}$ , noting that the quality  $Q_{ii}$  of the own connection was excluded from the questionnaires and thus also from the respective expectations; (c) apply one-sample t-tests to check for significant deviations between expected and observed values; and (d) verify per condition the hypotheses H1 and H2. For details see Section 8.3, in particular Figure 8.1 on the expected relations between  $Q_i$ ,  $Q_{ij}$  and  $Q_{ik}$ .

<sup>34</sup> The electronic questionnaire was kindly prepared by D. Guse and colleagues from the Quality and Usability Lab of the Technische Universität Berlin.

8.4.5 Results of Experiment LOT<sub>1</sub>

Figure 8.6 shows the visualization of the results and provides a summary of the analysis; the analysis in full detail is given in Appendix F.1.

## 1. Visualization of Results and Expectations



The errorbars show the mean values and 95% confidence intervals for the obtained ratings of  $Q_i$ ,  $Q_{ii}$ ,  $Q_{ij}$ , and  $Q_{ik}$ . The red crosses at the x-axis indicate which individual connections are impaired. The dashed lines reflect the expectations when Hypothesis 1 (mutual influence) holds, the dotted lines when Hypothesis 1 does not hold. Both dashed and dotted lines assume that Hypothesis 2 (mean function) holds. The reference condition Ref5 is repeated in each plot for better visual comparison.

Brief explanation of shown condition names (for full details see Table 8.7):

Ref5 = reference condition. P5 = packet loss impairment. B2, B2sym = narrow-band codec ("sym" = symmetric). B2symP5 = narrow-band codec (symmetric) and packet loss (asymmetric). E4, E4sym = listener echo ("sym" = symmetric). R = room reverberation. CN = speech-correlated noise.

## 2. Analysis Summary

Result for reference condition:

$Q_i$ ,  $Q_{ij}$ ,  $Q_{ik}$  are all essentially equal and optimal.  $\Rightarrow$  Ref5 serves as reference condition for this test.

Results concerning Hypothesis H<sub>1</sub>, see dotted lines.

Conditions	Result	Rationale
P5, E4, R, CN	✓	$Q_{ij}$ is significantly different from the expected value.
B2, B2sym, B2symP5, E4sym	×	$Q_{ij}$ is not significantly different from the expected value.

Results concerning Hypothesis H<sub>2</sub>, see dashed lines.

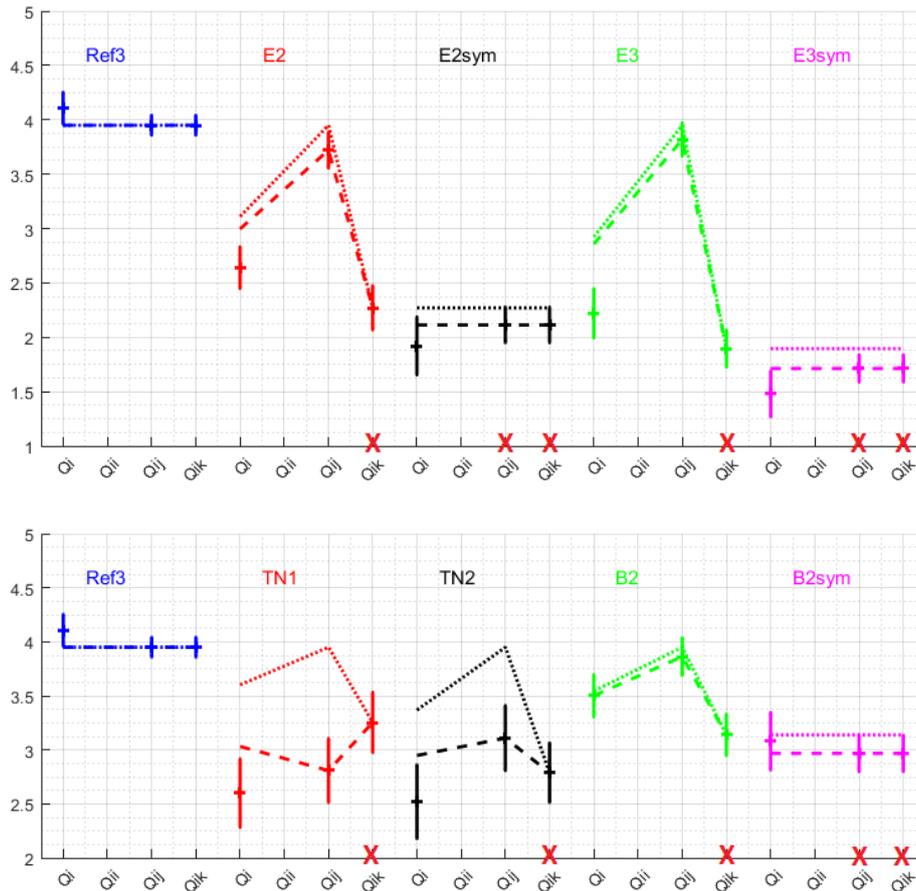
Conditions	Result	Rationale
B2, B2sym, E4sym	✓	$Q_i$ is not significantly different from the expected value.
P5, B2symP5, E4, R, CN	×	$Q_i$ is significantly different from the expected value.

Figure 8.6: Results for Experiment LOT<sub>1</sub>.

8.4.6 Results of Experiment LOT2

Figures 8.7 to 8.9 show the visualization of the results and provide summaries of the analysis; the analysis in full detail is given in Appendix F.2.

1. Visualization of Results and Expectations



The errorbars show the mean values and 95% confidence intervals for the obtained ratings of  $Q_i$ ,  $Q_{ii}$ ,  $Q_{ij}$ , and  $Q_{ik}$ . The red crosses at the x-axis indicate which individual connections are impaired. The dashed lines reflect the expectations when Hypothesis 1 (mutual influence) holds, the dotted lines when Hypothesis 1 does not hold. Both dashed and dotted lines assume that Hypothesis 2 (mean function) holds. The reference condition Ref3 is repeated in each plot for better visual comparison.

Brief explanation of shown condition names (for full details see Table 8.7):

Ref3 = reference condition. E2, E2sym, E3, E3sym = listener echo ("sym" = symmetric). TN1, TN2 = transmitted noise. B2, B2sym = narrow-band codec ("sym" = symmetric).

2. Analysis Summary

Result for reference condition:

$Q_i, Q_{ij}, Q_{ik}$  are all essentially equal and optimal.  $\Rightarrow$  Ref3 serves as reference condition for this test.

Results concerning Hypothesis H1, see dotted lines.

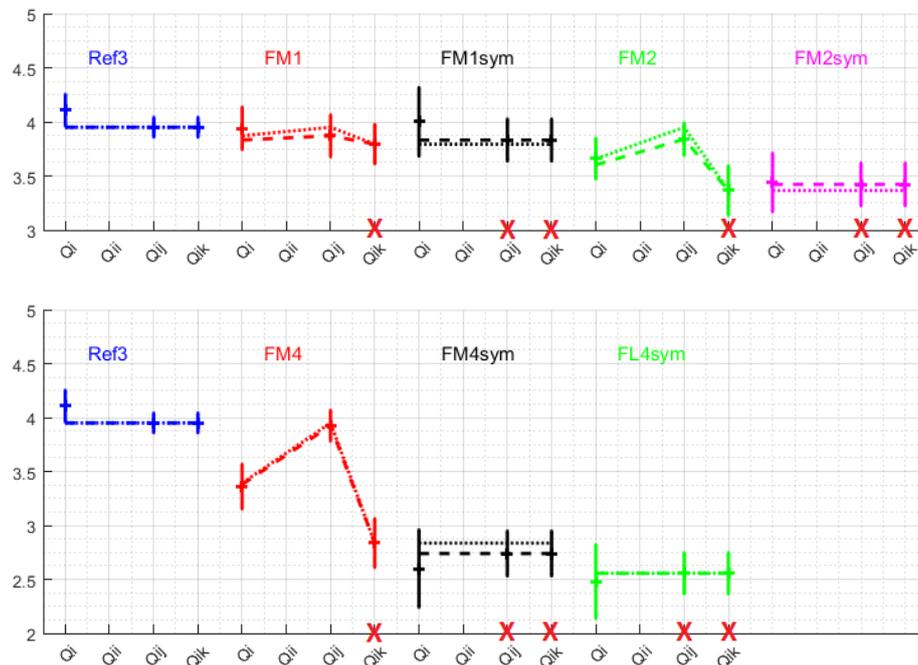
Condition	Result	Rationale
E2, E3sym, TN1, TN2, B2sym	✓	$Q_{ij}$ is significantly different from the expected value.
E2sym, E3, B2	×	$Q_{ij}$ is not significantly different from the expected value.

Results concerning Hypothesis H2, see dashed lines.

Condition	Result	Rationale
E2sym, B2, B2sym	✓	$Q_i$ is not significantly different from the expected value.
E2, E3, E3sym, TN1, TN2	×	$Q_i$ is significantly different from the expected value.

Figure 8.7: Results for Experiment LOT2, Part 1.

1. Visualization of Results and Expectations



The errorbars show the mean values and 95% confidence intervals for the obtained ratings of  $Q_i$ ,  $Q_{ii}$ ,  $Q_{ij}$ , and  $Q_{ik}$ . The red crosses at the x-axis indicate which individual connections are impaired. The dashed lines reflect the expectations when Hypothesis 1 (mutual influence) holds, the dotted lines when Hypothesis 1 does not hold. Both dashed and dotted lines assume that Hypothesis 2 (mean function) holds. The reference condition Ref<sub>3</sub> is repeated in each plot for better visual comparison.

Brief explanation of shown condition names (for full details see Table 8.7):

Ref<sub>3</sub> = reference condition. FM<sub>1</sub>, FM<sub>1sym</sub>, FM<sub>2</sub>, FM<sub>2sym</sub>, FM<sub>4</sub>, FM<sub>4sym</sub> = bandpass filtered microphone signal (“sym” = symmetric). FL<sub>4sym</sub> = bandpass filtered loudspeaker signal (symmetric).

2. Analysis Summary

Result for reference condition:

$Q_i, Q_{ij}, Q_{ik}$  are all essentially equal and optimal.  $\Rightarrow$  Ref<sub>3</sub> serves as reference condition for this test.

Results concerning Hypothesis H<sub>1</sub>, see dotted lines.

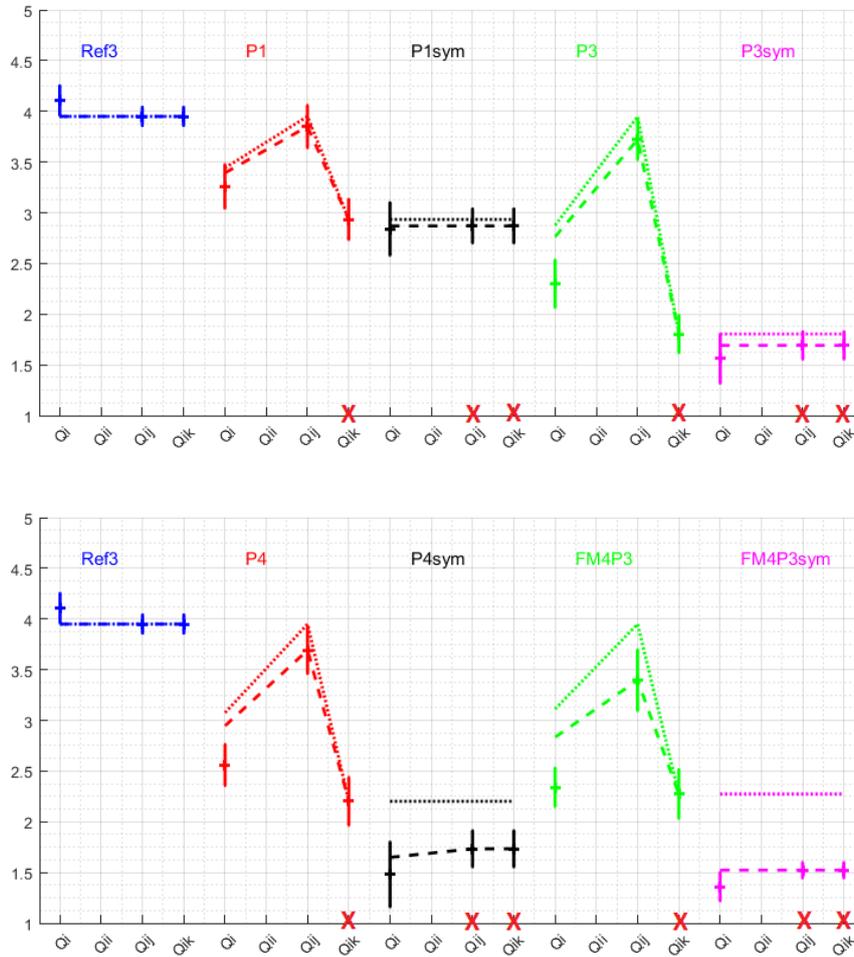
Condition	Result	Rationale
FM <sub>1</sub> , FM <sub>1sym</sub> , FM <sub>2</sub> , FM <sub>2sym</sub> , FM <sub>4</sub> , FM <sub>4sym</sub> , FL <sub>4sym</sub>	×	$Q_{ij}$ is not significantly different from the expected value.

Results concerning Hypothesis H<sub>2</sub>, see dashed lines.

Condition	Result	Rationale
FM <sub>1</sub> , FM <sub>1sym</sub> , FM <sub>2</sub> , FM <sub>2sym</sub> , FM <sub>4</sub> , FM <sub>4sym</sub> , FL <sub>4sym</sub>	✓	$Q_i$ is not significantly different from the expected value.

Figure 8.8: Results for Experiment LOT<sub>2</sub>, Part 2.

1. Visualization of Results and Expectations



The errorbars show the mean values and 95% confidence intervals for the obtained ratings of  $Q_i$ ,  $Q_{ii}$ ,  $Q_{ij}$ , and  $Q_{ik}$ . The red crosses at the x-axis indicate which individual connections are impaired. The dashed lines reflect the expectations when Hypothesis 1 (mutual influence) holds, the dotted lines when Hypothesis 1 does not hold. Both dashed and dotted lines assume that Hypothesis 2 (mean function) holds. The reference condition Ref3 is repeated in each plot for better visual comparison.

Brief explanation of shown condition names (for full details see Table 8.7):

Ref3 = reference condition. P1, P1sym, P3, P3sym, P4, P4sym = packet loss ("sym" = symmetric). FM4P3, FM4P3sym = bandpass filtered microphone signal and packet loss ("sym" = symmetric).

2. Analysis Summary

Result for reference condition:

$Q_i$ ,  $Q_{ij}$ ,  $Q_{ik}$  are all essentially equal and optimal.  $\Rightarrow$  Ref3 serves as reference condition for this test.

Results concerning Hypothesis H1, see dotted lines.

Condition	Result	Rationale
P3, P4, P4sym, FM4P3, FM4P3sym	✓	$Q_{ij}$ is significantly different from the expected value.
P1, P1sym, P3sym	×	$Q_{ij}$ is not significantly different from the expected value.

Results concerning Hypothesis H2, see dashed lines.

Condition	Result	Rationale
P1, P1sym, P3sym, P4sym	✓	$Q_i$ is not significantly different from the expected value.
P3, P4, FM4P3, FM4P3sym	×	$Q_i$ is significantly different from the expected value.

Figure 8.9: Results for Experiment LOT2, Part 3.

## 8.4.7 Discussion

*Review of Results* Table 8.9 summarizes the main findings for the two listening-only tests LOT1 and LOT2.

Experiment	Conditions		Hypothesis H1	Hypothesis H2
	Asymmetric	Symmetric		
LOT1	B2	B2sym, E4sym	×	✓
	P5, E4, R, CN	–	✓	×
	–	B2symP5	×	×
LOT2	B2, FM1, FM2, FM4, P1	E2sym, FM1sym, FM2sym, FM4sym, FL4sym, P1sym, P3sym	×	✓
	–	B2sym, P4sym	✓	✓
	E2, TN1, TN2, P3, P4, FM4P3	E3sym, FM4P3sym	✓	×
	E3	–	×	×

Hypothesis H1 is confirmed in 14 out of 31 tested cases, and rejected in 17 cases. This means that a clear evidence is found that a mutual influence of the individual connections exists, while such influence depends on the actual technical condition. Concerning a systematic behavior that could explain when a condition shows that influence or not, Table 8.9 shows that it is most likely not a function of the type of impairment, since conditions of the same type, e.g. packet loss P1 to P4 or echo E2 to E4, sometimes show that influence and sometimes not. The strength of an impairment, a possible explanation found for the audio-only conversation tests, is only partially sufficient to explain the mutual influence. This can be seen on the visualizations in Figures 8.2 to 8.5: Those conditions with quality scores closer to the reference condition do not show the mutual influence; and while there are also conditions with lower quality scores that do show the mutual influence (e.g. LOT1: E4, R, LOT2: P3, P4), other conditions with lower quality scores do not (e.g. LOT1: B2symP5, E4sym, LOT2: E3), or do show the mutual influence (e.g. LOT1: E4, R, LOT2: P3, P4). Furthermore, the selection of conditions allows to investigate symmetry effects in terms of differences between asymmetric and symmetric configurations. Thus, it is possible to check, if symmetry can explain the influence. However, symmetry does not do this, since Table 8.9 shows that impairments in both symmetric and asymmetric configurations are distributed among both cases H1 confirmed and H1 rejected.

Hypothesis H2 is confirmed in 17 of 31 tested cases, and rejected in 14 cases. This means that the *Single-Perspective Telemeeting Quality*  $Q_i$  can be expressed in many cases as a simple average of the *Individual Connection Quality* scores  $Q_{ij}$  of the individual connections. However, there are also many technical conditions in which this simple relation does not hold. In case of asymmetric configurations, the strength of an impairment appears to correlate with the need to express the relation between  $Q_i$  and  $Q_{ij}$  with a more sophisticated function, since Figures 8.2 to 8.5 show that conditions with *Single-Perspective Telemeeting Quality* scores  $Q_i < 3$  differ from the expected values. However, this behavior is not linked with the existence of a mutual influence,

Table 8.9: Overview of the results for the listening-only tests LOT1 & LOT2, concerning the two hypotheses H1 & H2. Confirmed hypotheses are denoted with a ✓, rejected hypotheses with a ×.

since Table 8.6 shows that the conditions cover all combinations of a confirmed or rejected hypothesis H1 and H2. Concerning symmetry as a further possible explanation, no consistent trend is visible: Some impairments in asymmetric and symmetric conditions do not support a symmetry effect, as either both confirm or reject H2 (LOT1: B2 vs. B2sym, LOT2: FM1 vs. FM1sym, FM2 vs. FM2sym, FM4 vs. FM4sym, P1 vs. P1sym, FM4P3sym vs. FM4P3sym). Other impairments do support a symmetry effect as they are distributed across the two cases *H2 confirmed* and *H2 rejected* (LOT2: B2 vs. B2sym, P3 vs. P3sym).

*Review of the Approach and Open Questions for Future Work* There are a number of open questions concerning the experimental method that this section will address now: rating on multiple scales and the limitation in terms of communicative aspects.

The aspect of multiple rating scales has been discussed in Section 8.3 for the audio-only conversation tests, and the same discussion applies here: it might be that test participants do form some average of  $Q_{ij}$  simply due to the presentation of both  $Q_i$  and  $Q_{ij}$  on the same questionnaire, despite the usage of different scale designs and questions; and future studies could validate this by different experimental approaches (e.g. split sessions, between-subject designs).

The aspect of communicative aspects is a non-trivial issue in the context of a listening-only test. One motivation for conducting a listening-only test is to simulate a situation in which the test participant is taking the role of a passive listener, a typical situation in real-life scenarios. However, this simulation only partially reflects the real-world case, given that the test participants are not really interested in the actual information shared in the recordings. While this potentially puts test participants into an analytic listening and rating mode, meaning they might be more critical about impairments, it also reduces the potential impact of the communicative situation on the quality rating. On the one hand, this means test participants might focus even more on the desired *Telecommunication Component* of *Telemeeting Quality*. On the other hand, this means the test is getting closer to a conventional listening-only test for traditional telephony scenarios, meaning that the contribution of any multiparty-specific aspects to the quality rating is reduced. In other words, the listening-only test potentially leads to a stronger focus on technical quality with a potentially reduced focus on multiparty aspects of quality. For future work, it may be difficult to empirically prove such considerations, even with dedicated experiments. For that reason, the proposal for first steps for future work is to conduct more listening-only tests by different laboratories. This would allow to exchange different results and interpretations between researchers and practitioners in the field in order to assess the practical relevance of these considerations.

## 8.5 Quality Impact of an Audiovisual Telemeeting System - Conversation Tests

This section reports on two experiments investigating the quality impact of *Technical System Capability* by means of audiovisual conversation tests, henceforth referred to as Experiments AVCT<sub>1</sub> and AVCT<sub>2</sub>.

### 8.5.1 Goals

*Motivation* The choice of investigating also the audiovisual communication modality is motivated by the observation that more and more PC-based and mobile-phone-based solutions for audiovisual telemeetings are available on the mass market. Thus, audiovisual telemeetings are – after audio-only telemeetings – the second most used form of telemeetings.

*Approach* According to the iterative research method (Section 1.4), the study consists of two experiments AVCT<sub>1</sub> and AVCT<sub>2</sub>. Conducting two experiments allowed to improve the experimental method in the second experiment based on the insights gained in the first experiment. The focus of the first experiment was to gain experience in the conduction of a multiparty audiovisual test and to obtain first insights into the quality perception of audiovisual telemeetings. The focus of the second experiment was to reach a good balance between the strength of audio-only and video-only degradations in the test.

Furthermore, the scope of the experiments was to investigate quality aggregation in terms of multiparty aggregation, i.e.  $Q_i$  as function of  $Q_{ij}$ . The scope was not do investigate quality aggregation in terms of modality aggregation, i.e. audiovisual quality as function of audio quality and video quality. For that reason, the approach was to ask test participants only for audiovisual quality, and not to collect ratings for audio-only and audio-video quality. This had the advantage to not confuse or even overwhelm test participants, which would be the case when asking them to judge quality for all three modalities (audio-only, video-only, audiovisual) for each of the two multiparty aggregation levels ( $Q_i, Q_{ij}$ ). Nevertheless, focusing on audiovisual quality still allowed to investigate the combination of audio-only and video-only impairments to audiovisual impairments to a certain extent by comparing the quality ratings for the corresponding technical audio-only, video-only and audiovisual degradations.

*Hypotheses* The hypotheses are the same two investigated in the previous sections:

- H1: There is a mutual influence of individual connections on the perception of *Individual Connection Quality*  $Q_{ij}$ , whereas this mutual influence is not constant, i.e. it depends on the actual test condition.
- H2: An aggregation of *Individual Connection Quality*  $Q_{ij}$  to *Single-Participant Telemeeting Quality*  $Q_i$  takes place, whereas this aggregation

a simple average operation:  $Q_i = 1/N \cdot \sum_{j=1}^N (Q_{ij})$ .

### 8.5.2 Experimental Factors

*Test Tasks* The task for test participants was in all two experiments a conventional conversation test task (“Have a conversation over the test system, and give ratings after each conversation”).

*Conversation Scenarios* Two types of conversation scenarios were used: a survival task game and a celebrity name guessing game. Originally coming from team building domain, survival tasks have been used in studies on face-to-face and mediated group communication. In this task, test participants are supposed to be in a survival scenario which resulted from an accident-like event, e.g. a plain crash in the desert. The participants’ task is to rank for a list of 12 to 15 items that are found in the debris how important those items are for surviving the scenario. Typical durations to complete that task are one to two hours, Thompson & Covert<sup>35</sup> for instance allowed test participants to complete the survival task within 77 minutes. A modified version featuring four scenarios (desert & winter, sea, and moon) was proposed and validated by Gros & Filandre<sup>36</sup> for the purpose of quality assessment tests for audiovisual multiparty telemeeting systems, which was also included into the ITU-T Recommendation P.1301<sup>37</sup>. One modification was to distribute the items among the participants, i.e. each participant found different items in that scenario, forcing each participant to contribute to the conversation. Another modification was to agree on six items useful for survival, shortening the task duration down to roughly four to five minutes<sup>38</sup>. In order to test more than four conditions with such scenarios, Skowronek & Spur<sup>39</sup> extended this set with three more scenarios (mountains, swamp, cave), and made minor adjustments to the original scenarios after a small pilot test with six participant groups. The full set of scenarios is electronically available.

The second task, a celebrity name guessing game, was proposed by Schoenenberg et al.<sup>40</sup> as a delay sensitive conversation task for two-party video-telephony, and the author adapted this game to the three party case as follows: In each round of this game, test participants are supposed to be a celebrity, and the task is to guess their own name. For that purpose, a test participant may ask only questions that can be answered by the other two interlocutors with yes or no, e.g. “Am I female?” or “Am I an actor?”. Furthermore, the asking test participant had a paper card with the celebrity’s initials, which he or she clipped to his or her shirt. The other interlocutors had the corresponding full name on a paper card that they were not allowed to hold into the camera. The initials helped the asking participant in guessing the name and helped all interlocutors to keep track of which celebrity is to be guessed; small numbers that were also written on the cards uniquely identified the paper cards with the solution. Participants take turns in asking the questions and may continue as long as the

<sup>35</sup> Lori Foster Thompson and Michael D. Covert. “Teamwork Online: The Effects of Computer Conferencing on Perceived Confusion, Satisfaction, and Postdiscussion Accuracy”. In: *Group Dynamics: Theory, Research, and Practice* 7.2 (2003), pp. 135–151

<sup>36</sup> Laetitia Gros and Gontran Filandre. *A study on tasks for assessing audiovisual quality of videoconferencing systems in multipoint conversation tests*. ITU-T Contribution COM 12 - C 259 - E. Geneva, Switzerland: International Telecommunication Union, Oct. 2011

<sup>37</sup> ITU-T. *Recommendation P.1301 - Subjective quality evaluation of audio and audiovisual telemeetings*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2012

<sup>38</sup> Laetitia Gros and Gontran Filandre. *A study on tasks for assessing audiovisual quality of videoconferencing systems in multipoint conversation tests*. ITU-T Contribution COM 12 - C 259 - E. Geneva, Switzerland: International Telecommunication Union, Oct. 2011

<sup>39</sup> Janto Skowronek and Maxim Spur. *3SurvivalCT - 3-party Survival task Conversation Test scenarios for conferencing assessment*. 2017. DOI: 10.5281/zenodo.345837

<sup>40</sup> Katrin Schoenenberg, Alexander Raake, and Pierre Lebreton. “On interaction behaviour in telephone conversations under transmission delay”. In: *Proceedings of the 6th International Workshop on Quality of Multimedia Experience QoMEX*. Sept. 2014, pp. 31–36

answers are “yes”, while both interlocutors were allowed to answer the questions. In case of a “no” answer, the next interlocutor has the turn in asking questions. In case the name was correctly guessed, the participant could choose the next card with initials, the others looked for the corresponding solutions on the hidden cards, and the game continued with the next participant. This was continued until each call reached a duration of four to five minutes, then the test supervisor interrupted the game round and stopped the call. The German instructions (game rules) and the full list of celebrities for the German cultural context are electronically available<sup>41</sup>.

Test participants conducted the two types of conversation scenarios in interleaved order, i.e. changing the scenario type after each test call. The motivation was that the above mentioned small pilot test with six participant groups suggested negative effects on the participants' mood if seven calls of the same scenario type were made in one block. Having seven survival tasks games in sequence started to get boring and partially even depressing; having seven celebrity name guessing rounds in sequence started to get frustrating if test participants were not particularly good in guessing the names. In contrast, the interleaved order of the two scenario types was experienced quite positively by the participants.

*Confederate as test participant* Due to the good practical experience in Experiment ACT<sub>3</sub>, a confederate was also present in the two audiovisual experiments AVCT<sub>1</sub> and AVCT<sub>2</sub>. The confederate's task was to guide through the games and to aim at an equal amount of contribution of all interlocutors. In particular the confederate solved misunderstandings and minimized non-intended discourses. To optimize the naturalness and balance of the conversations, especially of the name guessing game, also the confederate did not know the full names of the celebrities and always got new celebrity names to guess. The test participants were informed about the confederate with the information that he was there to help play the games.

*Test System* The audio part of the test system was the same as in Experiment ACT<sub>3</sub>: The Skype Testing Client allowed to use different codecs (SILK, G.711) and audio bandwidths (narrowband, wideband, & super-wideband), test participants used closed headsets (Beyerdynamic DT790) as audio end devices, which were connected to the computer via RME sound cards, PureData was inserted in the audiopath for introducing acoustic distortions, and the oneway end-to-end audio delay was 350ms. The sound reproduction levels were calibrated with an head and torso simulator (HATS) to 73dB SPL, the microphone signals were adjusted to -26dBov digital level for an acoustic sound pressure level of 89dB SPL at the HATS mouth reference point. Concerning the video part of the test system, the Skype Testing Client applied the H264 video codec, allowed to set different data bitrates, video resolutions and frame rates. The oneway end-to-end video delay was  $200 \pm 2$  ms, which means that the system's

<sup>41</sup> Janto Skowronek and Katrin Schoenenberg. 3CNG - 3-party Celebrity Name Guessing task for Conferencing Assessment. 2017. DOI: 10.5281/zenodo.345834

audiovisual asynchrony was 150ms with lagging audio. Depending on different studies (see for instance Vatakis and Spence<sup>42</sup> or Eg et al.<sup>43</sup> for corresponding overviews), this value is below or just above the observed detection thresholds for audiovisual asynchrony. Logitech Pro9000 Webcams served as cameras, participants saw the video directly on the laptop screens (Fujitsu Siemens S Series) at a viewing distance of 3.4 times the screen height<sup>44</sup>, and the display peak luminance was set to the maximum, which was about  $100 \pm 10 \text{ cd/m}^2$ , using the white area of the test picture in Figure 8.10. The room lighting (D65, 6504 K) was set such that 34 lux of incident light were measured at face position and  $24 \text{ cd/m}^2$  of reflected light behind the screens. All light calibrations were done with a Sekonic L-758 Cine light meter. Finally, Netem was again used for introducing packet loss.

*Technical Conditions* As for the audio-only tests, the technical conditions are determined by the two aspects *combination of impairments on the individual connections* and the *technical system characteristics* as such. Table 8.10 provides a respective overview of all test conditions.

Motivating the selection of the conditions, the common goal was – as for the audio-only tests – to optimally balance the number of technical characteristics and observations per condition and the test effort. The motivations for certain choices though differed between the two experiments AVCT<sub>1</sub> to AVCT<sub>2</sub>. In Experiment AVCT<sub>1</sub>, the decision was to test a representative but small set of technical conditions, given that the experiment’s primary goal was to gain first experience in conducting a multiparty-relevant audiovisual test. The choice fell on testing two audio-only and two video-only impairments as well as their combinations to form respective audiovisual impairments, all in an asymmetric configuration (only one interlocutor is impaired). In Experiment AVCT<sub>2</sub>, the decision was to test again conditions that are constructed from two audio-only and two video-only impairments. This time, however, those impairments were tested in both an asymmetric (only one interlocutor is impaired) and symmetric configuration (both interlocutors are impaired), as well as their combination to an audiovisual impairment in an asymmetric configuration. Furthermore, bearing the results of the first experiment in mind, the selection of conditions specifically aimed at impairments that are clearly perceived and that show a rather good balance between the strengths of audio-only and video-only impairments. This was achieved by running a small pilot test with four subject groups. In addition, echo and audiovisual packet loss were selected due to their practical relevance from the perspective of the project under which the tests were conducted.

### 8.5.3 Method

*Design* Similarly to Experiment ACT<sub>3</sub>, both audiovisual experiments used a mixed design, in which a number of technical conditions were

<sup>42</sup> Argiro Vatakis and Charles Spence. “Audiovisual synchrony perception for speech and music assessed using a temporal order judgment task”. In: *Neuroscience Letters* 393 (2006), pp. 40–44

<sup>43</sup> Ragnhild Eg et al. “Audiovisual robustness: exploring perceptual tolerance to asynchrony and quality distortion”. In: *Multimedia Tools and Applications* 74 (2015), pp. 345–365

<sup>44</sup> Since participants could have different body positions, the viewing distance here was defined as the distance between the front border of the table and the laptop screen

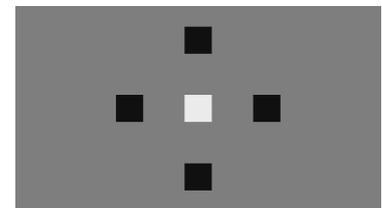


Figure 8.10: Test picture used for display calibration, provided by Garcia:

Marie-Neige Garcia. *Parametric Packet-based Audiovisual Quality model for IPTV services*. Springer, 2014

Experiment AVCT1: Audiovisual Conversation Test, Headphones (diotic), Superwideband codec						
Condition	Impairment combination			Impairment description	Technical parameters	#Obs
<i>AVRef</i>	0	0	0	0: no impairments	Reference: $SILK_{SWB}$ , $T_a = 350ms$ , $OLR = 10dB$ , $T_v = 200ms$ , $H264@500kbit/s$ , $FR = 30fps$ , $RES = 640x480$	41
<i>VF1</i>	0	0	<i>VF1</i>	<i>VF1</i> : reduced frame rate	$FR = 15fps$	66
<i>VR</i>	0	0	<i>VR</i>	<i>VR</i> : reduced resolution	$RES = 160x120$	67
<i>FM2</i>	0	0	<i>FM2</i>	<i>FM2</i> : bandpass filtered speech	microphone bandpass 400 – 7000Hz	63
<i>E2</i>	<i>TE2</i>	0	<i>LE2</i>	<i>TE2</i> : talker echo, <i>LE2</i> : listener echo	$TELR = 20dB$ , $WEPL' = 20dB$	65
<i>FM2 – VF1</i>	0	0	<i>FM2 – VF1</i>	<i>FM2</i> : bandpass filtered speech, <i>VF1</i> : reduced frame rate	microphone bandpass 400 – 7000Hz, $FR = 15fps$	18
<i>E2 – VF1</i>	<i>TE2</i>	0	<i>LE2 – VF1</i>	<i>TE2</i> : talker echo, <i>LE2</i> : listener echo <i>VF1</i> : reduced frame rate	$TELR = 20dB$ , $WEPL' = 20dB$ , $FR = 15fps$	23
<i>FM2 – VR</i>	0	0	<i>FM2 – VR</i>	<i>FM2</i> : bandpass filtered speech, <i>VF1</i> : reduced resolution	microphone bandpass 400 – 7000Hz, $RES = 160x120$	23
<i>E2 – VR</i>	<i>TE2</i>	0	<i>LE2 – VR</i>	<i>TE2</i> : talker echo, <i>LE2</i> : listener echo <i>VF1</i> : reduced resolution	$TELR = 20dB$ , $WEPL' = 20dB$ , $RES = 160x120$	22

Experiment AVCT2: Audiovisual Conversation Test, Headphones (diotic), Superwideband codec						
Condition	Impairment combination			Impairment description	Technical parameters	#Obs
<i>AVRef</i>	0	0	0	0: no impairments	$SILK_{SWB}$ , $T_a = 350ms$ , $OLR = 10dB$ , $T_v = 200ms$ , $H264 @ 500kbit/s$ , $FR = 30fps$ , $RES = 640x480$	41
<i>VF2</i>	0	0	<i>VF2</i>	<i>VF2</i> : reduced frame rate	$FR = 2fps$	43
<i>VF2sym</i>	0	<i>VF2</i>	<i>VF2</i>	<i>VF2</i> : reduced frame rate	$FR = 2fps$	34
<i>VB</i>	0	0	<i>VB</i>	<i>VB</i> : reduced bit rate	$H264@10kbit/s$ , $FR = 15fps$ , $RES = 160x120$	45
<i>VBsym</i>	0	<i>VB</i>	<i>VB</i>	<i>VB</i> : reduced bit rate	$H264@10kbit/s$ , $FR = 15fps$ , $RES = 160x120$	35
<i>E2</i>	<i>TE2</i>	0	<i>LE2</i>	<i>TE2</i> : talker echo, <i>LE2</i> : listener echo	$TELR = 20dB$ , $WEPL' = 20dB$	45
<i>FM4</i>	0	0	<i>FM4</i>	<i>FM4</i> : bandpass filtered speech	microphone bandpass 800 – 5000Hz	50
<i>FM4sym</i>	0	<i>FM4</i>	<i>FM4</i>	<i>FM4</i> : bandpass filtered speech	microphone bandpass 800 – 5000Hz	31
<i>FM4P3</i>	0	0	<i>FM4P3</i>	<i>FM4</i> : bandpass filtered speech, <i>P3</i> : audio packet loss	microphone bandpass 800 – 5000Hz, $P_{pl} = 15\% \mu_{10} = 1$ (=random PL)	45
<i>FM4P3sym</i>	0	<i>FM4P3</i>	<i>FM4P3</i>	<i>FM4</i> : bandpass filtered speech, <i>P3</i> : audio packet loss	microphone bandpass 800 – 5000Hz, $P_{pl} = 15\% \mu_{10} = 1$ (=random PL)	36
<i>FM4 – AVP3</i>	0	0	<i>FM4 – AVP3</i>	<i>FM4</i> : bandpass filtered speech, <i>AVP3</i> : audiovisual packet loss	microphone bandpass 800 – 5000Hz, $P_{pl} = 15\% \mu_{10} = 1$ (=random PL)	35
<i>FM4 – VF2</i>	0	0	<i>FM4 – VF2</i>	<i>FM4</i> : bandpass filtered speech, <i>VF2</i> : reduced frame rate	microphone bandpass 800 – 5000Hz, $FR = 2fps$	19
<i>FM4P3 – VF2</i>	0	0	<i>FM4P3 – VF2</i>	<i>FM4</i> : bandpass filtered speech, <i>P3</i> : audio packet loss <i>VF2</i> : reduced frame rate	microphone bandpass 800 – 5000Hz, $P_{pl} = 15\% \mu_{10} = 1$ (=random PL), $FR = 2fps$	17
<i>FM4 – VB</i>	0	0	<i>FM4 – VB</i>	<i>FM4</i> : bandpass filtered speech, <i>VB</i> : reduced bit rate	microphone bandpass 800 – 5000Hz, $H264@10kbit/s$ , $FR = 15fps$ , $RES = 160x120$	24
<i>FM4P3 – VB</i>	0	0	<i>FM4P3 – VB</i>	<i>FM4</i> : bandpass filtered speech, <i>P3</i> : audio packet loss <i>VB</i> : reduced bit rate	microphone bandpass 800 – 5000Hz, $P_{pl} = 15\% \mu_{10} = 1$ (=random PL), $H264@10kbit/s$ , $FR = 15fps$ , $RES = 160x120$	23

judged by all test participants (within-subject design) and other conditions were judged only by subsets of the test participants (between-subject design). This allowed to balance the overall number of conditions and the overlap of conditions between test participants (see corresponding discussion in Section 8.3). In AVCT1, the conditions

Table 8.10: Test conditions for the two audio-visual conversation tests AVCT1 and AVCT2. The conditions are defined by the shown combinations of the impairments that each interlocutor perceives from the three individual connections.

for the within-subject part were the reference condition as well as the two audio-only and two video-only impairments; the conditions for the between-subject part were the audiovisual conditions created by combining each audio-only and video-only impairment. In AVCT<sub>2</sub>, the conditions for the within-subject part were the reference, the echo and the audiovisual packet loss condition, as well as the two audio-only and two video-only impairments in asymmetric configurations; the conditions for the between-subject part were the audio-only and video-only impairments in symmetric configurations and the audiovisual conditions.

Concerning the number of conditions per test session, the good experience with 90 minutes test time from the audio-only tests and the conversation scenarios (four to five minutes per survival task and name guessing game) allowed for twelve test calls (six per task) and two training calls (one per task). Concerning the overall test time, the time and budget constraints allowed to aim for 24 groups (maximum 48 observations per condition due to two test participants and one confederate) in both experiments, from which 19 were actually conducted in Experiment AVCT<sub>1</sub> and all 24 in Experiment AVCT<sub>2</sub>.

*Test participants* The characteristics of the invited test participants are given in Table 8.11.

Participant characteristics	AVCT <sub>1</sub>	AVCT <sub>2</sub>
Number of participants (without confederate and any removed subjects)	38	49
Age: mean / minimum / maximum	30.8 / 22 / 58	26.7 / 19 / 54
Gender: female / male, in percent (absolute numbers)	63% / 37% (24 / 14)	53% / 47% (26 / 23)
Number female / male / mixed gender groups (not considering male confederate subject)	9 / 4 / 6	6 / 4 / 16
Multiparty experience in professional life (working, studying)	32% (12 of 38)	20% (10 of 49)
Multiparty experience in private life	13% (5 of 38)	39% (19 of 49)
Experience with video telephony	48% (17 of 38)	76% (37 of 49)
Groups with subjects knowing each other	2	2
Groups with subjects knowing the student co-worker	1	2
Number of subjects with experience with similar listening or telephony experiments	61% (23 of 38)	51% (25 of 49)
Hearing impairments (self-reported)	none	none
Visual impairments (self-reported)	none	none

Table 8.11: Test participant statistics for the two audiovisual conversation tests AVCT<sub>1</sub> & AVCT<sub>2</sub>.

*Procedure* In both experiments, participants came for a single 90 minutes session. After a general introduction, the test participants conducted first a “Get to know” call under reference condition, and everybody (confederate and test participants) introduced themselves to the others according to a number of questions asked by the confederate. After the “Get to know” call, the test participants had two training calls, in which the participants could get used to playing the two games and after which they answered a trial questionnaire. In both experiments, those training calls were conducted under the limited resolution condition *RES* and the echo condition *E2* (see Table 8.10). After the training calls, the test participants conducted

Experiments AVCT <sub>1</sub> & AVCT <sub>2</sub>		
Part	Question ID	Wording / Description, EN: English translation, DE: German original
1	Q1.1	EN: "What was your personal intuitive <b>overall impression (audio and video)</b> of the system's quality? – <i>extremely bad ... ideal</i> "
		DE: "Wie war Ihr persönlicher intuitiver <b>Gesamteindruck (Bild und Ton)</b> von der Qualität des Systems? – <i>schlecht ... ausgezeichnet</i> "
2	Q2.1	EN: "Your <b>own audiovisual connection</b> into the conference was <i>bad ... excellent</i> ."
		DE: "Ihre <b>eigene audiovisuelle Verbindung</b> in die Konferenzschaltung war <i>schlecht ... ausgezeichnet</i> ."
	Q2.2	EN: "The <b>audiovisual connection of Mr./Ms. Client X</b> was <i>bad ... excellent</i> ."
		DE: "Die <b>audiovisuelle Verbindung von Herrn/Frau Client X</b> in die Konferenzschaltung war <i>schlecht ... ausgezeichnet</i> ."
	Q2.3	EN: "The <b>audiovisual connection of Mr./Ms. Client Y</b> was <i>bad ... excellent</i> ."
		DE: "Die <b>audiovisuelle Verbindung von Herrn/Frau Client Y</b> in die Konferenzschaltung war <i>schlecht ... ausgezeichnet</i> ."
3	Q3.1	EN: "Please describe, if and which degradations or special characteristics you perceived. If you have other remarks, please write them down here as well. – <i>Free text</i> "
		DE: "Bitte beschreiben Sie kurz, ob und welche Störungen oder Besonderheiten Ihnen aufgefallen sind. Wenn Sie sonstige Anmerkungen haben, schreiben Sie diese bitte auch hier auf."

Table 8.12: Summary of the questionnaires used in the two audiovisual conversation tests AVCT<sub>1</sub> & AVCT<sub>2</sub>.

twelve test calls under different technical conditions according to the conversation scenarios (the two games) and answered a questionnaire after each call. Contrary to the audio-only conversation tests, no break was necessary, since the two games were cognitively much less demanding than the audio-only conversation scenarios. At the end of the session, the test supervisor conducted a short outtake interview with the test participants.

#### 8.5.4 Data Acquisition

*Questionnaires and Rating Scales* The questionnaire was with minor modifications essentially the same questionnaire used in Experiment ACT<sub>3</sub>. One modification emphasized that test participants should rate audiovisual quality, i.e. "audio and video", another modification addressed the interlocutors by the artificial account names of each test client that were visible on the screen.

Thus both tests measured *Speech Communication Quality with focus on the Telecommunication Component* for both *Single-Perspective Telemeeting Quality*  $Q_i$  and *Individual Connection Quality*  $Q_{ij}$ .

Table 8.12 provides English translations of the questions as well as the original German texts.

Table 8.13 gives an overview of the link between the *Measured Variables*, the actual measures used, and descriptions on how they were extracted from the questionnaires. For detailed explanation of these links see Section 8.3.

Furthermore, both tests also applied the two different rating scales: Absolute Category Rating (ACR) scale for  $Q_i$  and Extended Continuous (EC) scale for  $Q_{ij}$ , which required a scale transformation from the EC scale to the ACR scale according to Equation 8.1.

Experiments AVCT1 & AVCT2		
Measured Variable	Measure	Extraction from questionnaire
Individual Connection Quality		
<i>Speech Communication Quality</i> with focus on <i>Telecommunication Component</i>	<i>Connection Quality</i>	Directly measured with Q2.1, Q2.2, & Q2.3
Single-Perspective Telemeeting Quality		
<i>Speech Communication Quality</i> with focus on <i>Group-Communication Component</i>	<i>System's Quality</i>	Directly measured with Q1.1

*Data Processing and Validation* All measures were represented in such a way that high values reflect high quality. Furthermore, inspecting the questions did not reveal any specific outliers in terms of unusual rating behavior of test participants, except in one case in Experiment AVCT2, in which one participant was constantly rating all conditions and all questions as good. While this participant was removed from the data, an additional session could be scheduled, meaning that the data set actually contained 24 full groups plus one single test participant. Concerning the removal of data points in case a test participant forgot to answer all questions – an effect that happened in <5% in the audio-only conversation tests – this was only necessary in 2% of the cases for AVCT1 and in 1% of the cases for AVCT2.

Concerning the removal of data due to technical reasons, i.e. the high complexity of the system and the required amount of manual operations from the test supervisor between calls (see Section 8.3 for details), about 3% of the calls needed to be removed from the data. In addition, some of the data for two participants in Experiment AVCT1 and one participant in Experiment AVCT2 needed to be removed as well: Those participants reported in the open text field that they saw a bad quality of the own video picture, even though the own camera picture was displayed without any induced impairments. While it was not possible to reproduce a technical error in the system to verify that this perception stems from a technical error, the data in which the participants reported such an impairment was removed nevertheless.

*Impact of Conversation Scenarios* Since test participants conducted two different types of conversation scenarios, it is possible that the different types of content and the different conversational structure might have a significant influence on the quality ratings. To check for this, the author conducted per technical condition independent T-tests between the observations of the two scenario types.

For Experiment AVCT1 no significant impact of the conversation scenario was found; see Table G.1 in Appendix G for details. For Experiment AVCT2 a significant impact of the conversation scenarios was found for four out of 55 tested conditions, i.e. 7%; see Table G.2 for details. In two cases (conditions *FM4* and *FM4P3sym*), the quality of the own connection  $Q_{ii}$ , which was technically not impaired, was affected. In the other two cases (conditions *VB* and *FM4P3 – VF2*), the quality of the unimpaired connection to an interlocutor was affected. In all four cases the name guessing game showed significantly lower ratings than the survival task.

Table 8.13: Relation between the Measured Variables (on their conceptual level), the actual measures used for obtaining respective quantitative data, and extraction of the measures from the questionnaire answers.

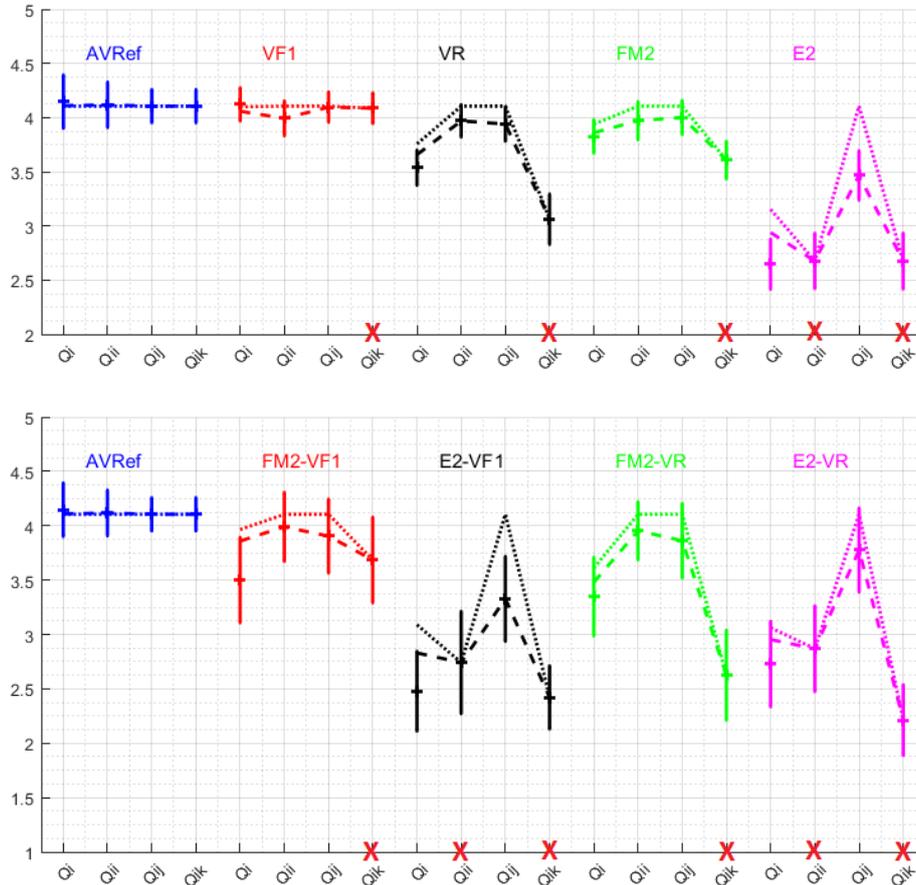
At first glance, this suggests that the name guessing game allows test participants to be more critical than the survival task. However, the impact was found actually for unimpaired connections and only in very few cases. Such unsystematic occurrence suggests that any such impact of the conversation scenarios is – at least for the present data – practically negligible. For that reason, all further analysis is conducted with the data of both conversation scenario types combined.

*Data Analysis and Presentation* The data analysis comprised the same four steps conducted for the audio-only conversation tests: (a) organize the collected data according to the algorithm described in Section 6.5; (b) determine the expected relations between the Single-Perspective Telemeeting Quality score  $Q_i$  and the Individual Connection Quality scores  $Q_{ii}$ ,  $Q_{ij}$  and  $Q_{ik}$ ; (c) apply one-sample t-tests to check for significant deviations between expected and observed values; and (d) verify per condition the hypotheses H1 and H2. For details see Section 8.3, and in particular Figure 8.1 on the expected relations between  $Q_i$ ,  $Q_{ii}$ ,  $Q_{ij}$  and  $Q_{ik}$ .

8.5.5 Results of Experiment AVCT<sub>1</sub>

Figure 8.11 shows the visualization of the results and provides a summary of the analysis; the analysis in full detail is given in Appendix G.1.

1. Visualization of Results and Expectations



The errorbars show the mean values and 95% confidence intervals for the obtained ratings of  $Q_i$ ,  $Q_{ii}$ ,  $Q_{ij}$ , and  $Q_{ik}$ . The red crosses at the x-axis indicate which individual connections are impaired. The dashed lines reflect the expectations when Hypothesis 1 (mutual influence) holds, the dotted lines when Hypothesis 1 does not hold. Both dashed and dotted lines assume that Hypothesis 2 (mean function) holds. The reference condition AVRef is repeated in each plot for better visual comparison.

Brief explanation of shown condition names (for full details see Table 8.10):

AVRef = reference condition. VF1 = reduced video frame rate. VR = reduced video resolution. FM2 = bandpass filtered microphone signal. E2 = echo. FM2-VF1, E2-VF1, FM2-VR, E2-VR = audiovisual impairments, combinations of VF1, VR, FM2, E2.

2. Analysis Summary

Result for reference condition:

$Q_i$ ,  $Q_{ii}$ ,  $Q_{ij}$ ,  $Q_{ik}$  are all essentially equal and optimal.  $\Rightarrow$  AVRef serves as reference condition for this test.

Results concerning Hypothesis H<sub>1</sub>, see dotted lines.

Condition	Result	Rationale
VF1, FM2, FM2-VF1, FM2-VR	×	$Q_{ii}$ & $Q_{ij}$ are not significantly different from the expected value.
E2, E2-VF1, VR	✓	$Q_{ij}$ is significantly different from the expected value.
E2-VR	×	$Q_{ij}$ is not significantly different from the expected value.

Results concerning Hypothesis H<sub>2</sub>, see dashed lines.

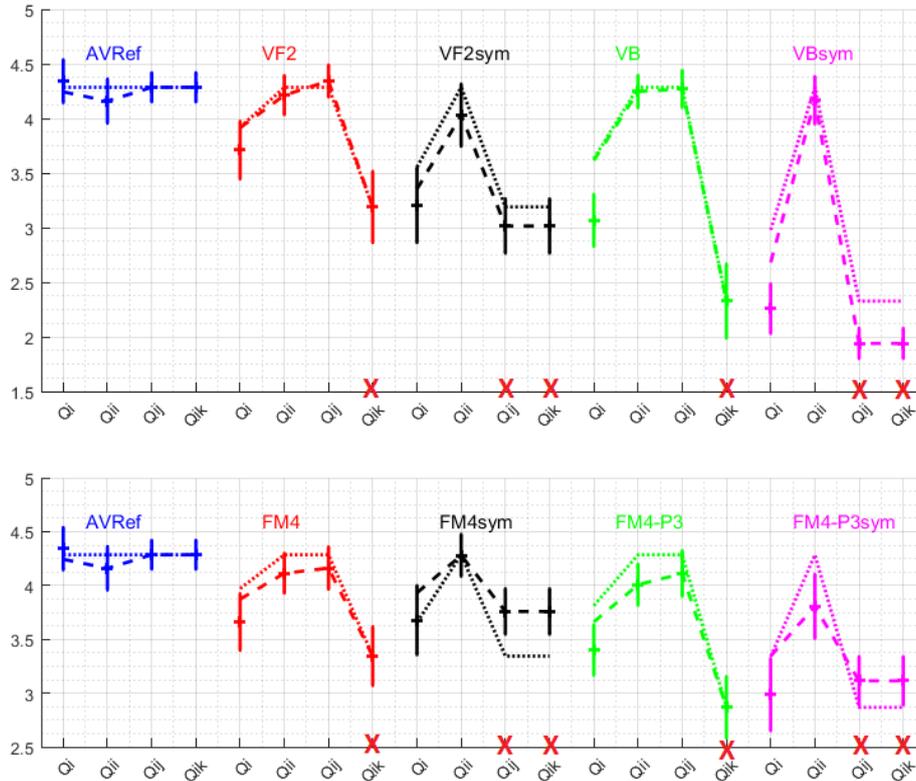
Condition	Result	Rationale
VF1, VR, FM2, E2-VF1, FM2-VR, E2-VR	✓	$Q_i$ is not significantly different from the expected value.
E2, FM2-VF1	×	$Q_i$ is significantly different from the expected value.

Figure 8.11: Results for Experiment AVCT<sub>1</sub>.

8.5.6 Results of Experiment AVCT2

Figures 8.12 and 8.13 show the visualization of the results and provide summaries of the analysis; the analysis in full detail is given in Appendix G.2.

1. Visualization of Results and Expectations



The errorbars show the mean values and 95% confidence intervals for the obtained ratings of  $Q_i$ ,  $Q_{ii}$ ,  $Q_{ij}$ , and  $Q_{ik}$ . The red crosses at the x-axis indicate which individual connections are impaired. The dashed lines reflect the expectations when Hypothesis 1 (mutual influence) holds, the dotted lines when Hypothesis 2 (mean function) holds. Both dashed and dotted lines assume that Hypothesis 2 (mean function) holds. The reference condition AVRef is repeated in each plot for better visual comparison.

Brief explanation of shown condition names (for full details see Table 8.10):

AVRef = reference condition. VF2, VF2sym = reduced video frame rate (“sym” = symmetric). VB, VBsym = reduced video bitrate (“sym” = symmetric). FM4, FM4sym = bandpass filtered microphone signal (“sym” = symmetric). FM4P3, FM4P3sym = bandpass filtered microphone signal and packet loss on audio signal (“sym” = symmetric).

2. Analysis Summary

Result for reference condition:

$Q_i$ ,  $Q_{ii}$ ,  $Q_{ij}$ ,  $Q_{ik}$  are all essentially equal and optimal.  $\Rightarrow$  AVRef serves as reference condition for this test.

Results concerning Hypothesis H1, see dotted lines.

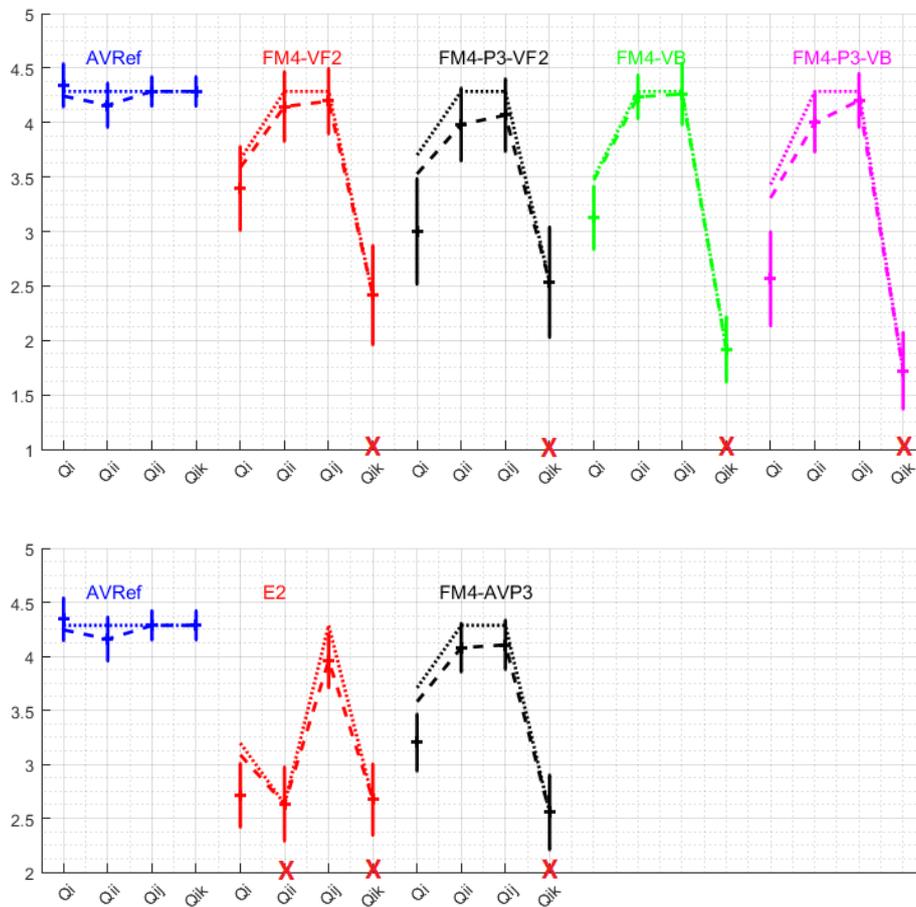
Condition	Result	Rationale
VF2, VF2sym, VB, FM4	×	$Q_{ii}$ & $Q_{ij}$ are not significantly different from the expected value.
FM4-P3	✓	$Q_{ii}$ is significantly different from the expected value.
VBsym, FM4sym	✓	$Q_{ij}$ is significantly different from the expected value.
FM4-P3sym	✓	$Q_i$ & $Q_{ij}$ are significantly different from the expected value.

Results concerning Hypothesis H2, see dashed lines.

Condition	Result	Rationale
VF2, VF2sym, FM4, FM4sym	✓	$Q_i$ is not significantly different from the expected value.
VB, VBsym, FM4-P3, FM4-P3sym	×	$Q_i$ is significantly different from the expected value.

Figure 8.12: Results for Experiment AVCT2, Part 1.

1. Visualization of Results and Expectations



The errorbars show the mean values and 95% confidence intervals for the obtained ratings of  $Q_i$ ,  $Q_{ii}$ ,  $Q_{ij}$ , and  $Q_{ik}$ . The red crosses at the x-axis indicate which individual connections are impaired. The dashed lines reflect the expectations when Hypothesis 1 (mutual influence) holds, the dotted lines when Hypothesis 1 does not hold. Both dashed and dotted lines assume that Hypothesis 2 (mean function) holds. The reference condition AVRef is repeated in each plot for better visual comparison.

Brief explanation of shown condition names (for full details see Table 8.10):

AVRef = reference condition. FM4-VF2, FM4-P3-VF2, FM4-VB, FM4-P3-VB = audiovisual impairments, combinations of FM4, FM4-P3, VF2, VB. E2 = echo. FM4-AVP3 = bandpass filtered speech and packet loss on audio and video signal.

2. Analysis Summary

Result for reference condition:

$Q_i$ ,  $Q_{ii}$ ,  $Q_{ij}$ ,  $Q_{ik}$  are all essentially equal and optimal.  $\Rightarrow$  AVRef serves as reference condition for this test.

Results concerning Hypothesis H1, see dotted lines.

Condition	Result	Rationale
FM4-VF2, FM4-P3-VF2, FM4-VB, FM4-AVP3	×	$Q_{ii}$ & $Q_{ij}$ are not significantly different from the expected value.
FM4-P3-VB	✓	$Q_{ii}$ is significantly different from the expected value.
E2	✓	$Q_{ij}$ is significantly different from the expected value.

Results concerning Hypothesis H2, see dashed lines.

Condition	Result	Rationale
FM4-VF2	✓	$Q_i$ is not significantly different from the expected value.
FM4-P3-VF2, FM4-VB, FM4-P3-VB, E2, FM4-AVP3	×	$Q_i$ is significantly different from the expected value.

Figure 8.13: Results for Experiment AVCT2, Part 2.

## 8.5.7 Discussion

*Review of Results* Table 8.14 summarizes the main findings for the two audiovisual conversation tests AVCT<sub>1</sub> and AVCT<sub>2</sub>.

Experiment	Conditions					Hypothesis H1	Hypothesis H2
	Audio-only		Video-only		Audiovisual		
	Asymmetric	Symmetric	Asymmetric	Symmetric	Asymmetric		
AVCT <sub>1</sub>	FM <sub>2</sub>	–	VF <sub>1</sub>	–	FM <sub>2</sub> -VR, E <sub>2</sub> -VR	×	✓
	–	–	–	–	FM <sub>2</sub> -VF <sub>1</sub>	×	×
	–	–	VR	–	E <sub>2</sub> -VF <sub>1</sub>	✓	✓
	E <sub>2</sub>	–	–	–	–	✓	×
AVCT <sub>2</sub>	FM <sub>4</sub>	–	VF <sub>2</sub>	VF <sub>2</sub> sym	FM <sub>4</sub> -VF <sub>2</sub>	×	✓
	–	–	VB	–	FM <sub>4</sub> -P <sub>3</sub> -VF <sub>2</sub> , FM <sub>4</sub> -VB, FM <sub>4</sub> -AVP <sub>3</sub>	×	×
	E <sub>2</sub> , FM <sub>4</sub> -P <sub>3</sub>	FM <sub>4</sub> sym, FM <sub>4</sub> -P <sub>3</sub> sym	–	VBsym	FM <sub>4</sub> -P <sub>3</sub> -VB	✓	×

Hypothesis H1 is confirmed in nine out of 22 tested cases, and rejected in 13 cases. This means that a clear evidence is found for a mutual influence of the individual connections, while such influence depends on the actual technical condition. Concerning a systematic behavior that could explain when a condition shows that influence or not, it is helpful to first recapitulate the results of the audio-only tests. The audio-only conversation tests and listening-only tests showed that the impairment type most likely does not explain the influence, while the strength of an impairment – at least partially – does. This could be verified since in the audio-only tests, multiple instantiations of the same impairment type at different strengths were tested. Now looking at the audiovisual tests here, however, only few such multiple instantiations were tested. For that reason, there is some evidence that the impairment type explains the mutual influence, but this is contradictory, according to Table 8.14: Those cases with different instantiations of the same impairment type (i.e. VF<sub>1</sub> and VF<sub>2</sub>) show consistently no mutual influence, which does not reject the impairment type as a possible explanation. However, considering cases with similar impairment types (i.e. VR and VB, both essentially causing lower video resolution) show either a mutual influence or not, which rejects the impairment type as a possible explanation. Concerning the strength of an impairment as explanation for a mutual influence, again the evidence is contradictory: Figure 8.11 supports this for the audio-only and video-only impairments in Experiment AVCT<sub>1</sub> (FM<sub>2</sub> vs. E<sub>2</sub>, VF<sub>1</sub> vs. VR), but not for the audiovisual impairments; Figures 8.12 and 8.13 support this only for the asymmetric audio-only impairments (FM<sub>4</sub> vs. FM<sub>4</sub>-P<sub>3</sub>), but not for the other impairments. Furthermore, the selection of conditions was specifically intended to investigate modality effects in terms of audio-only and video-only impairments as well as symmetry effects in terms of differences between asymmetric and symmetric configurations. Thus, it is possible

Table 8.14: Overview of the results for the audiovisual conversation tests AVCT<sub>1</sub> & AVCT<sub>2</sub>, concerning the two hypotheses H1 & H2. Confirmed hypotheses are denoted with a ✓, rejected hypotheses with a ×.

to check, if modality or symmetry can explain the influence. However, neither modality nor symmetry do this, since Table 8.14 shows that impairments from all modalities and impairments in both symmetric and asymmetric configurations are distributed among both cases *H1 confirmed* and *H1 rejected*.

Hypothesis H2 is confirmed in ten of 22 tested cases, and rejected in twelve cases. This means that in many cases the Single-Perspective Telemeeting Quality  $Q_i$  can be expressed as a simple average of the Individual Connection Quality Scores  $Q_{ij}$  of the individual connections. However, there are also many technical conditions for which this simple relation does not hold. As for the mutual influence, neither the type nor the strength of an impairment are sufficient to explain the need to express the relation between  $Q_i$  and  $Q_{ij}$  with a more sophisticated function. Conditions that require a more sophisticated function (E2, FM2-VF1, VB, VBsym, FM4-P3, FM4-P3sym) are of different impairment types (Table 8.14) and have quality scores distributed across the whole range (Figures 8.11 to 8.13). Similarly, also no consistent trend in terms of symmetry is visible, as impairments in asymmetric and symmetric conditions do either both confirm or reject H2 (FM4-P3 vs. FM-P3sym, VF2 vs. VF2sym, VB vs. VBsym), or they are distributed across the two cases *H2 confirmed* and *H2 rejected* (FM4 vs. FM4sym). Also no consistent trend in terms of modality is visible, as impairments from all modalities are distributed among both cases *H2 confirmed* and *H2 rejected*.

*Review of the Approach and Open Questions for Future Work* There are a number of open questions concerning the experimental method that this section will address now: rating on multiple scales and the limitation in terms of communicative aspects.

The aspect of multiple rating scales has been discussed in Section 8.3 for the audio-only conversation tests and Section 8.4 for the listening-only tests, and the same discussion (in short) applies here: it might be that test participants do form some average of  $Q_{ij}$  simply due to the presentation of both  $Q_i$  and  $Q_{ij}$  on the same questionnaire, despite the usage of different scale designs and questions; and future studies could validate this by different experimental approaches (e.g. split sessions, between-subject designs).

Also the aspect of the communicative situation was discussed in Section 8.3 for the audio-only conversation tests, and a similar discussion applies here: the experiments considered one degree of communication complexity in terms of the number of interlocutors, and future work could repeat this experiments with a different number of interlocutors. However, contrary to the audio-only conversation tests, test participants had to perform two different conversation tasks throughout a session, whereas those tasks are different in their conversational structure: in one task participants have a discussion, in the other task participants exchange questions and short yes-no answers. This means that the experiments considered two degrees of communication complexity in terms of the conversational structure.

Interestingly, the impact of those two tasks on the quality ratings was found to be non-significant for the present data, despite the conversational differences. Here, future work could apply conversational analysis methods (see Section 2.2) and correlate those results with the quality ratings to better investigate the impact of conversation tasks and quality perception, similar to the work by Schoenenberg<sup>45</sup>, who did such analysis for delay impairments.

### Summary

The goal of the research presented here was to obtain more insights into the influence of telecommunication aspects on telemeeting quality, more specifically the perceived system performance, expressed as *Speech Communication Quality* with a focus on the *Telecommunication Component*. In particular, the quality aggregation process from *Individual Connection Quality* scores  $Q_{ij}$  to the *Single-Perspective Telemeeting Quality* score  $Q_i$  was in focus of the investigations. The communicative aspects were indirectly considered in the tests as well by simulating three different types of communication scenarios: an audio-only telemeeting from the perspective of an active participant, simulated by means of an audio-only conversation test, an audio-only telemeeting from the perspective of a passive participant, simulated by means of an listening-only test, and an audiovisual telemeeting from the perspective of an active participant, simulated by means of an audiovisual conversation test.

Concerning the audio-only conversation tests, three experiments were conducted in which the experimental methodology of the third experiment was improved after the first two experiments. Here the following results were found: First, there is evidence for a mutual influence of the individual connections on the *Individual Connection Quality* scores  $Q_{ij}$ , whereas this influence is visible only for a limited number of the tested conditions. Second, for most conditions, the average of *Individual Connection Quality* scores  $Q_{ij}$  is sufficient to estimate the *Single-Perspective Telemeeting Quality* score  $Q_i$ , whereas there also are a number of conditions for which a more sophisticated aggregation than a simple mean is necessary. Third, the strength of an impairment appears to explain – at least partially – whether the two effects (mutual influence of connections and need for a more sophisticated aggregation) occur or not: the stronger the impairment, the more likely the effects occur. Fourth, improving the experimental methodology, which also allowed to increase the number of tested conditions, paid off in the sense that only the third experiment was sufficiently sensitive to detect that a simple average of  $Q_{ij}$  is not always sufficient to estimate  $Q_i$ . The main finding from those audio-only conversation tests can be formulated as follows: the relation between the two aggregation levels is a rather complex construct, which appears to depend on the strength of impairments or the level of quality, respectively.

Concerning the listening-only conversation tests, two experiments

<sup>45</sup> Katrin Schoenenberg. "The Quality of Mediated-Conversations under Transmission Delay". PhD Thesis. Technische Universität Berlin, Germany, 2016

were conducted in which the experimental methodology was improved after the first experiment. Here the following results were found: First, as for the audio-only conversation tests, part of the tested conditions showed a mutual influence of individual connections and a need for a more sophisticated aggregation, while other conditions did not show those effects. Second, slightly different from the audio-only conversation tests, the strength of an impairment is not sufficient to explain when such effects occur, as results are less consistent in that respect. Third, the presence of asymmetry appears to partially explain those effects as well, whereas here again the results are not very consistent either. The main finding from those listening-only tests can be formulated as follows: the relation between the two aggregation levels is a rather complex construct, which appears to depend on the one hand on the strength of impairments or the level of quality, and on the other hand on the presence of asymmetry.

Concerning the audiovisual conversation tests, two experiments were conducted in which the experimental methodology was – as before – improved after the first experiment. Here the following results were found: First, as for the audio-only conversation tests, part of the tested conditions showed a mutual influence of individual connections and a need for a more sophisticated aggregation, while other conditions did not show those effects. Second, slightly different from the audio-only conversation tests, neither the strength of an impairment nor the presence of asymmetry are sufficient to explain when such effects occur, as results are less consistent in that respect. Third, the modality of an impairment appears to partially explain those effects as well, whereas here again the results are not very consistent either. The main finding from those audiovisual conversation tests can be formulated as follows: the relation between the two aggregation levels is a rather complex construct, which appears to depend – at least partially – on the strength of impairments, the presence of asymmetry, and the modality of impairments, whereas no individual aspect alone is sufficient to explain the aggregation consistently.

## Perceptual Models of Telemeeting Quality

---

This chapter contains text passages that either stem from previous texts of the author or that are based on such source texts. For readability purposes those passages are not marked as quotations. The main source documents used for this chapter are:

- Janto Skowronek et al. "Method and Apparatus for Computing the Perceived Quality of a Multiparty Audio or Audiovisual Telecommunication Service or System". Patent application WO/2016/041593 (EP). Deutsche Telekom AG. Mar. 2016
- Janto Skowronek. *Initial model for audio-only communication based on available data*. Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, June 2013
- Janto Skowronek. *Improved model for audio-only communication*. Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, Dec. 2013
- Janto Skowronek. *Pilot test for audiovisual communication*. Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, Feb. 2014
- Janto Skowronek. *Second test for audiovisual communication*. Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, Oct. 2014
- Janto Skowronek. *Final Project Report – Quality of Multiparty Audio-Video Communication*. Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, Dec. 2014

Since additional documents served as further background material, see Appendix A for the full list of source documents and their contributions to this chapter.

---

### *What this chapter is about*

This chapter reports on the development and evaluation of a set of algorithms that model the quality aggregation process that was experimentally investigated in the previous chapter. In other words, the goal is to estimate by means of mathematical algorithmic operations the *Single-Perspective Telemeeting Quality*  $Q_i$  based on the *Individual Connection Quality* scores  $Q_{ij}$ .

### 9.1 Introduction

*Motivation* The motivation for the intended models is to obtain insights on how the individual connections contribute to the overall telemeeting experience. From a theoretical perspective, this provides evidence for the quality aggregation processes discussed in the conceptual model in Chapter 5. From an engineering perspective, this provides the basis for a future instrumental modeling framework. In such a framework the quality of individual connections  $Q_{ij}$  can be estimated with instrumental methods and then the telemeeting quality  $Q_i$  is computed based on those estimations.

*Approach* The approach is to develop and evaluate a set of computational models, which use as input the  $Q_{ij}$  ratings obtained in the perception experiments of the previous Chapter 8, and which compute estimations of the  $Q_i$  ratings obtained from said experiments. Since these models rely on perceptual data as input, they are called here *Perceptual Quality Models*, in order to differentiate them from the so-called *Instrumental Quality Models*, in which technical information or the signals are used as input (Section 4.5).

*Hypothesis* The presented modeling work is build around a fundamental modeling hypothesis, which can be formulated as:

H1: There exists a set of advanced models that outperform the simple mean operation, whereas the added value of those advanced models holds only for a subset of conditions.

This hypothesis is motivated by the results of Chapter 8, which suggest that in many cases, i.e. technical conditions, it is sufficient to estimate the *Single-Perspective Telemeeting Quality*  $Q_i$  as a simple mean of the *Individual Connection Quality* scores  $Q_{ij}$ , while for other conditions this mean operation is clearly not sufficient. Thus, the present modeling work is focused on finding a set of appropriate advanced models and on evaluating their performance with respect to the performance of a simple mean operation and to the potential dependency on the technical conditions.

## 9.2 Modeling Algorithms

In this work, the author developed and evaluated different modeling algorithms. Those algorithms are first inspired by three fundamental modeling ideas, then translated into a set of eight modeling functions, i.e. mathematical operations to estimate  $Q_i$  scores based on  $Q_{ij}$  scores, and finally further specified by a set of additional modeling factors, which essentially resemble a number of processing variations.

### 9.2.1 Fundamental Modeling Ideas

The first two fundamental modeling ideas (MI1 & MI2) can be formulated as:

MI1: The Single-Perspective Telemeeting Quality  $Q_i$  is a simple mean of the Individual Connection Qualities  $Q_{ij}$ .

MI2: The Single-Perspective Telemeeting Quality  $Q_i$  is the minimum value of the Individual Connection Qualities  $Q_{ij}$ .

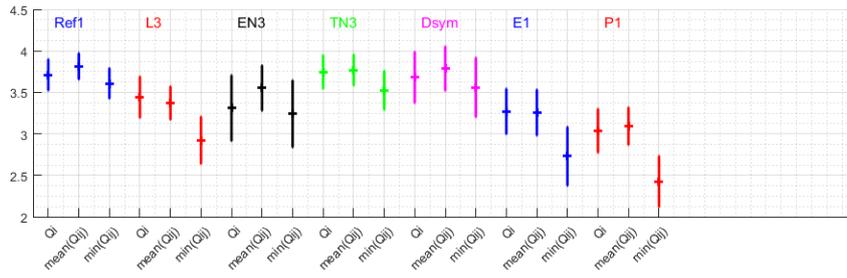
These ideas are motivated by the observation that for almost all conditions  $Q_i$  lies between the mean value and the minimum value of  $Q_{ij}$ . As Figures 9.1 to 9.3 show, in most of those conditions the mean of the  $Q_{ij}$  is closer to  $Q_i$ , but for some conditions the minimum value of  $Q_{ij}$  is closer to  $Q_i$ . In that respect, MI1 and MI2 can be regarded

as the baseline models due to their simplicity and the fact that they constitute for most conditions the upper and lower boundaries.

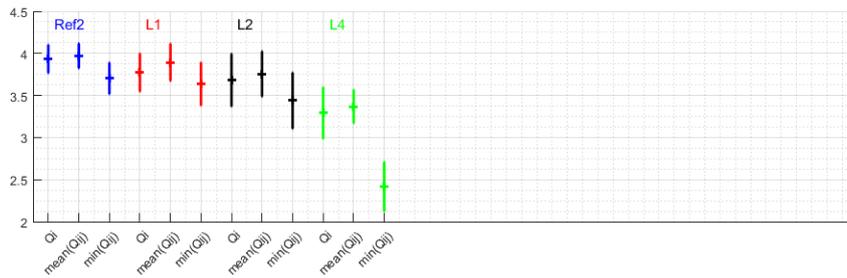
However, the noise conditions TN<sub>1</sub> and TN<sub>2</sub> in ACT<sub>3</sub> and LOT<sub>2</sub> and the echo conditions E<sub>2</sub> and E<sub>3</sub> in ACT<sub>3</sub> deviate from this; here  $Q_i$  is even lower than the minimum value of  $Q_{ij}$ . Looking more closely at the noise conditions in ACT<sub>3</sub> and LOT<sub>2</sub>, the individual connection causing the impairment can not be identified by the test participants as the noise is additive and does not provide any cues from which interlocutor it stems. This in turn suggests an inverse quality aggregation paradigm: instead of aggregating  $Q_{ij}$  to  $Q_i$ , participants form a judgment of  $Q_i$  and then *distribute* this judgment, or more precisely the quality impairment leading to this judgment, across the individual connections. Turning this relation around, a third modeling idea can be extracted as follows:

MI<sub>3</sub>: The *Single-Perspective Telemeeting Quality Impairment*  $Q_{I_i}$  is the sum of the *Individual Connection Quality Impairments*  $Q_{I_{ij}}$ , whereas  $Q_{I_i}$  can be expressed as the difference between a maximum possible quality  $Q_{max}$  and the *Single-Perspective Telemeeting Quality*  $Q_i$  and  $Q_{I_{ij}}$  as the difference between  $Q_{max}$  and the *Individual Connection Quality*  $Q_{ij}$ .

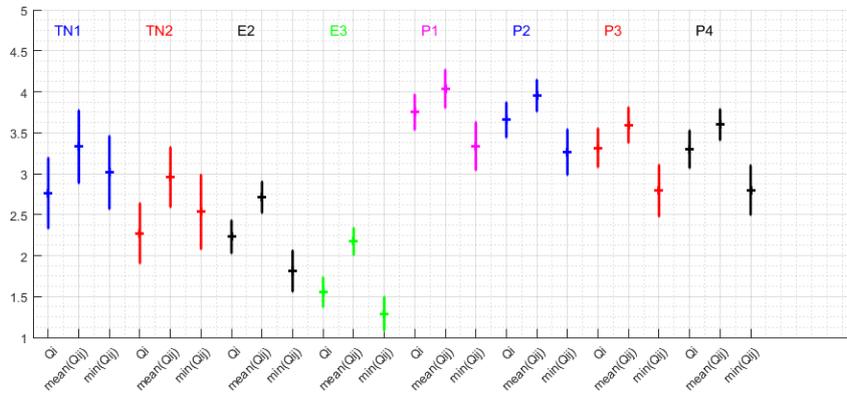
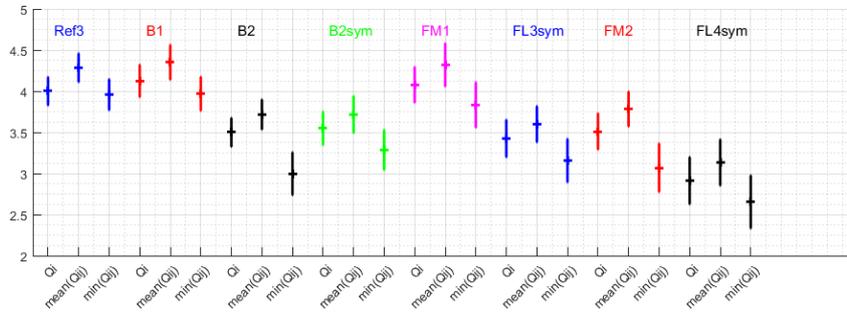
In other words, MI<sub>3</sub> assumes that not an aggregation of the actual quality values but an aggregation of quality impairments is taking place, which then lead to quality values.



(a) Experiment ACT1:  $mean(Q_{ij})$  is closer to  $Q_i$  than  $min(Q_{ij})$  for: L3 (loudness), TN3 (transmitted noise), Dsym (delay), E1 (echo), P1 (packet loss).  $min(Q_{ij})$  is closer to  $Q_i$  for: EN3 (noise in environment). Both are equally close for: Ref1 (reference).

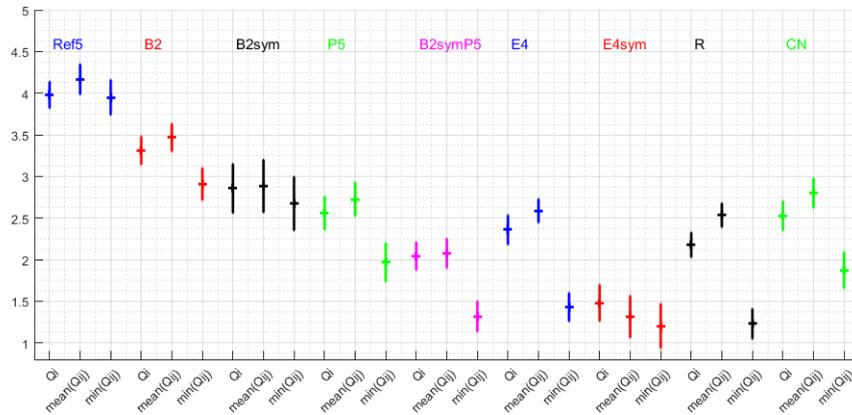


(b) Experiment ACT2:  $mean(Q_{ij})$  is closer to  $Q_i$  than  $min(Q_{ij})$  for: Ref2 (reference), L1, L2, L4 (loudness).  $min(Q_{ij})$  is never closer to  $Q_i$ .

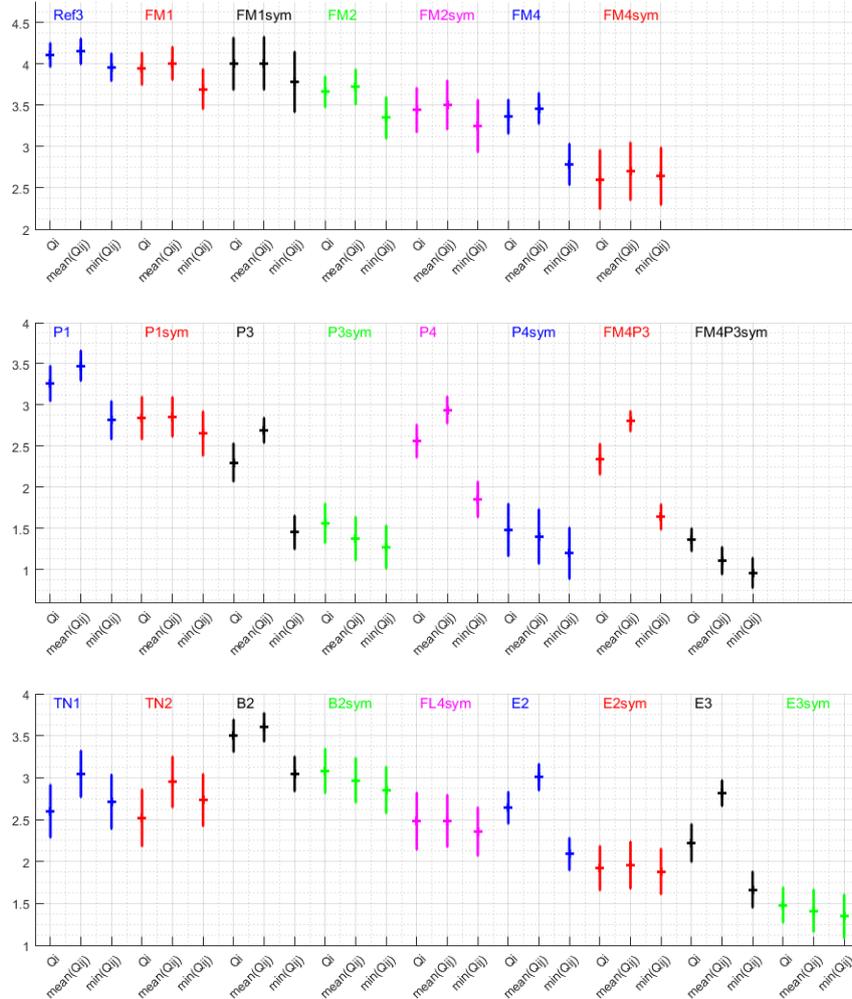


(c) Experiment ACT3:  $mean(Q_{ij})$  is closer to  $Q_i$  than  $min(Q_{ij})$  for: B2, B2sym (narrow band codec), FM1, FL3sym, FM2, FL4sym (bandpass filtered speech), P1, P2, P3, P4 (packet loss).  $min(Q_{ij})$  is closer to  $Q_i$  for: Ref3 (reference), B1 (wide band codec), TN1, TN2 (transmitted noise), E2, E3 (echo).

Figure 9.1: Motivation for the two fundamental modeling ideas MI1 ( $Q_i$  is mean of  $Q_{ij}$ ) and MI2 ( $Q_i$  is minimum value of  $Q_{ij}$ ), data of audio-only conversation tests.

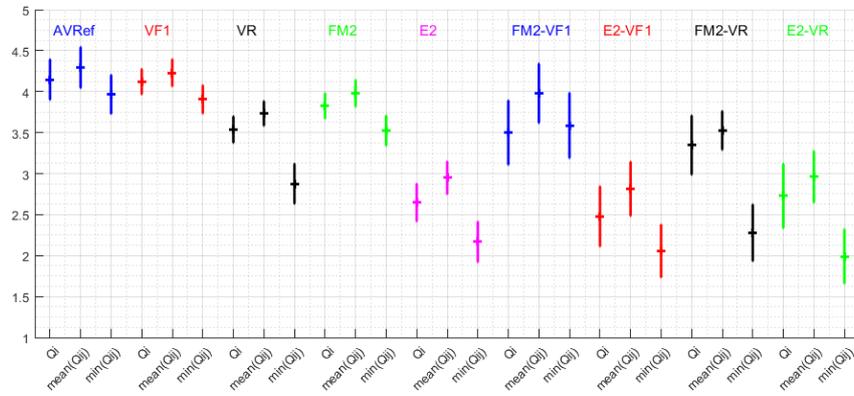


(d) Experiment LOT1:  $mean(Q_{ij})$  is closer to  $Q_i$  than  $min(Q_{ij})$  for: B2, B2sym (narrow band codec), P5 (packet loss), B2symP5 (narrow band codec & packet loss), E4, E4sym (echo), R (reverberation), CN (speech correlated noise).  $min(Q_{ij})$  is closer to  $Q_i$  for: Ref5 (reference).

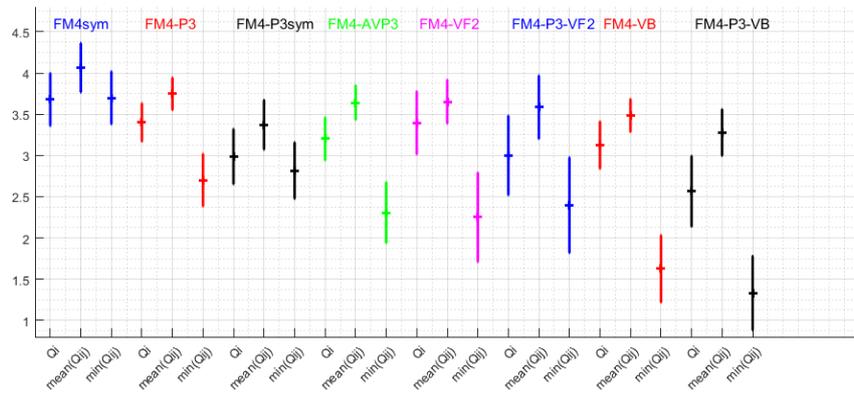
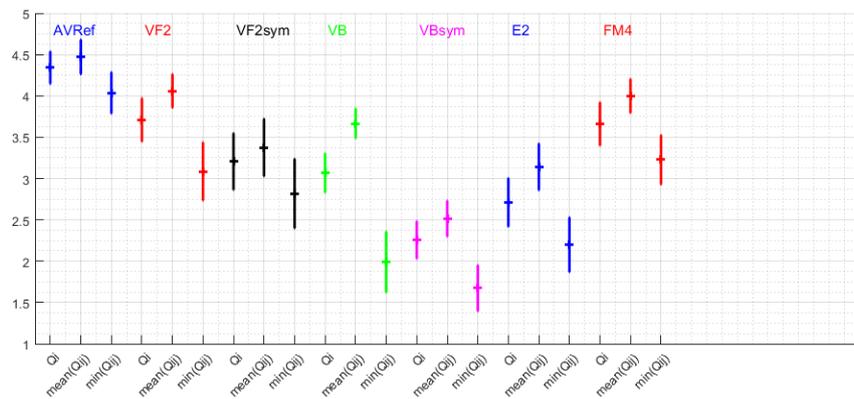


(e) Experiment LOT2:  $mean(Q_{ij})$  is closer to  $Q_i$  than  $min(Q_{ij})$  for: Ref3 (reference), FM1, FM1sym, FM2, FM2sym, FM4, FL4sym (bandpass filtered speech), P1, P1sym, P3, P3sym, P4, P4sym (packet loss), FM4P3, FM4P3sym (bandpass filtered speech & packet loss), B2, B2sym (narrow band codec), E2, E2sym, E3, E3sym (echo).  $min(Q_{ij})$  is closer to  $Q_i$  for: FM4sym (bandpass filtered speech), TN1, TN2 (transmitted noise).

Figure 9.2: Motivation for the two fundamental modeling ideas MI1 ( $Q_i$  is mean of  $Q_{ij}$ ) and MI2 ( $Q_i$  is minimum value of  $Q_{ij}$ ), data of listening-only tests.



(f) Experiment AVCT1:  $mean(Q_{ij})$  is closer to  $Q_i$  than  $min(Q_{ij})$  for: AVref (reference), VF1 (reduced frame rate), VR (reduced resolution), FM2 (bandpass filtered speech), E2 (echo), E2-VF1, FM2-VR, E2-VR (audiovisual combinations).  $min(Q_{ij})$  is closer to  $Q_i$  for: FM2-VF1 (bandpass filtered speech & reduced frame rate).



(g) Experiment AVCT2:  $mean(Q_{ij})$  is closer to  $Q_i$  than  $min(Q_{ij})$  for: AVref (reference), VF2, VF2sym (reduced video frame rate), VB, VBSym (reduced bit rate), E2 (echo), FM4 (bandpass filtered speech), FM4-P3 (bandpass filtered speech & packet loss on audio stream), FM4-VP3 (bandpass filtered speech & packet loss on audio & video stream), FM4-VF2 (bandpass filtered speech & reduced frame rate), FM4-P3-VF2 (bandpass filtered speech & packet loss on audio stream & reduced frame rate), FM4-VB (bandpass filtered speech & reduced bit rate), FM4-P3-VB (bandpass filtered speech & packet loss on audio stream & reduced bit rate).  $min(Q_{ij})$  is closer to  $Q_i$  for: FM4sym, FM4-P3sym.

Figure 9.3: Motivation for the two fundamental modeling ideas MI1 ( $Q_i$  is mean of  $Q_{ij}$ ) and MI2 ( $Q_i$  is minimum value of  $Q_{ij}$ ), data of audiovisual conversation tests.

### 9.2.2 Modeling Functions

Based on the fundamental modeling ideas and inspired by suggestions from the team members of the project under which the present work was conducted<sup>1</sup>, the author developed and tested eight different modeling functions F1 to F8. These functions estimate the *Single-Perspective Telemeeting Quality* score  $\widehat{Q}_i$  based on the *Individual Connection Quality* scores  $Q_{ij}$ . Table 9.1 shows an overview of these functions, while the following text will provide the corresponding rationale for selecting these functions.

<sup>1</sup> Janto Skowronek et al. "Method and Apparatus for Computing the Perceived Quality of a Multiparty Audio or Audiovisual Telecommunication Service or System". Patent application WO/2016/041593 (EP). Deutsche Telekom AG. Mar. 2016

Function	Computation	Parameter Constraints
F1 "baseline mean"	$\widehat{Q}_i = \text{mean}(Q_{ix}) = 1/3 \cdot (Q_{ii} + Q_{ij} + Q_{ik})$	
F2 "weighted mean"	$\widehat{Q}_i = w_{ii} \cdot Q_{ii} + w_{ij} \cdot Q_{ij} + w_{ik} \cdot Q_{ik}$	$w_{ii} + w_{ij} + w_{ik} = 1$
F3 "mutual influence"	$\widehat{Q}_i = \sigma_{ij}(Q_{ij}) \cdot \sigma_{ik}(Q_{ik}) \cdot Q_{ii}$ $+ \sigma_{ii}(Q_{ii}) \cdot \sigma_{ik}(Q_{ik}) \cdot Q_{ij}$ $+ \sigma_{ii}(Q_{ii}) \cdot \sigma_{ij}(Q_{ij}) \cdot Q_{ik}$ with $\sigma_{ix}(Q_{ix}) = \frac{1}{1 + e^{-s_{ix} \cdot (Q_{ix} - o_{ix})}}$ , $x \in [i, j, k]$	$s_{ix} \in [0..10]$ and $o_{ix} \in [0..10]$
F4 "quality dependent influence"	$\widehat{Q}_i = w_1 \cdot \text{mean}(Q_{ix}) \cdot \sigma_1(\text{mean}(Q_{ix}))$ $+ w_2 \cdot \min(Q_{ix}) \cdot \sigma_2(\text{mean}(Q_{ix}))$ with $\sigma_1(Q_{ix}) = \frac{1}{1 + e^{-s_1 \cdot (Q_{ix} - o_1)}}$ with $\sigma_2(Q_{ix}) = \frac{1}{1 + e^{+s_2 \cdot (Q_{ix} - o_2)}}$	$[w_1, w_2, s_1, s_2, o_1, o_2] > 0$
F5 "influence of outliers"	$\widehat{Q}_i = w_{ii,n} \cdot w_{ii,s} \cdot Q_{ii} + w_{ij,n} \cdot w_{ij,s} \cdot Q_{ij} + w_{ik,n} \cdot w_{ik,s} \cdot Q_{ik}$ with $w_{ix,n} = \begin{cases} \frac{1-\beta}{\#IL-1} \cdot \#outliers + \frac{\#IL-\beta}{\#IL-1}$ , if $Q_{ix}$ is outlier 1, if $Q_{ix}$ is not an outlier with $w_{ix,s} = \begin{cases} \frac{1-\delta}{\gamma} \cdot \text{std}(Q_{ix}) + \delta$ , if $Q_{ix}$ is outlier 1, if $Q_{ix}$ is not an outlier with $Q_{ix}$ is outlier if $Q_{ix} < \text{mean}(Q_{ix}) - \alpha \cdot \text{std}(Q_{ix})$ or $Q_{ix} > \text{mean}(Q_{ix}) + \alpha \cdot \text{std}(Q_{ix})$	$[\alpha, \beta, \gamma, \delta] > 0$
F6 "weighted minimum"	$\widehat{Q}_i = w \cdot \min(Q_{ix}) + C$	$w > 0, C \in \mathbb{R}$
F7 "baseline minimum"	$\widehat{Q}_i = \min(Q_{ix})$	
F8 "weighted mean of impairments"	$\widehat{Q}_i = w_{ii} \cdot Q_{ii} + w_{ij} \cdot Q_{ij} + w_{ik} \cdot Q_{ik} + w_{max} \cdot Q_{max}$ with $Q_{max} = \text{maximum quality score in whole data set}$	$[w_{ii}, w_{ij}, w_{ik}] > 0, w_{max} \leq 0$

Table 9.1: Modeling functions F1 to F7.

**Functions F1 and F7** The rationale for F1 and F7 is straight forward: they realize the first two fundamental modeling ideas *simple mean* of  $Q_{ij}$  and *simple minimum* value of  $Q_{ij}$ .

**Functions F2 and F6** The rationale for F2 and F6 is to extend F1 and F7 with parameters that can be chosen to improve prediction

performance. Thus, F2 is a weighted mean of  $Q_{ij}$  and F6 a weighted minimum value of  $Q_{ij}$  with an additional offset parameter. Furthermore, F2 is also interesting from a conceptual perspective, as the single weightings per individual connection resemble the relative importance of the individual connections. This actually means that F2 allows to include inherently the mutual influence of individual connections that was observed in Chapter 8.

*Function F3* The rationale for F3 is to explicitly model that mutual influence: each  $Q_{ij}$  score is weighted with a term which is in turn a function of the  $Q_{ij}$  scores of the respective other individual connections. The idea for this function is to model that strong impairments, i.e. connections with low quality scores, negatively effect the other connections. In other words, the contribution of a  $Q_{ij}$  score to the overall  $Q_i$  score is reduced when the other  $Q_{ij}$  scores are low, and vice versa. The choice fell on the sigmoid function  $\sigma_{ix}(Q_{ix}) = \frac{1}{1+e^{-s_{ix} \cdot (Q_{ix}-o_{ix})}}$ ,  $x \in [i, j, k]$ , with one parameter  $s_{ix}$  describing the steepness and one parameter  $o_{ix}$  describing the zero-crossing point. Figure 9.4 visualizes the  $\sigma$  function for different values of  $s_{ix}$  and  $o_{ix}$ .

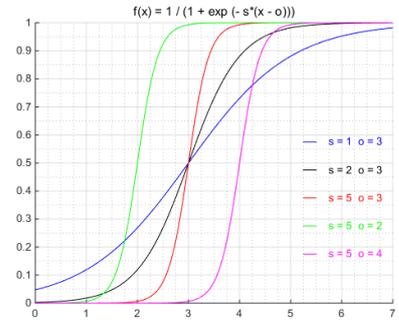


Figure 9.4: Visualization of the  $\sigma$  function for different values of  $s_{ix}$  and  $o_{ix}$ .

*Function F4* The rationale for F4 is to combine the two modeling functions F1 and F7. The idea is to exploit the observation that in many cases in Figures 9.1 to 9.3 the mean of  $Q_{ij}$  is close to  $Q_i$  when  $Q_i$  is high and the minimum value is close when  $Q_i$  is low. This leads to a quality dependent weighting of the mean and the minimum value, whereas this quality dependency is realized as a function of a pre-estimation of  $Q_i$ , favoring the mean of  $Q_{ij}$  for higher values of the pre-estimate and the minimum value of  $Q_{ij}$  for lower values of the pre-estimate. This in turn is realized by using two sigmoid functions  $\sigma_1$  and  $\sigma_2$  with inverse orientations (see Figure 9.5 for a visualization), and by using the simple mean of  $Q_{ij}$  as pre-estimation of  $Q_i$ .

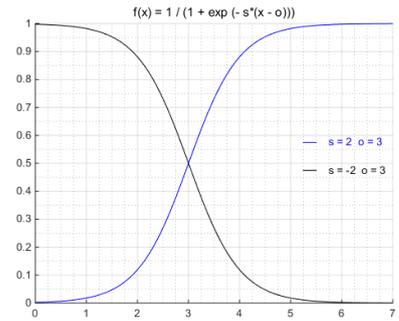


Figure 9.5: Visualization of the  $\sigma$  function with inverse orientations by inverting the sign of  $s_{ix}$ .

*Function F5* The rationale for F5 is to model the mutual influence of individual connections as a function of outliers: If an *Individual Connection Quality* score is an outlier, i.e. the score is substantially different from the other individual connections, then this outlier gets a reduced weighting. From a conceptual perspective, this reflects the possibility that when participants perceive one particularly different connection, he or she will tend to attribute the quality to the corresponding interlocutors or the corresponding device or environment. This means that the participant does not or just in parts attribute the quality to the system, which in turn means, the contribution of that connection has a lower contribution to the *Single-Perspective Telemeeting Quality* score. Hence, the weighting of that connection is reduced.

The implementation chosen here is first to detect an outlier by threshold operation:  $Q_{ij}$  is an outlier if it lies outside the standard deviation around the mean of all  $Q_{ij}$  scores of that participant and call. Then, the effect of outliers is realized such that two aspects are covered.

The first aspect is that the more outliers are in a telemeeting, the lower is the effect of outliers. This is modeled here as a linearly increasing function of number of detected outliers:  $\frac{1-\beta}{\#IL-1} \cdot \#outliers + \frac{\#IL \cdot \beta}{\#IL-1}$ , with  $\beta$  determining the minimum possible weight in case that there is only one outlier, and  $\#IL$  being the total number of interlocutors or connections, respectively. This particular realization converges to a weight of one in case that all connections are outliers and in case that the number of interlocutors is large; both cases meaning that the impact of a single outlier becomes negligible. The second aspect is that the more the different  $Q_{ij}$  are spread across the quality range, the lower is the effect of outliers. This is modeled here as a linearly increasing function of the spread, here expressed by the standard deviation, of all  $Q_{ij}$  scores of that participant and call:  $\frac{1-\delta}{\gamma} \cdot std(Q_{ix}) + \delta$ , with  $\gamma$  determining the value of the standard deviation for which the weight should be equal to one, and  $\delta$  determining the minimum weight in case that the standard deviation is equal to zero.

*Function F8* The rationale for F8 is the third fundamental modeling idea, i.e.  $Q_i$  is a weighted mean of the quality impairments on the individual connections instead of the plain quality scores of those connections. Function F8, which expresses this relation in terms of plain quality scores, can be derived from this as follows. Starting point is the weighted sum of quality impairments:

$$QI_i = w_{ii} \cdot QI_{ii} + w_{ij} \cdot QI_{ij} + w_{ik} \cdot QI_{ik} \quad (9.1)$$

Then, quality impairments can be described as the difference of the maximum possible quality and the observed quality:

$$QI_x = Q_{max} - Q_x \quad (9.2)$$

Substituting the terms in Equation 9.1 with Equation 9.2 leads to:

$$\begin{aligned} Q_{max} - Q_i &= w_{ii} \cdot (Q_{max} - Q_{ii}) + w_{ij} \cdot (Q_{max} - Q_{ij}) + w_{ik} \cdot (Q_{max} - Q_{ik}) \\ Q_i &= Q_{max} - w_{ii} \cdot Q_{max} - w_{ij} \cdot Q_{max} - w_{ik} \cdot Q_{max} + w_{ii} \cdot Q_{ii} + w_{ij} \cdot Q_{ij} + w_{ik} \cdot Q_{ik} \\ Q_i &= (1 - w_{ii} - w_{ij} - w_{ik}) \cdot Q_{max} + w_{ii} \cdot Q_{ii} + w_{ij} \cdot Q_{ij} + w_{ik} \cdot Q_{ik} \end{aligned} \quad (9.3)$$

Finally, Function F8 is then realized by allowing for an additional degree of freedom by replacing  $(1 - w_{ii} - w_{ij} - w_{ik})$  with a free parameter  $w_{max}$ , which, however, is limited to negative values or zero. Setting  $w_{max} \leq 0$  allows F8 to compute predictions of  $Q_i$  that are smaller than the minimum of  $Q_{ij}$ , a behavior that was found for some of the noise and echo conditions in ACT<sub>3</sub> and LOT<sub>2</sub>.

### 9.2.3 Modeling Factors

While the modeling functions constitute in a manner of speaking the core of the modeling algorithms, the following number of further aspects, here called modeling factors, differentiate the individual modeling algorithms that have been implemented in this work.

*Raw scores vs. mean opinion scores* The first modeling factor concerns the question, whether the raw *Single-Perspective Telemeeting Quality* scores  $Q_i$  per participant and call are predicted based on the *Individual Connection Quality* scores  $Q_{ij}$ , or whether the mean opinion scores (MOS) of the *Single-Perspective Telemeeting Quality*  $MOS(Q_i)$  are predicted based on the MOS of the *Individual Connection Quality* scores  $MOS(Q_{ij})$ , whereas MOS are computed as the mean of ratings across all participants and calls per technical condition. The present work focuses only on the raw quality scores, since the number of conditions per experiment is rather low, which in turn can lead to overtraining effects if MOS are used (see also Section 9.5.2 for a more detailed discussion on this).

*Training and testing on whole data, per subset or per technical condition* Another modeling factor concerns the model training and evaluation procedure explained in the next section. Considering that some modeling functions have free parameters that can be trained, such a training is possible for four different subsets of the data:

- Training on the whole data.
- Training per technical condition. This is useful for modeling attempts in which the ground-truth data (here the  $Q_i$  ratings) shows different behavior between technical conditions.
- Training per subset of the data, whereas a subset is freely definable and does not need to be the same than a technical condition. This is useful for modeling attempts in which subsets can contain multiple technical conditions, such as audio-only, video-only, and audiovisual impairments.
- Training per combination of technical condition and subset.

Similarly, model performance can be tested on the same four levels; thus 16 combinations of training and test levels are possible. The present work, however, focuses only on the two levels *whole data* and *per condition*, leading to four combinations of training and test levels.

*Assignment of scores to weights* Since the modeling functions F2, F3 and F8 use weights that are (per training data set) constant but not equal for all individual connections, a well-defined way to assign the *Individual Connection Quality* scores to the individual weights is required. In Skowronek et al.<sup>2</sup>, we explained this need with the following example of Function F2: “Suppose that the computation of  $Q$  for a conference call with three participants is  $Q = w_A * Q_A + w_B * Q_B + w_C * Q_C$ , where  $w_A$ ,  $w_B$ , and  $w_C$  are three different fixed weights, and  $Q_A$ ,  $Q_B$  and  $Q_C$  are the individual connection quality scores. In order to run this computation, there must be a method that determines which of the three individual connection quality scores is actually  $Q_A$ , which is  $Q_B$ , and which is  $Q_C$ , and hence which are assigned the respective weights  $w_A$ ,  $w_B$ , and  $w_C$ .”

<sup>2</sup> Janto Skowronek et al. “Method and Apparatus for Computing the Perceived Quality of a Multiparty Audio or Audiovisual Telecommunication Service or System”. Patent application WO/2016/041593 (EP). Deutsche Telekom AG. Mar. 2016

In this work, two alternatives for this assignment are considered. The first alternative is essentially a rule-based approach. It follows the systematic method for defining technical conditions by means of the impairments perceived by each interlocutor (see Chapter 8) and the corresponding technical analysis and data organization algorithm (see Sections 6.3 & 6.5). Applying this systematic method here means

- a) to always assign the quality scores of the own connection to the first variable  $Q_{ii}$ ,
- b) to assign – in case of asymmetric configurations with an unimpaired connection – the unimpaired connection to the second variable  $Q_{ij}$  and the impaired connection to the third variable  $Q_{ik}$ ,
- c) to assign – in case of asymmetric configurations with two different impairments – the presumably weaker impairment to the second variable  $Q_{ij}$  and the presumably stronger impairment to the third variable  $Q_{ik}$ , and
- d) to arbitrarily assign – in case of symmetric configurations including the reference condition – one of the two connections to the second variable  $Q_{ij}$  and the other connection to the third variable  $Q_{ik}$ .

This assignment is inherently conducted when applying the data organization algorithm of Section 6.5.

The second alternative is to always assign the quality scores of the own connection to the first variable  $Q_{ii}$ , and to assign the other (here: two) connections to the remaining variables by sorting the corresponding quality scores in ascending order. This means here, the minimum value of those two scores is assigned to the second variable  $Q_{ij}$  and the maximum value is assigned to the third variable  $Q_{ik}$ .

### 9.3 Model Training and Evaluation Procedure

#### 9.3.1 Motivation and Basic Principle

Some model functions have parameters that can freely be set to provide optimal model performance, which means, those modeling functions can be trained by means of data fitting or machine learning approaches. The approach used here is to apply a parameter training and model performance procedure that applies a three step approach typically used in the machine learning domain: first, split the data into training and test sets; second optimize the parameters for the training set; and third, evaluate the performance of those trained models with the test set, using appropriate performance measures. Since training and test results depend on the actual splitting of the data into the training and test sets, the common approach is to conduct repetitions of this procedure to obtain a more stable estimate of the true model performance. In the data fitting and machine learning domain, multiple alternatives are known to realize this repetition: for instance k-fold cross validation, leave-one-out validation, and bootstrapping. The author opted for the bootstrap method, which

has been developed by Efron<sup>3</sup>. The motivation is that the present data sets are quite small for robust model fitting procedures; and the bootstrap method is specifically suitable for estimating the true value of a statistical estimate, which is here the true mean of the model performance, for small sample sizes. While Zoubir<sup>4</sup> for instance provides a more extensive introduction of the bootstrap method, the basic idea of this method is to construct many variations of a data sample (typically in the order of 100 to 1000) by randomly *drawing* individual observations of that original sample, whereas individual observations may be drawn multiple times for each new data sample. In order to apply this method for the model evaluation task at hand, the author opted for a modification<sup>5</sup> as follows: The approach is to conduct many repetitions in which the data is randomly split into training or test sets, whereas the split is – in contrast to k-fold cross validation or leave-one-out method – done independently from the previous iteration. This allows that the training and test sets in each iteration may include again some of the observations of the previous iteration, which resembles the idea of bootstrapping. At the same time this ensures in each iteration that no individual observations are present in both training and test sets.

### 9.3.2 Performance Measures

Two well known measures are used to assess the performance of the different modeling functions: the Pearson correlation coefficient  $Rho$  and the root mean square error  $RMSE$ , both computed between the estimated and true scores of the *Single-Perspective Telemeeting Quality*  $Q_i$ . Furthermore, since the bootstrap repetition procedure is used to obtain a good estimate of the true modeling performance, which means the two performance measures are computed per repetition, it is possible to compute for both measures mean values ( $meanRho$ ,  $meanRMSE$ ) and 95% confidence intervals ( $ci95Rho$ ,  $ci95RMSE$ ). The mean values give an estimation of the expected performance, the confidence interval provides insights how stable this estimation is. To compare the performance of the eight different modeling functions F1 to F8, the author used errorbar plots (mean values and confidence intervals) to visualize differences in the performance of the functions, and formal statistical analysis (ANOVA and PostHoc tests) to determine whether differences are significant.

### 9.3.3 Selection of the Number of Bootstrap Repetitions

*Motivation* The number of bootstrap repetitions is a free parameter in the model evaluation procedure, which needs to be carefully chosen for two reasons. The first reason concerns practical issues: In the literature typical values for this number range between 100 and 1000, whereas it depends on the actual data, whether this number of repetitions is sufficient to estimate the true value. However, considering that for six of the eight modeling functions each bootstrap repetition requires a training, which in turn is calling a parameter search routine,

<sup>3</sup> Bradley Efron. "Bootstrap Methods: Another Look at the Jackknife". In: *The Annals of Statistics* 7.1 (1979), pp. 1–26

<sup>4</sup> A.M. Zoubir and B. Boashash. "The Bootstrap and its Application in Signal Processing". In: *IEEE Signal Processing Magazine* (Jan. 1998), pp. 56–76

<sup>5</sup> This modification has been used in the literature by others as well as by the author, e.g.:

Catherine Champagne et al. "A bootstrap method for assessing classification accuracy and confidence for agricultural land use mapping in Canada". In: *International Journal of Applied Earth Observation and Geoinformation* 29 (2014), pp. 44–52

Janto Skowronek, Martin McKinney, and Steven van de Par. "A demonstrator for automatic music mood estimation". In: *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*. 2007

computation times for such large numbers of repetitions can become very long. For that reason, a good balance between the accuracy of the performance estimate and the computation time is desired.

The second reason concerns methodological issues: While a larger number of bootstrap repetitions means to obtain a more stable estimate of the model performance, i.e. the mean of the performance measures Rho and RMSE, it also means that any desired sensitivity to *differentiate* between the modeling functions can be achieved, as long as the number of repetitions is large enough. The reason is that the confidence interval, which is used to distinguish between the performance of the different modeling functions, is per definition monotonically decreasing with an increasing number of observations  $N$ , due to the following relation<sup>6</sup>:

$$\begin{aligned}
 CI &= t_{n-1} \cdot SE \\
 &= t_{n-1} \cdot \frac{STD}{\sqrt{N}} \\
 &= t_{n-1} \cdot \frac{\sqrt{\frac{\sum(x-\bar{x})^2}{N-1}}}{\sqrt{N}} \\
 &= t_{n-1} \cdot \frac{\sqrt{\sum(x-\bar{x})^2}}{\sqrt{N-1} \cdot \sqrt{N}}
 \end{aligned} \tag{9.4}$$

with  $SE$  standard error,  $STD$  standard deviation, and  $t_{n-1}$  the value of the  $t$ -distribution for a degree of freedom of  $n - 1$ .

For that reason, limiting the number of repetitions  $N$  would lead to more conservative results, as confidence intervals are larger and thus the performance of different modeling functions must differ more before significant differences can be observed.

*Selection Process and Criteria* To make an informed selection of the number of bootstrap repetitions, the author computed the performance measures for the two functions F1 and F7 from one up to 1000 repetitions. The two functions do not require training, which makes it feasible to run large numbers of repetitions. Furthermore, the estimated telemeeting quality scores  $\widehat{Q}_i$  of these two functions represent for most technical conditions the upper and lower boundaries of the true *Single-Perspective Telemeeting Quality* scores  $Q_i$  (see Figures 9.1 to 9.3). Thus, the relation between modeling performance and number of bootstrap conditions for F1 and F7 is representative for this relation for all modeling functions.

The selection of the number of bootstrap repetitions is based on analyzing two aspects: the stability of estimating the true mean performance values, and the sensitivity to differentiate the modeling functions. Concerning the stability, the deviation of the mean performance values between the chosen and the maximum tested number of repetitions should be acceptable. This is the case when the deviation is in the order of the confidence intervals around the corresponding mean performance values. Concerning the sensitivity, the mean performance values of each function should not lie inside the confidence intervals of the respective other function, as this indicates significantly different modeling performance between the two functions.

<sup>6</sup> Andy Field. *Discovering Statistics using SPSS*. 3rd ed. SAGE publications, 2009

*Selection Result* Figures 9.6 to 9.8 provide the detailed analysis for the seven experiments ACT1 to AVCT2. Based on these figures, the number of bootstrap repetitions is set to 20, as this value fulfills the selection criteria for all except two cases.

---

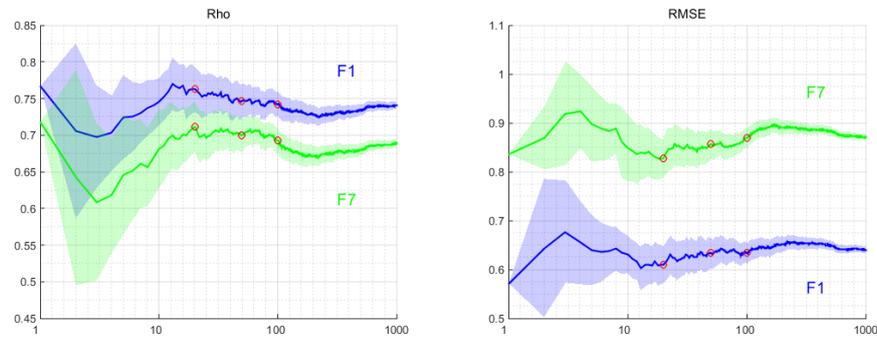


---

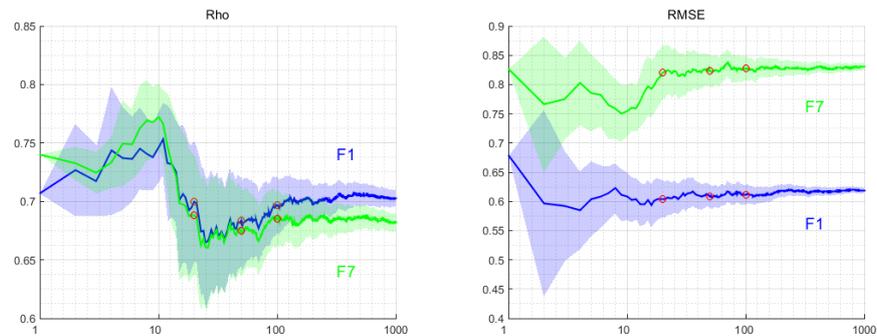
### 1. Visualization:

The plots show the two performance measures Rho and RMSE as a function of the number of bootstrap repetitions. The results are shown for the two baseline modeling functions F1 “simple mean” and F7 “simple minimum value”. The solid lines show the mean values of each performance measure, the shaded areas indicate the 95% confidence intervals around the mean values. The red circles indicate the mean values for 20, 50 and 100 bootstrap repetitions.

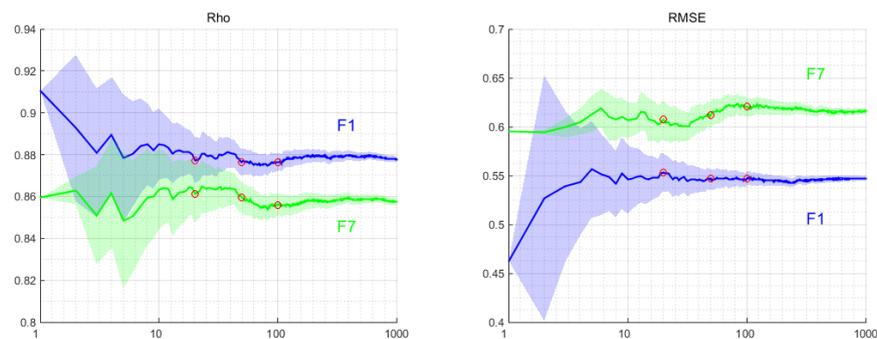
#### 2.1 Results Experiment ACT1.



#### 2.2 Results Experiment ACT2.



#### 2.3 Results Experiment ACT3.




---



---

#### 3.1 Analysis concerning stable estimates of the true mean values:

The mean values beyond 100 repetitions slightly change with increasing number of repetitions, which indicates that 100 repetitions already provide stable estimates. However, the mean values for 50 and 20 repetitions are also relatively close; the differences are in the order of the respective confidence intervals. Thus, the number of repetitions can be reduced down to 20 with acceptable deviations from the true mean.

#### 3.2 Analysis concerning sensitivity to differentiate the modeling functions:

With the exception of Rho in Experiment ACT2, the mean values of each modeling function do not lie within the confidence interval of the respective other modeling function for neither 20, 50 or 100 repetitions. Thus, significant differences between the two modeling functions can be expected. This means that already 20 repetitions provide in most cases a sufficient sensitivity to differentiate the modeling functions. In the case of Rho in Experiment ACT2, about 300 repetitions would be necessary; a value that is too large to be practically feasible.

---



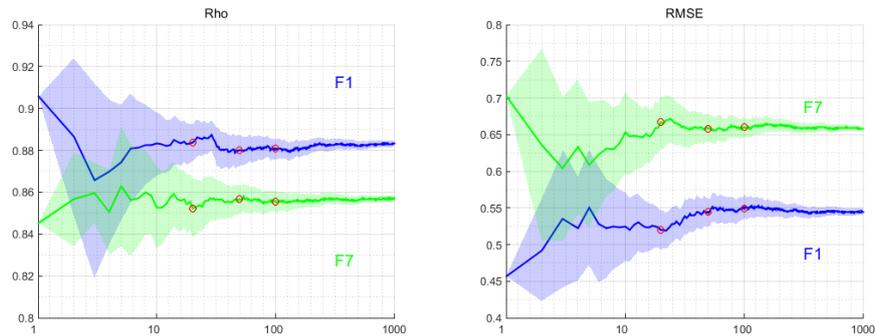
---

Figure 9.6: Analysis for selecting the number of bootstrap repetitions, results of audio-only conversation tests.

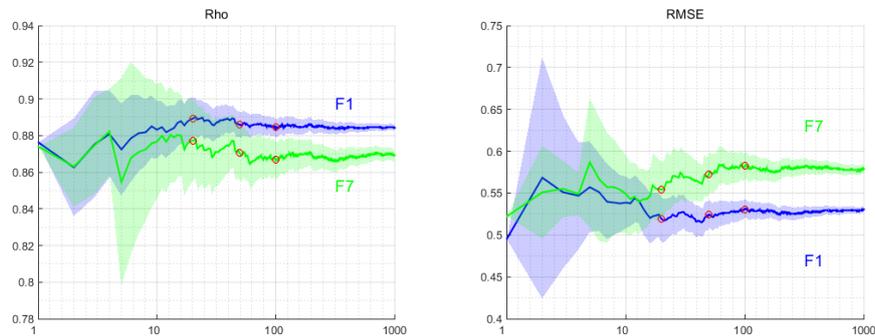
1. Visualization:

The plots show the two performance measures Rho and RMSE as a function of the number of bootstrap repetitions. The results are shown for the two baseline modeling functions F1 “simple mean” and F7 “simple minimum value”. The solid lines show the mean values of each performance measure, the shaded areas indicate the 95% confidence intervals around the mean values. The red circles indicate the mean values for 20, 50 and 100 bootstrap repetitions.

2.1 Results Experiment LOT1.



2.2 Results Experiment LOT2.



3.1 Analysis concerning stable estimates of the true mean values:

The mean values beyond 100 repetitions slightly change with increasing number of repetitions, which indicates that 100 repetitions already provide stable estimates. However, the mean values for 50 and 20 repetitions are also relatively close; the differences are in the order of the respective confidence intervals. Thus, the number of repetitions can be reduced down to 20 with acceptable deviations from the true mean.

3.2 Analysis concerning sensitivity to differentiate the modeling functions:

With the exception of Rho in Experiment LOT2, the mean values of each modeling function do not lie within the confidence interval of the respective other modeling function for neither 20, 50 or 100 repetitions. Thus, significant differences between the two modeling functions can be expected. This means that already 20 repetitions provide in most cases a sufficient sensitivity to differentiate the modeling functions. In the case of Rho in Experiment LOT2, about 60 repetitions would be necessary, while for 20 repetitions only the mean value of F1 lies in the confidence intervals of F7, but not the mean value of F7 in the confidence interval of F1.

Figure 9.7: Analysis for selecting the number of bootstrap repetitions, results of listening-only tests.

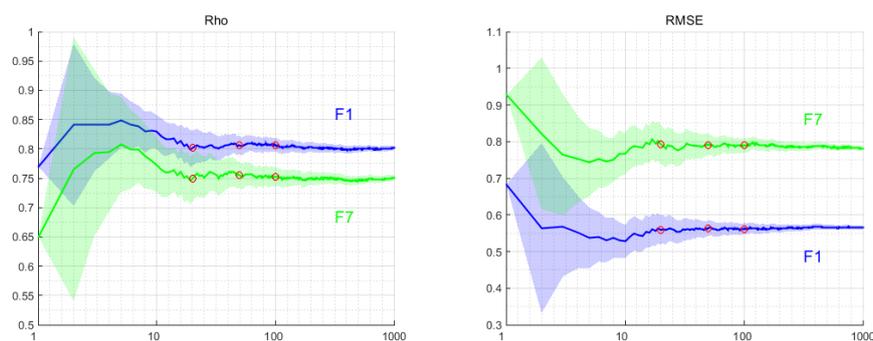
---

### 1. Visualization:

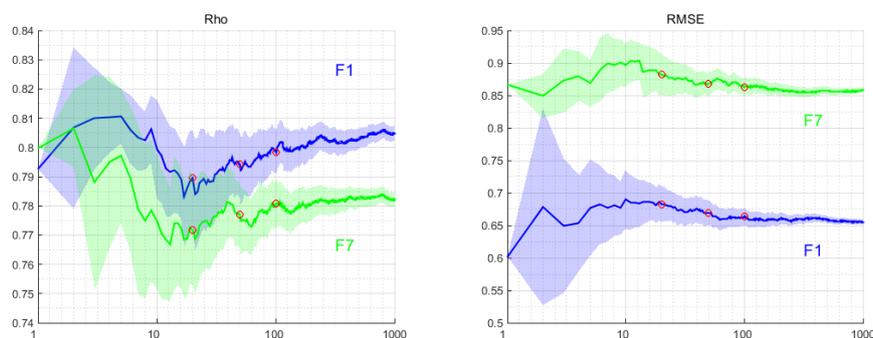
The plots show the two performance measures Rho and RMSE as a function of the number of bootstrap repetitions. The results are shown for the two baseline modeling functions F1 “simple mean” and F7 “simple minimum value”. The solid lines show the mean values of each performance measure, the shaded areas indicate the 95% confidence intervals around the mean values. The red circles indicate the mean values for 20, 50 and 100 bootstrap repetitions.

---

#### 2.1 Results Experiment AVCT1.



#### 2.2 Results Experiment AVCT2.



#### 3.1 Analysis concerning stable estimates of the true mean values:

The mean values beyond 100 repetitions slightly change with increasing number of repetitions, which indicates that 100 repetitions already provide stable estimates. However, the mean values for 50 and 20 repetitions are also relatively close; the differences are in the order of the respective confidence intervals. Thus, the number of repetitions can be reduced down to 20 with acceptable deviations from the true mean.

#### 3.2 Analysis concerning sensitivity to differentiate the modeling functions:

With the exception of Rho in Experiment AVCT2, the mean values of each modeling function do not lie within the confidence interval of the respective other modeling function for neither 20, 50 or 100 repetitions. Thus, significant differences between the two modeling functions can be expected. This means that already 20 repetitions provide in most cases a sufficient sensitivity to differentiate the modeling functions. In the case of Rho in Experiment AVCT2, about 40 repetitions would be necessary, while at 20 repetitions the mean values of each function are just touching the confidence intervals of the other function.

---

Figure 9.8: Analysis for selecting the number of bootstrap repetitions, results of audiovisual conversation tests.

#### 9.3.4 *Algorithm to Split the Data into Training and Test Sets*

Within each bootstrap repetition, the data is randomly split into training and test sets. To do this, the author implemented an algorithm that splits the data according to three boundary conditions:

1. It tries to achieve the target ratio between training and test set for the individual technical conditions represented in the data.
2. It tries to achieve the target ratio between training and test set for individual subsets of the data, whereas those subsets are freely definable and do not need to be the same than the technical conditions.
3. It allows to link individual observations independently from subsets of technical conditions such that all observations linked by such a variable are either assigned to the training or to the test set. The foremost and actually applied use case is to consider the test subjects as the linking variable, which means all observations belonging to one subject are either assigned to the training or test data.

Given this approach, the modeling framework allows to run all 16 combinations of the four training and test levels mentioned in Section 9.2: the whole data set, per subset, per condition and per combination of subset and condition.

#### 9.3.5 *Selection of Proportion of Training and Test Data*

*Motivation* Similar to the number of bootstrap repetitions, the proportion between the sizes of the training and test sets can influence the modeling performance. Usually the larger the amount of training data is, the higher the modeling performance – on average – gets. However, increasing the size of the training data means reducing the size of the test data, which in turns increases the confidence intervals. Thus a good balance needs to be found in improving the average performance and limiting the uncertainty – the confidence intervals – around this. Typical values for this proportion are 70% or 80% training data; and the author also opted for a proportion of 80%. Nevertheless, it would be good to validate, if this value is also appropriate for the present data.

*Validation Process* To validate the selection of a proportion of 80%, the author computed the performance measures for the two functions F2 and F6 from proportion values between 0.1 and 0.9. These two functions are the closest alternatives to the baseline functions F1 and F7 that require training. A candidate proportion is appropriate when it provides a good balance between achieving optimal mean performance and limiting the uncertainty around them. Concerning the optimal performance, this is considered here to be the case when the mean performance is among the best performing values across the seven studies. Concerning the limitation of uncertainty, this is

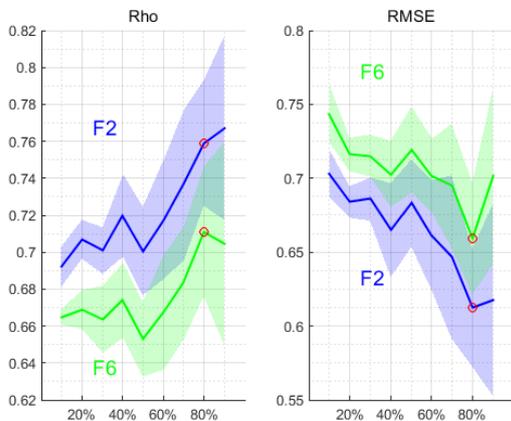
considered here to be the case when the mean performance values of each function do not lie inside the confidence intervals of the respective other function.

*Selection Result* Figures 9.9 and 9.10 provide the detailed analysis for the seven experiments ACT<sub>1</sub> to AVCT<sub>2</sub>. Based on these figures, a proportion of 80% training data is considered here to be appropriate, as this value achieves the desired balance for all but one experiment (ACT<sub>2</sub>).

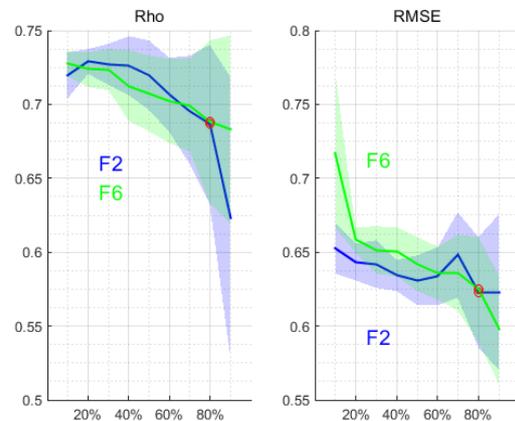
### 1. Visualization:

The plots show the two performance measures Rho and RMSE as a function of the number of bootstrap repetitions. The results are shown for the two baseline modeling functions F2 “weighted mean” and F6 “weighted minimum value”. The solid lines show the mean values of each performance measure, the shaded areas indicate the 95% confidence intervals around the mean values. The red circles indicate the mean values for a proportion of 80% training data.

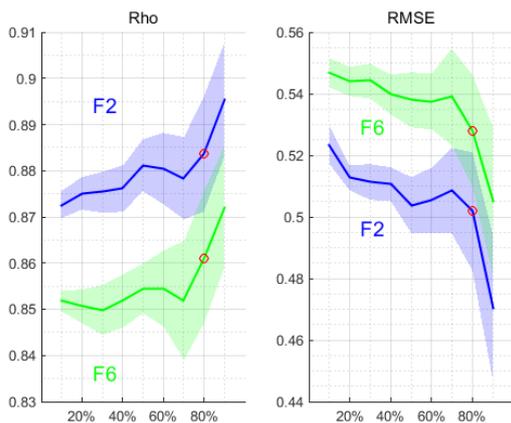
#### 2.1 Results Experiment ACT<sub>1</sub>.



#### 2.2 Results Experiment ACT<sub>2</sub>.



#### 2.3 Results Experiment ACT<sub>3</sub>.



### 3. Analysis

For Experiments ACT<sub>1</sub> and ACT<sub>3</sub>, a proportion of 80% gives for both functions F2 and F6 and both measures Rho and RMSE the optimal results, and the mean performance values of each function are outside the confidence intervals of the respective other function.

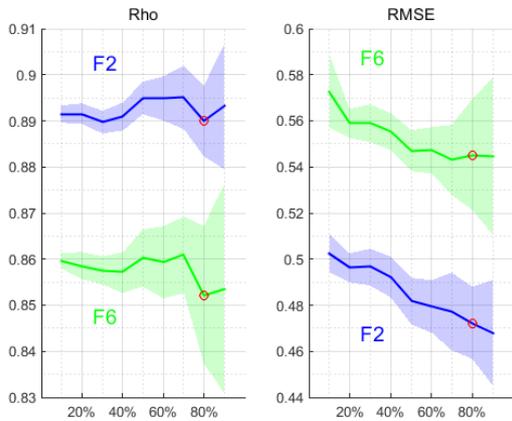
For Experiment ACT<sub>2</sub>, the results concerning the mean performance are inconsistent, as Rho is best for small proportions and RMSE is best for large proportions, and there is in almost all cases a overlap of the mean performance values of each function with the confidence intervals of the respective other function.

Figure 9.9: Analysis for selecting the proportion of training data, results of audio-only conversation tests.

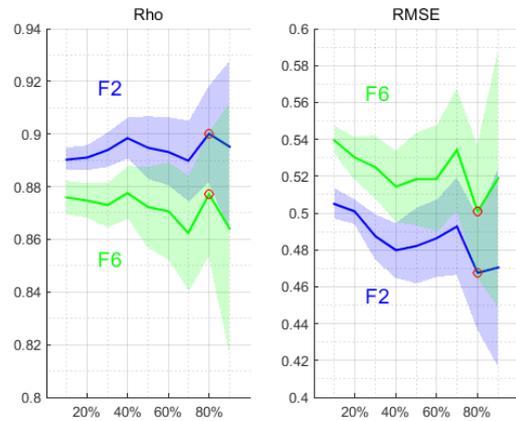
1. Visualization:

The plots show the two performance measures Rho and RMSE as a function of the number of bootstrap repetitions. The results are shown for the two baseline modeling functions F2 “weighted mean” and F6 “weighted minimum value”. The solid lines show the mean values of each performance measure, the shaded areas indicate the 95% confidence intervals around the mean values. The red circles indicate the mean values for a proportion of 80% training data.

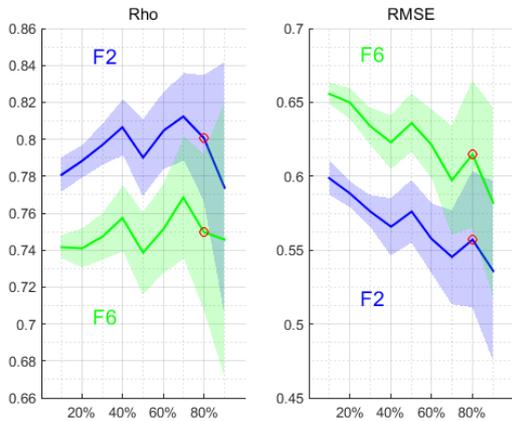
2.1 Results Experiment LOT1.



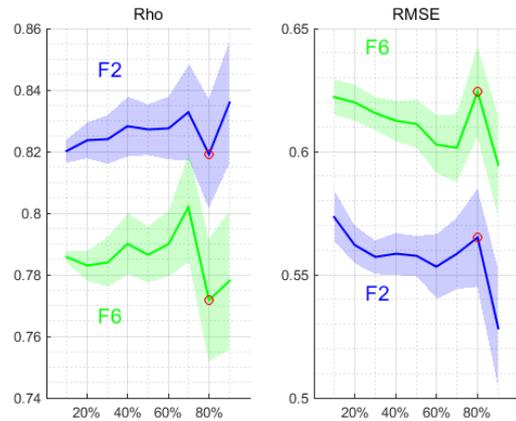
2.2 Results Experiment LOT2.



2.1 Results Experiment AVCT1.



2.2 Results Experiment AVCT2.



3. Analysis

For all four experiments, a proportion of 80% gives for both functions F2 and F6 and both measures Rho and RMSE the optimal results, and the mean performance values of each function are outside – or in case of LOT2 at the border of – the confidence intervals of the respective other function.

Figure 9.10: Analysis for selecting the proportion of training data, results of listening-only and audiovisual conversation tests.

### 9.4 Model Performance

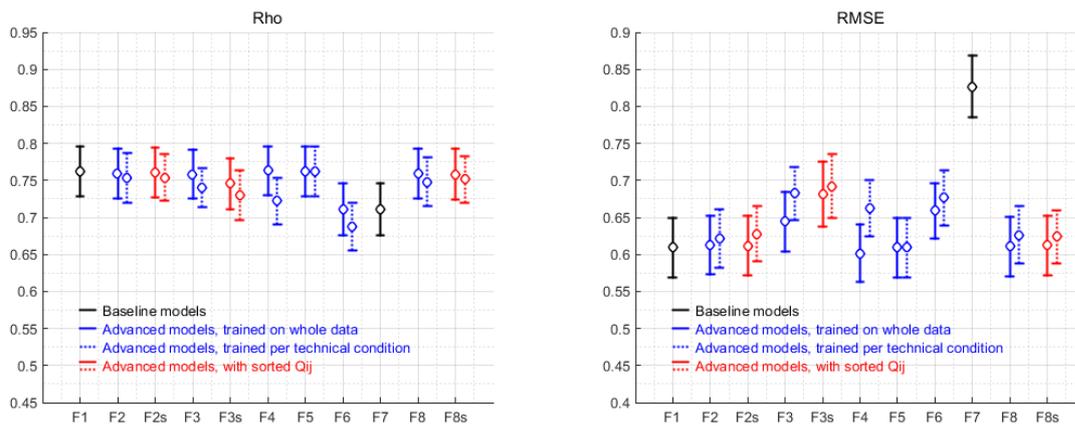
Figures 9.11 to 9.17 show per experiment the visualization of the model performance across the technical conditions. The figures also provide summaries of the statistical analysis conducted to identify significant differences between the modeling functions (F1 – F8) and the modeling factors (Section 9.2.3): training on whole data vs. training per condition, rule-based vs. sorting-based assignment of weighting parameters for functions F2, F3 & F8. Furthermore, the figures also provide summaries of the analysis results conducted per technical condition, in order to verify if individual functions perform differently for individual technical conditions. The analysis in full detail is given in Figures H.1 to H.52 in Appendix H.

---

#### 1. Results across technical conditions

---

1.1 Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



#### 1.2 Observations and Conclusions:

Based on detailed analysis in Figure H.1, for A & B using ANOVAs & PostHoc tests, for C using t-Tests.

- A) No advanced model significantly outperforms the baseline mean F1.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition reduces performance for F4 for Rho.

---

#### 2. Results per technical condition:

Based on detailed analysis in Figures H.2 to H.5, for A & B using ANOVAs & PostHoc tests, for C using t-Tests.

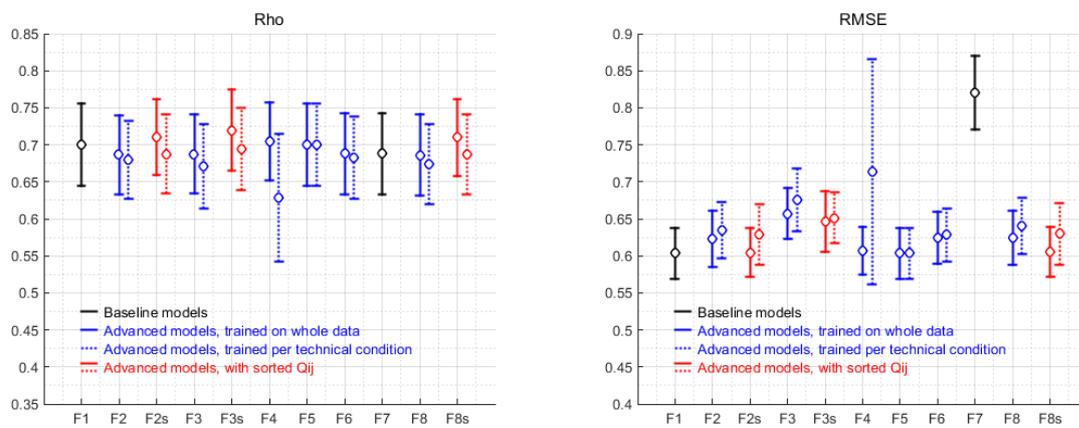
Effects	Condition(s)
A) No advanced model significantly outperforms the baseline mean F1.	All (Ref <sub>1</sub> , L <sub>3</sub> , EN <sub>3</sub> , TN <sub>3</sub> , Dsym, E <sub>1</sub> , P <sub>1</sub> )
B) Equal performance for sorting-based and rule-based assignment of $Q_{ij}$ .	All
C.1) Training per condition reduces performance for F4.	Dsym
C.2) Training per condition does not improve performance.	All other

---

Figure 9.11: Modeling performance across all technical conditions of Experiment ACT1.

**1. Results across technical conditions**

**1.1 Visualizing the results:** Errorbar plots showing the mean values & 95% confidence intervals.



**1.2 Observations and Conclusions:**

Based on detailed analysis in Figure H.6, for A & B using ANOVAs & PostHoc tests, for C using t-Tests.

- A) No advanced model outperforms the baseline mean F1.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition leads for F4 to reduced and instable (large confidence intervals) performance.

**2. Results per technical condition:**

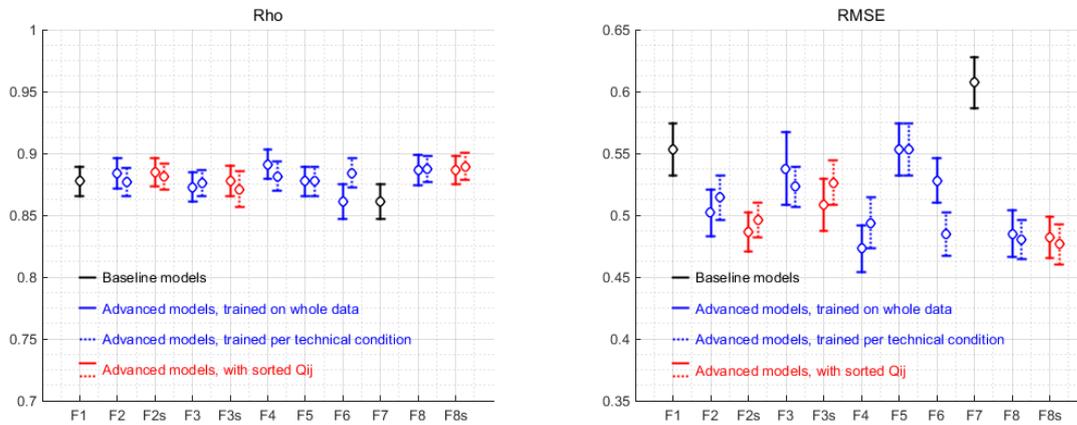
Based on detailed analysis in Figures H.7 to H.8, for A & B using ANOVAs & PostHoc tests, for C using t-Tests.

Effects	Condition(s)
A) No advanced model significantly outperforms the baseline mean.	All (Ref2, L1, L2, L4)
B) Equal performance for sorting-based and rule-based assignment of $Q_{ij}$ .	All
C.1) Training per condition destabilizes performance (large confidence intervals) for F4.	L1
C.2) Training per condition does not improve performance.	All other

Figure 9.12: Modeling performance across all technical conditions of Experiment ACT2.

**1. Results across technical conditions**

**1.1 Visualizing the results:** Errorbar plots showing the mean values & 95% confidence intervals.



**1.2 Observations and Conclusions:**

Based on detailed analysis in Figure H.9, for A & B using ANOVAs & PostHoc tests, for C using t-Tests.

- A) Six advanced models (F2s, F4 & F8 for both training modes, F8s & F2 trained on whole data and F6 trained per condition) outperform the baseline mean F1. F4 is the best, though not sig. different from the other five.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition improves performance only for F6.

**2. Results per technical condition:**

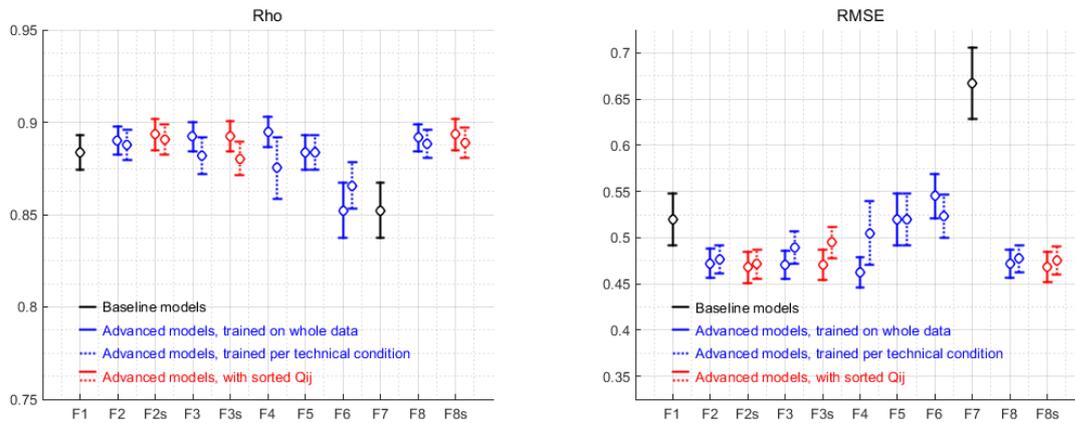
Based on detailed analysis in Figures H.10 to H.18, for A & B using ANOVAs & PostHoc tests, for C using t-Tests.

Effects	Condition(s)
A.1) No advanced model significantly outperforms the baseline mean.	Ref3, B1, B2, B2sym, FM1, FL3sym, FM2, FL4sym, P1, P2, P3
A.2) No advanced model significantly outperforms the baseline mean. F2 is slightly better (not sig.)	P4
A.3) F3, F6 & F8 all trained per condition outperform the baseline mean F1 but not the baseline minimum F7.	TN1
A.4) F3, F3s, F6 & F8 (all trained per condition) outperform the baseline mean F1 but not the baseline minimum F7.	TN2
A.5) F2s, F3, F3s, F4, F6, F8 & F8s outperform the baseline mean F1 but not the baseline minimum F7.	E2
A.6) F2, F2s, F3, F3s, F4, F6, F8 & F8s outperform the baseline mean F1 but not the baseline minimum F7.	E3
B) Equal performance for sorting-based and rule-based assignment of $Q_{ij}$ .	All
C.1) Training per condition reduces performance for F3.	B2
C.2) Training per condition improves performance for F3.	P1
C.3) Training per condition improves performance for F3, F6 & F8.	TN1
C.4) Training per condition improves performance for F6, F8 & F8s.	TN2
C.5) Training per condition improves performance for F2, F3s, F6 & F8.	E3
C.6) Training per condition does not improve performance.	All other

Figure 9.13: Modeling performance across all technical conditions of Experiment ACT3.

**1. Results across technical conditions**

**1.2 Visualizing the results:** Errorbar plots showing the mean values & 95% confidence intervals.



**1.2 Observations and Conclusions:**

Based on detailed analysis in Figure H.19, for A & B using ANOVAs & PostHoc tests, for C using t-Tests.

- A) Three advanced models (F2s, F4 & F8s for both training modes) significantly outperform the baseline mean F1. F4 is the best, though not sig. different from the other two. Furthermore, F2 & F8 are slightly better than F1, though not significantly.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition reduces performance or F3s & F4.

**2. Results per technical condition:**

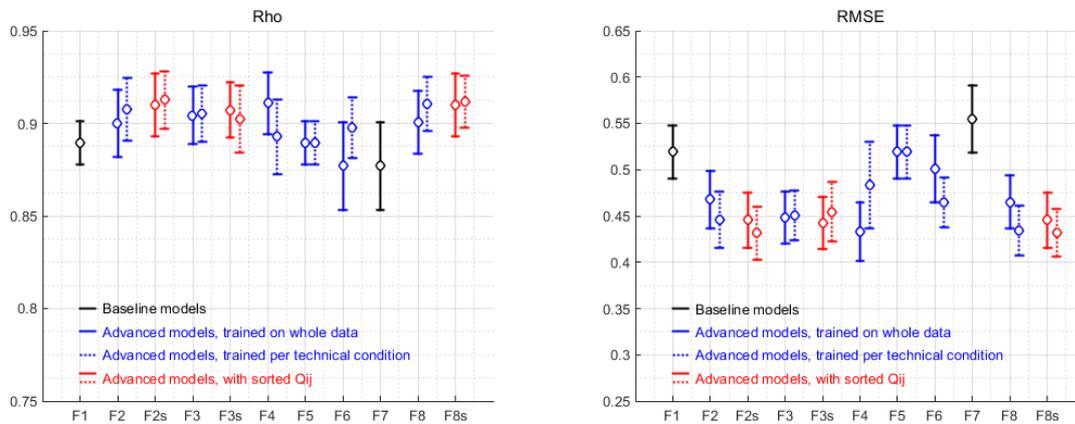
Based on detailed analysis in Figures H.20 to H.24, for A & B using ANOVAs & PostHoc tests, for C using t-Tests.

Effects	Condition(s)
A.1) No advanced model significantly outperforms the baseline mean F1.	Ref5, B2, B2sym, P5, B2symP5, E4, E4sym, CN
A.2) F2, F2s, F3, F3s, F4, F6, F8 & F8s significantly outperform the baseline mean F1.	R
B) Equal performance for sorting-based and rule-based assignment of $Q_{ij}$ .	All
C.1) Training per condition destabilizes performance (large confidence intervals) for F4.	Ref5, E4sym
C.2) Training per condition does not improve performance.	All other

Figure 9.14: Modeling performance across all technical conditions of Experiment LOT1.

**1. Results across technical conditions**

**1.1 Visualizing the results:** Errorbar plots showing the mean values & 95% confidence intervals.



**1.2 Observations and Conclusions:**

Based on detailed analysis in Figure H.25, for A & B using ANOVAs & PostHoc tests, for C using t-Tests.

- A) Seven advanced models (F2 & F8 all trained per condition, F4 trained on whole data, and F2s, F3, F3s, F8s for both training modes) outperform the baseline mean F1 and the baseline minimum F7. F4 (trained on whole data) is the best, though not significantly different from the other six.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

**2. Results per technical condition:**

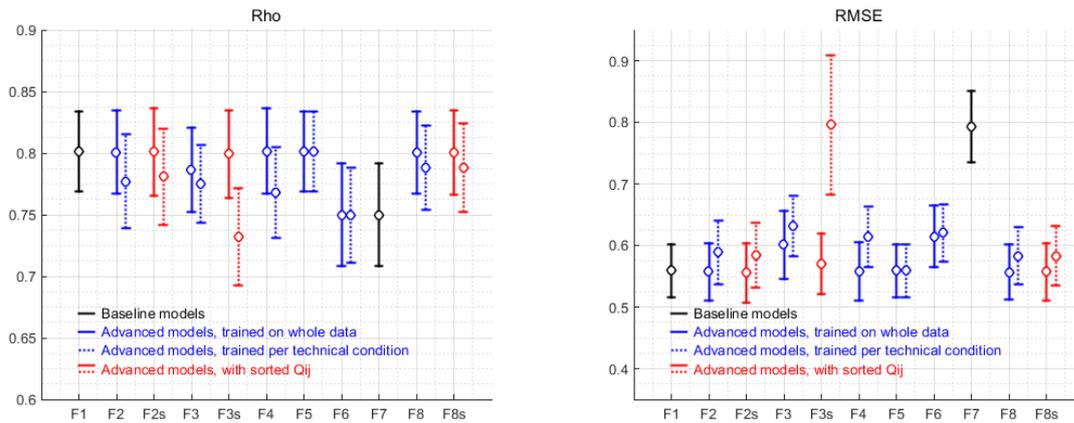
Based on detailed analysis in Figures H.26 to H.37, for A & B using ANOVAs & PostHoc tests, for C using t-Tests.

Effects	Condition(s)
A.1) No advanced model significantly outperforms the baseline mean F1.	Ref3, TN1, TN2, B2, B2sym, FL4sym, FM1, FM1sym, FM2, FM2sym, FM4, FM4sym, FM4P3sym, E2sym, E3, E3sym, P1, P1sym, P3sym
A.2) F2s, F3, F3s, F4, F6, F8 & F8s significantly outperform the baseline mean F1. F4 & F6 (trained on whole data), F3 (trained per condition) and F2s, F3s & F8s outperform the baseline minimum F7.	FM4P3
A.3) F3s (trained on whole data) and F2, F2s, F3, F4, F8 & F8s significantly outperform the baseline mean F1.	E2
A.4) F2, F2s, F3, F3s, F4, F6, F8 & F8s significantly outperform the baseline mean F1. F2 (trained per condition) and F3, F3s & F4 (all trained on whole data) significantly outperform the baseline minimum F7.	P3
A.5) F2, F2s, F3, F3s, F4, F6, F8 & F8s significantly outperform the baseline mean F1.	P4
A.6) F3s & F4 (both trained on whole data) outperform the baseline mean F1, but not the baseline minimum F7.	P4sym
B.1) Sorting-based assignment of $Q_{ij}$ outperforms rule-based assignment for one function (F2s better than F2).	FM4P3
B.2) Equal performance for sorting-based and rule-based assignment of $Q_{ij}$ .	All other
C.1) Training per condition destabilizes performance (large confidence intervals) for F4.	E2sym, E3
C.2) Training per condition reduces performance for F4.	FM4sym
C.3) Training per condition improves performance for F2.	TN1
C.4) Training per condition improves performance for F3.	FM2sym, P3sym
C.5) Training per condition improves performance for F3s.	B2sym
C.6) Training per condition improves performance for F3, F3s, F8 & F8s.	FM2
C.7) Training per condition improves performance for F6.	FL4sym, FM4P3sym, P4sym
C.8) Training per condition improves performance for F6, but destabilizes performance (large confidence intervals) for F3.	E3sym
C.9) Training per condition improves performance for F8.	FM4P3
C.10) Training per condition does not improve performance.	Ref3, TN2, B2, FM1, FM1sym, FM4, E2, P1, P1sym, P3, P4

Figure 9.15: Modeling performance across all technical conditions of Experiment LOT2.

**1. Results across technical conditions**

1.1 Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



**1.2 Observations and Conclusions:**

Based on detailed analysis in Figure H.38, for A & B using ANOVAs & PostHoc tests, for C using t-Tests.

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition reduces performance for F3s.

**2. Results per technical condition:**

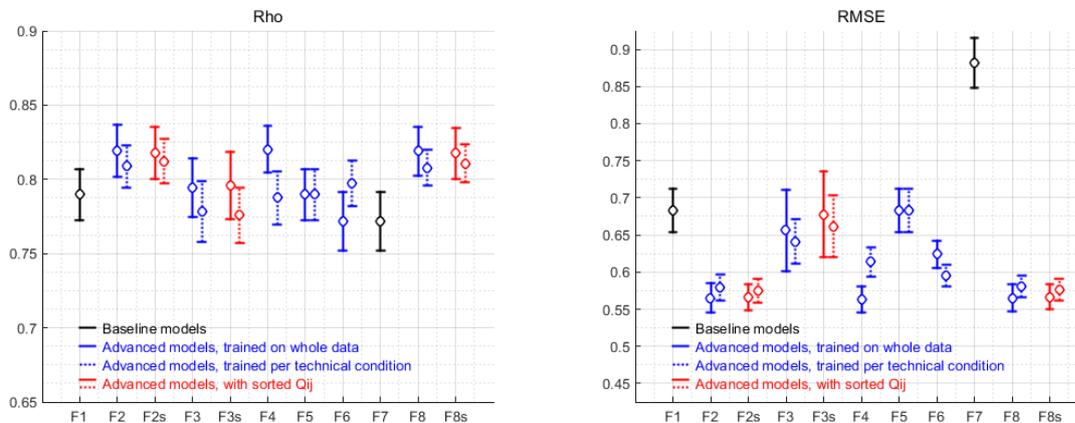
Based on detailed analysis in Figures H.39 to H.43, for A & B using ANOVAs & PostHoc tests, for C using t-Tests.

Effects	Condition(s)
A) No advanced model significantly outperforms the baseline mean F1.	All (AVRef, VF1, VR, FM2, E2, FM2-VF1, E2-VF1, FM2-VR, E2-VR)
B) Equal performance for sorting-based and rule-based assignment of $Q_{ij}$ .	All
C.1) Training per condition destabilizes performance (large confidence intervals) for F3 & F3s.	VR, E2-VR
C.2) Training per condition destabilizes performance (large confidence intervals) for F4.	FM2-VF1
C.3) Training per condition reduces performance for F3s.	FM2
C.4) Training per condition reduces performance for F3 & F3s.	E2-VF1
C.5) Training per condition does not improve performance.	AVRef, VF1, E2, FM2-VR

Figure 9.16: Modeling performance across all technical conditions of Experiment AVCT1.

## 1. Results across technical conditions

### 1.1 Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



### 1.2 Observations and Conclusions:

Based on detailed analysis in Figure H.44, for A & B using ANOVAs & PostHoc tests, for C using t-Tests.

- Six advanced models (F2, F2s, F4, F8 & F8s for both training modes and F6 trained per condition) outperform the baseline mean F1.
- Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- Training per condition improves performance for F6 and reduces performance for F4.

## 2. Results per technical condition:

Based on detailed analysis in Figures H.45 to H.52, for A & B using ANOVAs & PostHoc tests, for C using t-Tests.

Effects	Condition(s)
A.1) No advanced model significantly outperforms the baseline mean F1.	AVRef, VF2sym, E2, FM4sym, FM4P3sym, FM4-VF2, FM4-VB
A.2) No advanced model significantly outperforms the baseline mean F1. F2, F2s, F4, F6, F8 & F8s are slightly better (not sig.) than F1.	VF2
A.3) No advanced model significantly outperforms the baseline mean F1. F2, F2s, F4, F8 & F8s are slightly better (not sig.) than F1.	FM4P3
A.4) F8 significantly outperforms the baseline mean F1, but not the baseline minimum F7. F2, F2s, F4, F8 & F8s are slightly better (not sig.) than F1, but not F7.	VBsym
A.5) F2, F2s, F3, F4, F6, F8 & F8s significantly outperform the baseline mean F1 and minimum F7.	VB
A.6) F2, F2s, F4, F6, F8 & F8s significantly outperform the baseline mean F1 and minimum F7.	FM4-AVP3
A.7) F2s significantly outperforms the baseline mean F1 and minimum F7. F2, F4, F8 & F8s are slightly better (not sig.) than F1 and F7.	FM4
A.8) F2, F2s, F4, F8 & F8s (all trained on whole data) significantly outperform the baseline mean F1 and minimum F7.	FM4P3-VB
A.9) F2, F2s, F4, F8 & F8s (all trained per condition) and F6 significantly outperform the baseline mean F1, but not the minimum F7.	VB
B) Equal performance for sorting-based and rule-based assignment of $Q_{ij}$ .	All
C.1) Training per condition destabilizes performance (large confidence intervals) for F3 & F3s.	FM4-VF2
C.2) Training per condition destabilizes performance (large confidence intervals) for F4.	FM4P3-VB
C.3) Training per condition stabilizes performance (small confidence intervals) for F3.	FM4P3-VF2
C.4) Training per condition destabilizes performance for F3s and stabilizes performance for F3.	FM4
C.5) Training per condition destabilizes Rho but stabilizes RMSE for F3s, improves performance for F6, and reduces performance for F4.	FM4-VB
C.6) Training per condition does reduces performance for F4.	VB
C.7) Training per condition improves performance for F3.	AVRef
C.8) Training per condition improves performance for F3s.	FM4P3
C.9) Training per condition improves performance for F6.	VBsym
C.10) Training per condition does not improve performance.	VF2, VF2sym, E2, FM4sym, FM4P3sym, FM4AVP3

Figure 9.17: Modeling performance across all technical conditions of Experiment AVCT2.

## 9.5 Discussion

### 9.5.1 Review of Results

*Model performance* With correlation coefficients up to 0.95 and root mean square errors (RMSE) down to 0.45 on the five-point absolute category rating scale, the author concludes that it is possible to adequately model the Single-Perspective Telemeeting Quality  $Q_i$  based on the Individual Connection Quality scores  $Q_{ij}$ . This conclusion is based on the fact that the models aim at predicting the raw quality scores and not Mean Opinion Scores, meaning that the target values for prediction are noisy to a certain extent. More precisely, the confidence intervals for  $Q_i$  in Figures 9.1 to 9.3 are for most conditions in the order of 0.3 to 0.7 points on that five-point scale; and the achieved model RMSE values are in that same range. Concerning future work, however, in which such models would be applied for the prediction of Mean Opinion Scores, smaller RMSE values are necessary; especially when considering that a deviation of 0.5 points on the MOS scale is in practice considered as a substantial difference.

*Modeling Hypothesis* Table 9.2 provides an overview that shows for which conditions at least one advanced model could significantly outperform the baseline mean and/or the baseline minimum.

The results confirm the modeling hypothesis since there indeed exists a set of advanced models that outperform the simple mean operation. In addition, the added value of those advanced models indeed holds only for a subset of conditions, since the added value was found for only a limited number of conditions and in only four of the seven considered experiments: ACT<sub>3</sub>, LOT<sub>1</sub>, LOT<sub>2</sub> & AVCT<sub>2</sub>.

Furthermore, the errorbar visualization of the results showed that in several instances, advanced models performed better than the baseline models (confidence intervals were just not overlapping), but missed the significance thresholds when conducting the PostHoc tests. Those cases are also included in Table 9.2. This accounts for the possibility that those cases could have been significant if the settings of the performance training and evaluation procedure, i.e. number of bootstrap repetitions and training-test ratio, would have been chosen differently (see Section 9.3).

*Modeling factor: Training mode* From the t-Tests comparing the two training modes per condition (results C in Figures 9.11 to 9.17) it can be concluded that the added value of training per technical condition is not consistent. The reason is that across technical conditions, training per condition has for most functions no significant effect. Exceptions are: Function F<sub>4</sub>, which is negatively affected in Experiment ACT<sub>1</sub>, ACT<sub>2</sub>, LOT<sub>1</sub>, AVCT<sub>2</sub>, and in tendency also in LOT<sub>2</sub>; Function F<sub>3s</sub>, which is negatively affected in Experiment LOT<sub>1</sub> and AVCT<sub>1</sub>; and Function F<sub>6</sub>, which is positively affected in Experiment ACT<sub>3</sub> and AVCT<sub>2</sub>. And also looking at the t-Test results per technical

Exp.	Conditions	Function (Training Mode)						sig. outperform			
		(b)=both, (c)=per condition, (w) = whole data						F1	F7		
ACT3	across all	F2(w), F2s(b),		F4(b),	F6(c),	F8(b),	F8s(w)	✓	✓		
	P4	F2(w)						×	×		
	TN1		F3(c),		F6(c),	F8(c)		✓	×		
	TN2		F3(c), F3s(c),		F6(c),	F8(c)		✓	×		
	E2		F2s(b), F3(b),	F3s(b),	F4(b),	F6(b),	F8(b),	F8s(b)	✓	×	
	E2		F2(b), F2s(b),	F3(b),	F3s(b),	F4(b),	F6(b),	F8(b),	F8s(b)	✓	×
LOT1	across all		F2s(b),		F4(b),		F8s(b)	✓	✓		
	R	F2(b),	F2s(b),	F3(b),	F3s(b),	F4(b),	F6(b),	F8(b),	F8s(b)	✓	✓
LOT2	across all	F2(c),	F2s(b),	F3(b),	F3s(b),	F4(w),	F8(c),	F8s(b)	✓	✓	
	FM4P3		F2s(b),	F3(c),	F3s(b),	F4(w),	F6(w),	F8s(b)	✓	✓	
				F3(w),		F4(c),	F6(c),	F8(b)	✓	×	
	E2	F2(b),	F2s(b),	F3(b),	F3s(c),	F4(b),	F8(b),	F8s(b)	✓	✓	
	P3			F3(w),	F3s(w),	F4(w)			✓	✓	
		F2(b),	F2s(b),	F3(c),	F3s(c),	F4(c),	F6(b),	F8(b),	F8s(b)	✓	×
	P4	F2(b),	F2s(b),	F3(b),	F3s(b),	F4(b),	F6(b),	F8(b),	F8s(b)	✓	✓
P4sym				F3s(w),	F4(w)			✓	×		
AVCT2	across all	F2(b),	F2s(b),		F4(b),	F6(c),	F8(b),	F8s(b)	✓	✓	
	VF2	F2(b),	F2s(b),		F4(b),	F6(b),	F8(b),	F8s(b)	×	×	
	FM4P3	F2(b),	F2s(b),		F4(b),		F8(b),	F8s(b)	×	×	
	VBsym						F8(b)		✓	×	
		F2(b),	F2s(b),		F4(b),		F8(b),	F8s(b)	×	×	
	VB	F2(b),	F2s(b),	F3(b),	F4(b),	F6(b),	F8(b),	F8s(b)	✓	✓	
	FM4-AVP3	F2(b),	F2s(b),		F4(b),		F8(b),	F8s(b)	✓	✓	
	FM4	F2(b),			F4(b),		F8(b),	F8s(b)	×	×	
	FM4P3-VB	F2(w),	F2s(w),		F4(w),		F8(w),	F8s(w)	✓	✓	
	VB	F2(c),	F2s(c),		F4(c),	F6(b),	F8(c),	F8s(c)	✓	×	

Table 9.2: Data sets in which advanced modeling functions outperform, either significantly or in tendency (i.e. not sig.), the baseline models.

condition, the effect of training per condition has for the vast majority of combinations of function and condition no significant effect either. The few exceptions, which are given in Table 9.3, show in addition a mix of positive and negative effects.

While these results do not justify the added value of a training per condition, a training per condition can still be useful for tuning individual modeling functions. Referring to the results concerning the added value of the advanced models, i.e. Table 9.2, the slight performance differences caused by the condition-specific training can push the performance of an advanced model beyond the significance threshold, which was often the case for Function F6.

*Modeling factor: Assignment of Weights to Q<sub>ij</sub> scores* The results B in Figures 9.11 to 9.17 show that the performance of the two assignment strategies are essentially equal in almost all cases. Only in one case, condition FM4P3 in Experiment LOT2, the sorting-based assignment of Q<sub>ij</sub> outperforms rule-based assignment for Function F2 (i.e. F2s

Effect	No. of Conditions in Experiment						
	ACT1	ACT2	ACT3	LOT1	LOT2	AVCT1	AVCT2
F4 ↓	1	1		1	2	1	3
F3s ↓						4	2
↑		1		2			1
F3 ↓		1		1	3		2
↑		2		2			2
F6 ↑		3		2			1
F2 ↑		1		2			
F8 ↑		3		2			
F8s ↑				1			

Table 9.3: Effect of training mode on individual conditions.

better than F2). From an engineering perspective, this is a very positive result. It means that a sophisticated pre-analysis of the individual connections, which is required to apply the rule-based assignment, is not – at least for the present data – necessary, as it can be replaced with the sorting-based assignment strategy, which is technically easy to implement.

*Recommendation for an Optimal Model* The data confirmed the hypothesis that advanced models can outperform a simple mean, while this holds – as expected – for a limited number of cases. Furthermore, there is not one single model that shows consistently the best performance. This leads to the unsatisfactory situation that a practitioner would need to try out the different models for each application scenario. However, if resources do not allow a practitioner to do this, a recommendation for a single model would be helpful.

For that purpose, the author recommends Modeling Function F4 “quality dependent influence”, for two reasons: first, F4 is often the best or at least among the best performing models, although the differences to other functions are not significant; second, and most importantly, F4 essentially represents a compromise between the two baseline functions F1 and F7, which in turn were found to mark the upper and lower boundaries for the target values  $Q_i$  (see Figures 9.1 to 9.3) in the most cases. Concerning the training mode, F4 should not be trained on individual conditions, as this either reduced the performance of F4 (larger mean values of RMSE) or even destabilized the performance (larger confidence intervals of RMSE and RHO).

### 9.5.2 *Review of the Approach and Open Questions for Future Work*

There are a number of open questions that this section will address now: the aspect of predicting Mean Opinion Scores (MOS) per condition instead of raw scores, the impact of the training and test procedure, and the aspect of using closed-form and rule-based modeling approaches.

*Prediction of MOS per condition instead of raw scores* In the present work, the models attempted to predict the raw quality scores of all individual observations. An alternative that is usually applied in quality modeling is not to predict such raw scores but Mean Opinion Scores (MOS) computed per technical condition. The main advantage of such a modeling approach is its potential to stabilize and improve results, as the averaging of raw scores per condition reduces the noise caused by individual differences between test participants in terms of their rating behavior. The disadvantage is the danger of overtraining effects leading to too optimistic average results, which can be accompanied by increased performance variation around that average performance.

Such overtraining effects occur if the amount of data is small with respect to the number of model parameters to be trained. The mini-

imum requirement for any reasonable modeling is that the number of data points available for training must not be smaller than the amount of model parameters. Otherwise the task of the parameter fitting procedure to determine the values of the model parameters is ill-defined. Similarly, if the amount of data is equal or only slightly larger than the number of parameters, then the model parameters might very well reflect the actual data points and show excellent performance for test data that is similar to the training data available, causing an optimistic average performance. However, if the test data consists of data points that are dissimilar to the training data, then the performance of the model will break down, causing an increased variation around the optimistic performance.

Coming back to the idea of computing MOS per condition, the present data sets would be reduced in most cases to quite low numbers of data points, see Table 9.4. Given that the degree of freedom (number of free model parameters) ranges from 2 for Function F<sub>2</sub> to 6 for Functions F<sub>3</sub> and F<sub>4</sub>, the number of available data points per experiment is small enough to cause overtraining effects.

Future work could test the performance of the modeling functions proposed here for MOS by conducting large scale studies covering many more conditions to avoid the danger of overtraining. However, those studies need to be distributed across individual experiments, in order to limit effort for test participants, which in turn requires techniques that allow a combination of data across experiments. One possible approach is presented in Garcia<sup>7</sup>, which uses a number of anchor conditions throughout all experiments. In this approach, the results are transformed per experiment such that the anchor conditions are equal to the quality scores obtained in a so-called reference test, which in turn was specifically designed to cover the whole expected range of impairments. Another approach is presented in Pinson & Wolf<sup>8</sup>, which also uses a number of anchor conditions throughout all experiments. In this approach, however, the results are not mapped to a reference test, but to a grand mean that is computed from the anchor conditions across all experiments.

*Impact of the training and evaluation procedure* The discussions in Section 9.3 showed that differences in the results can be expected if another training and evaluation procedure is used. Obviously, one possible direction of future work is to evaluate such procedure-dependent effects, for instance by using a different repetition method than the modified bootstrap approach used here, or by using a different number of iterations, or by using a different ratio of training and test data.

On the one hand, following this direction of optimizing the training and test procedure might help in increasing the sensitivity to better differentiate between the individual modeling functions. On the other hand, it might be good to decide on the amount of effort to be spent on the optimization of the procedure after more knowledge on the relation between technical conditions and the performance

Experiment	No. conditions
ACT <sub>1</sub>	7
ACT <sub>2</sub>	4
ACT <sub>3</sub>	16
LOT <sub>1</sub>	9
LOT <sub>2</sub>	24
AVCT <sub>1</sub>	9
AVCT <sub>2</sub>	15

Table 9.4: Number of conditions for the seven experiments ACT<sub>1</sub> to AVCT<sub>2</sub>.

<sup>7</sup> Marie-Neige Garcia. *Parametric Packet-based Audiovisual Quality model for IPTV services*. Springer, 2014

<sup>8</sup> Margaret Pinson and Stephen Wolf. *Techniques for evaluating objective video quality models using overlapping subjective data sets*. Technical Report TR-09-457. US Department of Commerce, National Telecommunications and Information Administration (NTIA), 2008

differences is available. This is motivated by the observation that the results of the present and the previous chapter show strong but complex dependencies of the technical conditions, suggesting that those dependencies might impact the results more than any optimizations of the training and test procedure.

*Closed-form and rule-based modeling approaches* The individual modeling functions investigated can be considered as closed-form modeling approaches, as each function uses one single closed-form mathematical operation. However, this approach might not be optimal, given the condition dependency of the model performance, i.e. different modeling functions show optimal performance for different conditions, although performance differences are often small or even marginal. For that reason, future work may investigate rule-based approaches, in which a set of rules may be used for predicting  $Q_i$  based on  $Q_{ij}$ . Such approaches could for instance use such rules to choose the optimum closed-form modeling functions for different technical conditions or for different other aspects that characterize telemeetings such as usage scenario, number of interlocutors, or communication modality. Alternatively, *truly* rule-based approaches could predict  $Q_i$  from tables with stored values, whereas those values are chosen by a set of rules that analyze the actual combination of the  $Q_{ij}$  ratings. There are approaches to automatically generate such if-then rules from data; the introduction by Chiu<sup>9</sup> could serve as a starting point for future work in that direction.

### Summary

The goal of the research presented here was to develop and evaluate a model that estimates the perceptual ratings of *Single-Perspective Quality*  $Q_i$  based on the perceptual ratings of the *Individual Connection Qualities*  $Q_{ij}$ . The data sets stem from the seven experiments of Chapter 8: ACT<sub>1</sub>, ACT<sub>2</sub>, ACT<sub>3</sub>, LOT<sub>1</sub>, LOT<sub>2</sub>, AVCT<sub>1</sub>, AVCT<sub>2</sub>.

Motivated by the results of Chapter 8 the modeling work was driven by the following underlying hypothesis: There exists a set of advanced models that outperform the simple mean operation, whereas the added value of those advanced models holds only for a subset of conditions.

Furthermore, revisiting the perceptual data sets revealed three different behaviors, leading to three fundamental modeling ideas:

1. The *Single-Perspective Telemeeting Quality*  $Q_i$  is a simple mean of the *Individual Connection Qualities*  $Q_{ij}$ ;
2. the *Single-Perspective Telemeeting Quality*  $Q_i$  is the minimum value of the *Individual Connection Qualities*  $Q_{ij}$ ;
3. the *Single-Perspective Telemeeting Quality Impairment*  $QI_i$  is the sum of the *Individual Connection Quality Impairments*  $QI_{ij}$ , whereas  $QI_i$  can be expressed as the difference between a maximum possible

<sup>9</sup> Stephen Chiu. "Extracting fuzzy rules from data for function approximation and pattern classification". In: *Fuzzy Information Engineering: A Guided Tour of Applications*. Ed. by Didier Dubois, Henri Prade, and Ronald R. Yager. John Wiley&Sons, 1997

quality  $Q_{max}$  and the *Single-Perspective Telemeeting Quality*  $Q_i$  and  $Q_{ij}$  as the difference between  $Q_{max}$  and the *Individual Connection Quality*  $Q_{ij}$ .

Based on those basic ideas, eight modeling functions were developed and tested: F1 “baseline mean”, F2 “weighted mean”, F3 “mutual influence”, F4 “quality dependent influence”, F5 “influence of outliers”, F6 “weighted minimum”, F7 “baseline minimum”, and F8 “weighted mean of impairments”. Here, Functions F1 and F7 were considered as baseline functions as they do not contain any parameters that can be fitted, and the remaining Functions F2 to F6 and F8 were considered as advanced modeling functions as they contained parameters that can be fitted.

Furthermore, a number of algorithmic variants were investigated as well: Concerning a first set of variants, Functions F2, F3, and F8 essentially represented weighted sums of  $Q_{ij}$  and required a method to decide which  $Q_{ij}$  are assigned to which weights. Here two methods were tested, a rule-based assignment based on the methodology in Chapter 6, and a ranking-based assignment exploiting a sorting of the  $Q_{ij}$  values. Concerning a second set of variants, all modeling functions were trained and tested either across the whole data set for each of the seven experiments, or per technical condition.

The used model training and evaluation procedure stems from the machine learning domain and is based on the bootstrap method; a method that is known to work well for small data sets, which is the case here. The method applied here is a modification of the bootstrap method, in which the available data is randomly split into training and test data (here: 80% training, 20% test), the training data is used for fitting the model parameters, and the test data to evaluate the model performance. This is done for a number of repetitions (here: 20), and the performance of each iteration is stored. Two performance measures were used: the Pearson correlation coefficient  $Rho$  and the root mean square error  $RMSE$  between estimated and true values of  $Q_i$ . Since  $Rho$  and  $RMSE$  were computed per iteration, the results were aggregated across the iterations by computing simple statistics: mean and 95% confidence intervals.

The following results were found: First, the underlying modeling hypothesis is confirmed, as the advanced modeling functions outperform the baseline modeling functions for a limited number of data sets and conditions.

Second, no single winning function could be identified, as different functions showed optimal performance in different cases.

Third, nevertheless, modeling function F4 “quality dependent influence” can be recommended as the best compromise, as it was often the best or among the best functions and it combines the two fundamental trends that  $Q_i$  is either close to the mean or to the minimum value of  $Q_{ij}$ . Function F4 is computed as the weighted combination of the mean and the minimum value of  $Q_{ij}$ , whereas this weighting depends on a pre-estimation of  $Q_i$  based again on the mean of  $Q_{ij}$ .

Fourth, the method of assigning the weights had hardly any impact,

which means for real-life applications that a complicated rule-based assignment is not necessary and that the easy-to-implement sorting-based assignment is sufficient.

Fifth, training per technical condition had hardly any impact, which means for real-life applications that a complicated estimation of the technical conditions for optimal training is not justified.

Sixth, the effort of improving the experimental methods between the different perception tests paid off in the sense that the advantage of advanced modeling functions was increased. More precisely, the number of advanced modeling functions that outperformed the baseline functions was higher for experiments with an improved method (ACT<sub>3</sub>, LOT<sub>2</sub>, AVCT<sub>2</sub>) compared to the corresponding experiment with the non-optimal method (ACT<sub>1</sub>, ACT<sub>2</sub>, LOT<sub>1</sub>, AVCT<sub>1</sub>).

## **Part IV**

# **Summary and Conclusions**

## 10

# Summary and Conclusions

### *What this chapter is about*

This chapter reflects on the individual research objectives in light of the results presented in the previous chapters; and it concludes with a high-level discussion of the main findings.

### *10.1 Research Objectives Revisited*

This section describes how the results presented in Chapters 5 to 9 met the individual research objectives that have been defined in Section 1.3. Due to the alignment of individual research objectives and chapters, the main results of each chapter are briefly restated and put into relation to the corresponding originally formulated objective. Furthermore, a critical review of the achieved results and proposals for future work per objective conclude each discussion.

#### *10.1.1 Conceptual Model of Telemeeting Quality – Chapter 5*

*Objective Restated and Results Achieved* The objective was to derive an overview of aspects and processes that may play a role when test participants form a quality judgment of telemeeting systems. More specifically, the model should summarize the known fundamentals of telemeeting quality perception and should help to identify open issues and research questions for future research. Furthermore, the model should serve as a tool to put individual research activities on telemeeting quality assessment into a broader context which would help to interpret results and draw conclusions properly.

The developed conceptual model describes quality perception as a multi-stage process, in which the individual stages represent different perceptual and cognitive processes from sensory input to the description of the perceived quality. In this model, most aspects have been introduced in prior work and reflect the fundamental perception process perspective of quality by Jekosch, here rephrased for telemeetings: *telemeeting quality* is the result of a comparison between the perceived and desired character of a telemeeting.

Starting from this, the model proposes a number of new model stages that reflect the two main aspects investigated throughout the

present thesis: the aspect of the special communicative situation, i.e. to have a group conversation over a telecommunication system; and the aspect of the possibility to perceive asymmetric conditions, i.e. individual interlocutors are connected with different equipment and connection characteristics. To this aim, the model introduced on the one hand the concepts of a *Telecommunication* and *Group-Communication Component*, and on the other hand, the concepts of different *Quality Aggregation Levels* differentiating between individual connections and the whole telemeeting.

With these concepts and the corresponding processing stages, the conceptual model indeed provides an overview of the relevant aspects and processes. In addition, this model also helps to identify open questions for future work in two directions: the first direction is to further investigate the validity and practical relevance of the model's processing stages; the second direction is to dive into the details of the presented concepts and to investigate various individual research questions that are relevant in practice. One example is to investigate the impact of the experimental methodology (test instructions, questionnaire, test design, etc.) on the perception of the two components; another example is to investigate the relation between the different quality aggregation levels.

Furthermore, the model puts individual research activities into a broader context, as it allows to decouple the interaction between technical and communicative aspects of quality. Using the provided concepts and nomenclature, researchers can be very specific when conducting, analyzing and interpreting telemeeting quality experiments: Will test participants actually judge an individual connection that is particularly different from the others or will they judge the whole system? Will test participants actually focus on technical aspects, i.e. audio and video signal quality, or will they judge how the system facilitates group-communication?

*Limitations and Future Work* A first limitation of the model is its already mentioned conceptual nature. Here, more insights on the practical relevance beyond the scope of the present text are needed. A second limitation is that the model currently excludes one important aspect of the user's satisfaction with today's telemeeting systems: the usability and here especially the call setup phase. Here, the possible complexity and often occurring problems during call setup can substantially contribute to the perceived quality of a system. Hence future work should specifically investigate this aspect and correspondingly extend the conceptual model.

#### 10.1.2 *Methodology for the Perceptual Assessment of Telemeeting Quality – Chapter 6*

*Objective Restated and Results Achieved* The objective was to develop a set of multiparty specific test methods for conducting quality experiments with test participants and synthesize them into a set of

recommendations that can be used by practitioners in the field.

One major part of this objective has been addressed with the development of the ITU-T Recommendation P.1301. While the author substantially contributed to this document by a series of contributions (see Appendix A), the recommendation was a joint effort of various experts from different laboratories. In that respect, the objective was met in the sense that the present research activities contributed to this recommendation, and the present text provides a guideline on how to use this recommendation.

Furthermore, given that Recommendation P.1301 has some limitations when it comes to detailed aspects (for example treatment and analysis of asymmetric conditions), another part of the present objective was to extend the methodology beyond the recommendation. Here, the present text provides a detailed guideline on how to efficiently design and conduct quality assessment tests with asymmetric conditions and how to systematically structure the data for proper analysis.

*Limitations and Future Work* A first limitation concerning the experimental method can be derived by revisiting the results of the perception and modeling experiments. Here, the comparisons between the individual experiments with their slightly modified experimental methods suggest that experimental details on a rather fine level appear to have a strong impact on the results. Related to this, a second limitation is that the present results do not allow to specify the impact of individual experimental details, as the experiments usually differed in a number of aspects simultaneously. That means, future work is necessary to obtain more empirical results in order to further verify the experimental methods presented here, and – ideally – to identify which individual methodological aspects are really crucial and which can be chosen more freely.

### 10.1.3 *Perception of Telemeeting Quality with focus on Group-Communication Aspects – Chapter 7*

*Objective Restated and Results Achieved* The objective was to investigate the impact of a number of communicative aspects on multiparty telemeeting quality. More specifically, empirical results should provide knowledge on the relation between relevant communication aspects and the perceived quality.

In one study the author investigated the impact of *Communication Complexity* in interaction with the *Technical System Capabilities* on three aspects of quality perception: *Quality of Experience*, representing the overall experience; *Speech Communication Quality* focusing on the *Telecommunication Component*; and *Cognitive Load*, representing the *Group-Communication Component*. The first finding of that study was that *Technical System Capability* showed a significant impact on *Cognitive Load*. This means, a technically more advanced system can reduce cognitive load under certain conditions, which was in

this study the presence of spatial audio sound reproduction. The second and more important finding was that both *Communication Complexity* and *Technical System Capability* showed an effect on *Quality of Experience*. This means, not only a technically more advanced system, but also a less complex communicative situation can improve quality perception.

In a second study the author investigated the impact of *Involvement* on three slightly different aspects of quality perception: *Quality of Experience*, *Cognitive Load*, and *Speech Communication Quality*, this time focusing on the *Group-Communication Component*. The first finding was that the impact of *Involvement* on *Cognitive Load* has an unexpected direction: test participants appear to perceive actually a lower *Cognitive Load* when they feel more involved. One explanation could be that feeling more involved makes the assessment task more interesting for test participants, which in turn means that their need to force themselves to stay concentrated throughout the test is reduced. The second and more important finding was that *Involvement* has an impact on *Quality of Experience*, which was visible when comparing the two fundamental test paradigms listening-only and conversation tests. While the impact of the two test paradigms on quality ratings is known, this study helps to better explain this effect in terms of the *Involvement* of test participants. Furthermore, the choice of either listening-only or conversation tests should be taken with great care given its impact on results. This is even more emphasized by another important finding, which was that the impact of the test paradigm (listening-only vs. conversation test) can be stronger than the impact of the technical conditions under test (here packet loss), even if those condition differences are strong enough to reach statistical significance.

*Limitations and Future Work* A first limitation of the study on *Communication Complexity* is that it followed a listening-only test paradigm. Here, future work is necessary to evaluate if the found effects can be reproduced with a conversation test paradigm, as conversation tests are known to be closer to real communication scenarios. Another limitation of this study was that the tested technical conditions are quite idealized compared to real life. Spatial sound reproduction and full-band channels are not yet fully established in the market and existing solutions face various impairments not tested here. Here, future work should verify the found effects in test contexts that are closer to real-life scenarios. A limitation of the study on *Involvement* is that it focused on one type of degradation, i.e. packet loss. Here, future work should verify the found effects for other degradations as well.

#### 10.1.4 Perception of Telemeeting Quality with focus on Telecommunication Aspects – Chapter 8

*Objective Restated and Results Achieved* The objective is to investigate the impact of various technical system characteristics on multi-party telemeeting quality. More specifically, empirical results should provide knowledge on the quality relation between the individual connections and the whole telemeeting.

The results of a series of seven experiments confirmed two hypotheses concerning the *Single-Perspective Telemeeting Quality*, referring to the whole telemeeting, and the *Individual Connection Quality*, referring to the individual connections. The first hypothesis was that there exists a mutual influence of the individual connections on quality perception, whereas this influence holds only for a limited number of technical conditions. The second hypothesis was that for most conditions the *Single-Perspective Telemeeting Quality* is an average (arithmetic mean) of the *Individual Connection Quality* scores, whereas there is a limited number of technical conditions for which such a simple mean is not sufficient.

Interestingly, these results were found for all three different test scenarios that were applied across the seven experiments, i.e. audio-only conversation test, listening-only test, and audiovisual conversation test. Differences between those test scenarios, however, could be found when it comes to the potential explanation on whether a technical condition confirms or rejects the two hypotheses. For the audio-only conversation tests, the strength of an impairment appears to contribute to the found effects. For the listening-only tests, the strength of an impairment and, in addition, the presence of asymmetry appear to contribute to the found effects. For the audiovisual conversation tests, impairment strength and asymmetry are not sufficient to explain the found effects; here the modality of the impairment appears to contribute as well.

In summary, the mutual relation between *Individual Connection Quality* scores and their relation to the *Single-Perspective Quality* are a rather complex construct. Looking across the different test paradigms, it appears to depend – at least partially – on the strength of impairments, the presence of asymmetry, and the modality of impairments, whereas no individual aspect alone is sufficient to explain the relations consistently.

*Limitations and Future Work* A first limitation of the studies concerns the experimental methodology, more specifically the fact of asking for both *Single-Perspective Telemeeting Quality*  $Q_i$  and *Individual Connection Quality*  $Q_{ij}$  on the same questionnaire. Here, it might be that test participants form some average of the *Individual Connection Quality* scores simply due to the presentation of both  $Q_i$  and  $Q_{ij}$  on the same questionnaire, despite the usage of different scale designs and questions. Here, future studies should validate this by different experimental approaches.

A second limitation concerns the communicative situation. Despite the simulation of three different test scenarios, i.e. audio-only communication, passive listening, and audiovisual communication, the number of interlocutors was kept constant at a value of three, and the number of conversation scenario types was one in the audio-only conversation tests, and listening-only tests and two in the audiovisual tests. Here, future work would be needed to further validate the found results for a different number of interlocutors and different conversation scenario types.

#### 10.1.5 *Perceptual Models of Telemeeting Quality – Chapter 9*

*Objective Restated and Results Achieved* The objective was to develop a set of perceptual models which predict how test participants form a quality judgment on individual aspects that they perceive. In addition, the model should provide tangible results beyond the conceptual work done beforehand, and it should provide empirical evidence for the practical relevance of the conceptual model. Furthermore, by showing positive performance, the model should serve as indication for the validity of the experimental test methods, which are used to obtain the data for modeling.

The developed computational models computed an estimation of the *Single-Perspective Telemeeting Quality* scores based on the *Individual Connection Quality* scores. The models showed promising performance in terms of the Pearson correlation coefficient and the root mean square error between predicted and true *Single-Perspective Telemeeting Quality* scores. However, these results strongly depend on the individual experiments, and even individual conditions and modeling functions, as the correlation values were in a range of approximately 0.7 to 0.95 and root mean square errors in a range of approximately 0.45 to 0.7 on a five-point scale. Despite such dependencies, one modeling function (F4 “quality dependent influence”) appeared to provide the best compromise: This function is computed as the weighted combination of the mean and the minimum value of the *Individual Connection Quality* scores, whereas this weighting depends on a pre-estimation of the *Single-Perspective Telemeeting Quality* score, based again on the mean of the *Individual Connection Quality* scores.

Concerning the link to the conceptual model, the computational models show that it is possible to estimate, with a promising performance, the *Single-Perspective Telemeeting Quality* score based on the *Individual Connection Quality* scores. This supports the idea of having a multiparty decomposition stage in the conceptual model, which enables a person to consider the quality of individual connections, and a multiparty aggregation stage, which enables a person to combine the quality of individual connections.

Concerning the validity of the experimental methods, the modeling results strongly support the necessity of optimized experimental methods. Here, the effort of improving the experimental methods

between the different perception tests increased the advantage of the investigated advanced modeling functions compared to the investigated baseline modeling functions (mean and minimum value). More precisely, the number of advanced modeling functions that outperformed the baseline functions was higher for each experiment with an improved method (ACT<sub>3</sub>, LOT<sub>2</sub>, AVCT<sub>2</sub>) compared to the corresponding experiment with the non-optimized method (ACT<sub>1</sub>/ACT<sub>2</sub>, LOT<sub>1</sub>, AVCT<sub>1</sub>).

*Limitations and Future Work* A first limitation concerns the data characteristics used for modeling. Typical approaches in the quality modeling domain first compute per technical condition a Mean Opinion Score (MOS) across all observations of that condition, before those MOS values are fed into the quality model. The motivation is to decrease the noise inherent in the model input by averaging over the ratings of test participants. In the present research, however, the amount of data was rather small in terms of the number of conditions covered, despite its substantial size in terms of the number of observations per condition and the overall effort to run the complex tests. This means that the approach of computing MOS upfront was not reasonable in most cases, as the number of conditions in most experiments was in the order of the number of parameters for some models. This in turn would lead to too small data sets for proper model training. Here, future work should check the performance of MOS-based modeling, suggesting a two stage procedure. The first stage will be to work with those combinations of the present data sets and model functions, for which a MOS-based approach can be considered as reasonable. The second stage will be to conduct new experiments in which the number of covered conditions is sufficiently large for robust MOS-based modeling approaches.

A second limitation concerns the approach of applying closed-form mathematical operations to compute the *Single-Perspective Telemeeting Quality* scores. This approach appeared to be limited, given the mutual dependency on the data sets, the technical conditions in those sets, and the actual modeling functions used. Here, future work should investigate alternative rule-based approaches, whereas two directions are possible. The first direction is to manually define a set of rules based on the available knowledge gained in this or similar research projects. However, such an approach might still show similar limitations in the model performance, especially when considering the difficulties of explaining the results of the perception experiments in Chapter 8 by straight-forward aspects such as the type, strength or modality of impairments, or the presence of symmetric or asymmetric conditions. The second direction is to investigate data-driven rule-based approaches, which might improve model performance but which might also reduce the interpretability of the rules.

## 10.2 *Final Conclusions*

The overall research goal was to investigate the fundamentals of multiparty telemeeting quality and to provide a set of prototypical methods and models for the quality assessment and prediction of multiparty telemeeting systems. To this aim, the individual research objectives and corresponding chapters investigated telemeeting quality from different perspectives: The work on the conceptual model in Chapter 5 provides a theoretical framework for the fundamentals of multiparty telemeeting quality, and the experiments in Chapters 7 and 8 provide empirical results for a number of individual aspects of telemeeting quality and their mutual interaction. The work on the more general experimental method in Chapter 6 and the detailed improvements of the actual experiments in Chapters 7 and 8 provide a set of prototypical methods which showed good sensitivity to investigate individual aspects of multiparty telemeeting quality. The work on the development of perceptual models in Chapter 9 and in parts also the results in Chapter 8 provide a set of prototypical approaches to computationally estimate telemeeting quality based on individual aspects, here the impact of the individual connections.

Furthermore, the set of methods and models shall serve as a foundation for further research on the Quality of Experience (QoE) of multiparty telemeetings. To this aim, the present text discussed both promising results and limitations of the individual achievements and suggested directions for future work. Focusing on potential commonalities among those discussions, one general underlying conclusion can be formulated that holds independently of the actual directions that future studies would follow: The present research represents an example of an iterative approach in which – step by step – the highly complex construct of multiparty telemeeting quality has been decomposed as far as it was possible in the course of this research project. In that respect, the present research can indeed serve as a prototypical foundation for future work on the QoE of multiparty telemeetings or on similarly complex QoE research topics for novel telecommunication and multimedia systems.

Finally, the set of methods and models shall directly be applicable for practitioners in the field. Concerning the applicability of methods, the present text provided both general guidelines on the methodologies and detailed descriptions on the actually conducted experiments. This gives practitioners a starting point for conducting experimental studies on the perception of multiparty telemeeting quality. The present research also emphasizes that experimental details can influence results, meaning that the presented methods should not be applied to other test cases without prior validation.

Concerning the applicability of models, and in particular the computational algorithms, a distinction must be made between application in research scenarios and application in operating telemeeting services. Such a distinction is necessary as the computational algorithms represented perceptual models in the sense that they used perceptual

ratings on individual aspects of a telemeeting, here the individual connections, as input in order to estimate perceptual ratings of overall telemeeting quality. Such models can in principle directly be used in research scenarios for the purpose of further investigating the quality relationships between individual connections and the whole telemeeting. However, such models can – by definition – not directly be applied to quality modeling in operating telemeeting systems, as these models are not instrumental models that take technical or signal information as inputs. Instead, the perceptual models developed in this thesis can serve as the foundation for the development of instrumental models, in which the current input, the perceptual ratings, are replaced by technical or signal information.

As the very final remark, of all possible directions of future work, the most logical next step in the author's view is to develop a set of instrumental quality models which transfer the knowledge of the present research to the technical application. Later on, such models could be further improved by including the impact of the communication aspects, which would require the development of technical approaches to estimate aspects such as communication complexity or involvement.

## **Part V**

# **Appendices**

# A

## *Overview of Conducted Studies and Documents that Contributed to this Thesis*

### *What this chapter is about*

This appendix gives an overview on all documents (published, published with limited access, and unpublished), that report on intermediate results of the present research. To this aim, three tables show which document contributes to which chapter of the present text to which extent. The extent of each contribution is coded as either *Major*, *Aspects*, or *Background*. *Major* means, the document has a major and direct contribution to the chapter, or is highly relevant for the chapter, or substantial parts of the original texts are reused in the chapter; *Aspects* means, the document addresses individual or minor aspects of a chapter, or some minor parts of the original texts are reused in the chapter; and *Background* means, the document served as background knowledge for a chapter.

## A.1 Document Overview

Document	Chapter 5 Conceptual Model	Chapter 6 Methodology	Chapter 7 Perception Group- Communication Aspects	Chapter 8 Perception Telecommunica- tion Aspects	Chapter 9 Perceptual Models
Janto Skowronek et al. "Speech recordings for systematic assessment of multi-party conferencing". In: <i>Proceedings of Forum Acusticum 2011</i> . Aalborg, Denmark: European Acoustical Society, June 2011	—	Background	Aspects	—	—
Janto Skowronek and Alexander Raake. "Einfluss von Bandbreite und räumlicher Sprachwiedergabe auf die kognitive Anstrengung bei Telefonkonferenzen in Abhängigkeit von der Teilnehmeranzahl". In: <i>Fortschritte der Akustik (DAGA2011)</i> - 37. <i>Deutsche Jahrestagung für Akustik</i> . Deutsche Gesellschaft für Akustik. Düsseldorf, Germany, Mar. 2011, pp. 873-874	—	—	Aspects	—	—
Janto Skowronek and Alexander Raake. "Investigating the effect of number of interlocutors on the quality of experience for multiparty audio conferencing". In: <i>Proceedings of the 12th Annual Conference of the International Speech Communication Association (Inter-speech2011)</i> . International Speech Communication Association. Florence, Italy, Aug. 2011, pp. 829-832	—	—	Major	—	—
Janto Skowronek. "Internetumfrage zur Wahrnehmung heutiger Telefonkonferenzen im geschäftlichen Umfeld". In: <i>Fortschritte der Akustik (DAGA2012)</i> - 38. <i>Deutsche Jahrestagung für Akustik</i> . Darmstadt: Deutsche Gesellschaft für Akustik, 2012, pp. 899-900	Background	—	Background	Background	—
Janto Skowronek, Julian Herlinghaus, and Alexander Raake. "Quality Assessment of Asymmetric Multiparty Telephone Conferences: A Systematic Method from Technical Degradations to Perceived Impairments". In: <i>Proceedings of the 14th Annual Conference of the International Speech Communication Association (Inter-speech 2013)</i> . International Speech Communication Association. Lyon, France, Aug. 2013, pp. 2604-2608	—	Major	—	Major	—
Janto Skowronek, Katrin Schoenenberg, and Alexander Raake. "Experience with and insights about the new ITU-T standard on quality assessment of conferencing systems". In: <i>Proceedings of the international conference of acoustics (AIA-DAGA 2013)</i> . Merano, Mar. 2013, pp. 444-447	—	Major	—	—	—
Janto Skowronek, Falk Schiffner, and Alexander Raake. "On the influence of involvement on the quality of multiparty conferencing". In: <i>4th International Workshop on Perceptual Quality of Systems (PQS 2013)</i> . International Speech Communication Association. Vienna, Austria, Sept. 2013, pp. 141-146	—	—	Major	—	—
Janto Skowronek, Katrin Schoenenberg, and Gunilla Berndtsson. "Multimedia Conferencing and Telemeetings". In: <i>Quality of Experience - Advanced Concepts, Applications, Methods</i> . Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 217-228	Aspects	Background	—	—	—
Janto Skowronek and Alexander Raake. "Assessment of Cognitive Load, Speech Communication Quality and Quality of Experience for spatial and non-spatial audio conferencing calls". In: <i>Speech Communication</i> (2015), pp. 154-175. DOI: 10.1016	—	—	Major	—	—
Janto Skowronek, Anne Weigel, and Alexander Raake. "Quality of Multiparty Telephone Conferences from the Perspective of a Passive Listener". In: <i>Fortschritte der Akustik - 41. Jahrestagung für Akustik (DAGA)</i> . Nürnberg, Germany, Mar. 2015	—	—	—	Major	—
Janto Skowronek and Alexander Raake. "Conceptual model of multiparty conferencing and telemeeting quality". In: <i>Proceedings of the 7th International Workshop on Quality of Multimedia Experience (QoMEX 2015)</i> . Pilos, Greece, May 2015, pp. 1-6	Major	Aspects	—	—	—

Figure A.1: Overview of conducted studies and documents that contributed to this thesis. Part 1 – Publications.

Document	Chapter 5 Conceptual Model	Chapter 6 Methodology	Chapter 7 Perception Group-Communication Aspects	Chapter 8 Perception Telecommunication Aspects	Chapter 9 Perceptual Models
Janto Skowronek and Alexander Raake. <i>Development of scalable conversation test scenarios for multi-party telephone conferences</i> . ITU-T Contribution COM 12 - C187 - E. Geneva, Switzerland: International Telecommunication Union, Jan. 2011	—	—	Aspects	—	—
Janto Skowronek et al. <i>Initial insights for P.AMT based on reviewing existing recommendations</i> . ITU-T Contribution COM 12 - C 284 - E. Geneva, Switzerland: International Telecommunication Union, Oct. 2011	—	Background	Background	Background	—
Janto Skowronek. <i>Topics for Question 18/12 in view of insights from a field survey</i> . ITU-T Contribution COM 12 - C 285 - E. Geneva, Switzerland: International Telecommunication Union, Oct. 2011	Background	Background	—	—	—
Janto Skowronek and Alexander Raake. <i>Listening test stimuli using scalable multiparty conversation test scenarios</i> . ITU-T Contribution COM 12 - C 286 - E. Geneva, Switzerland: International Telecommunication Union, Oct. 2011	—	Background	Aspects	—	—
Janto Skowronek and Alexander Raake. <i>Number of interlocutors and QoE in multiparty telephone conferences</i> . ITU-T Contribution COM 12 - C 287 - E. Geneva, Switzerland: International Telecommunication Union, Oct. 2011	—	—	Aspects	—	—
Janto Skowronek et al. <i>Proposed main body and normative annex for draft recommendation P.AMT</i> . ITU-T Contribution COM 12 - C318 - E. Geneva, Switzerland: International Telecommunication Union, May 2012	Background	Major	Background	Background	—
Janto Skowronek et al. <i>Proposed non-normative appendices for draft recommendation P.AMT</i> . ITU-T Contribution COM 12 - C319 - E. Geneva, Switzerland: International Telecommunication Union, May 2012	Background	Major	Background	Background	—
Janto Skowronek and Alexander Raake. <i>Update on number of Interlocutors and QoE in multiparty telephone conferences</i> . ITU-T Contribution COM 12 - C035 - E. Geneva, Switzerland: International Telecommunication Union, Mar. 2013	—	—	Aspects	—	—
Janto Skowronek and Alexander Raake. <i>Report on a study on asymmetric multiparty telephone conferences for a future update of P.1301</i> . ITU-T Contribution COM 12 - C134 - E. Geneva, Switzerland: International Telecommunication Union, Dec. 2013	—	Major	—	Major	—
Janto Skowronek et al. "Method and Apparatus for Computing the Perceived Quality of a Multiparty Audio or Audiovisual Telecommunication Service or System". Patent application WO/2016/041593 (EP). Deutsche Telekom AG. Mar. 2016	Aspects	—	—	—	Major

Figure A.2: Overview of conducted studies and documents that contributed to this thesis. Part 2 – Documents with limited access: ITU Contributions and Patent Application.

Document	Chapter 5 Conceptual Model	Chapter 6 Methodology	Chapter 7 Perception Group- Communication Aspects	Chapter 8 Perception Telecommunica- tion Aspects	Chapter 9 Perceptual Models
Janto Skowronek. <i>Document summarizing model review</i> . Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, Apr. 2013	—	—	—	—	Background
Janto Skowronek. <i>Document describing scenarios</i> . Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, Apr. 2013	—	Major	—	—	—
Janto Skowronek. <i>Initial model for audio-only communication based on available data</i> . Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, June 2013	—	—	—	—	Aspects
Janto Skowronek. <i>Documentation on system setup</i> . Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, June 2013	—	—	—	Aspects	—
Janto Skowronek. <i>Improved model for audio-only communication</i> . Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, Dec. 2013	—	Aspects	—	Major	Major
Janto Skowronek. <i>Pilot test for audiovisual communication</i> . Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, Feb. 2014	—	—	—	Major	—
Janto Skowronek. <i>Initial model for audiovisual communication</i> . Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, May 2014	—	—	—	—	Major
Janto Skowronek. <i>Second test for audiovisual communication</i> . Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, Oct. 2014	—	—	—	Major	—
Janto Skowronek. <i>Improved model for audiovisual communication</i> . Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, Nov. 2014	—	Aspects	—	—	Major
Janto Skowronek. <i>Focus groups on quality perception and attribution</i> . Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, Nov. 2014	Background	—	—	—	—
Janto Skowronek. <i>Final Project Report – Quality of Multiparty Audio-Video Communication</i> . Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, Dec. 2014	—	Aspects	—	Major	Major

Figure A.3: Overview of conducted studies and documents that contributed to this thesis. Part 3 – Unpublished documents: Project Reports.

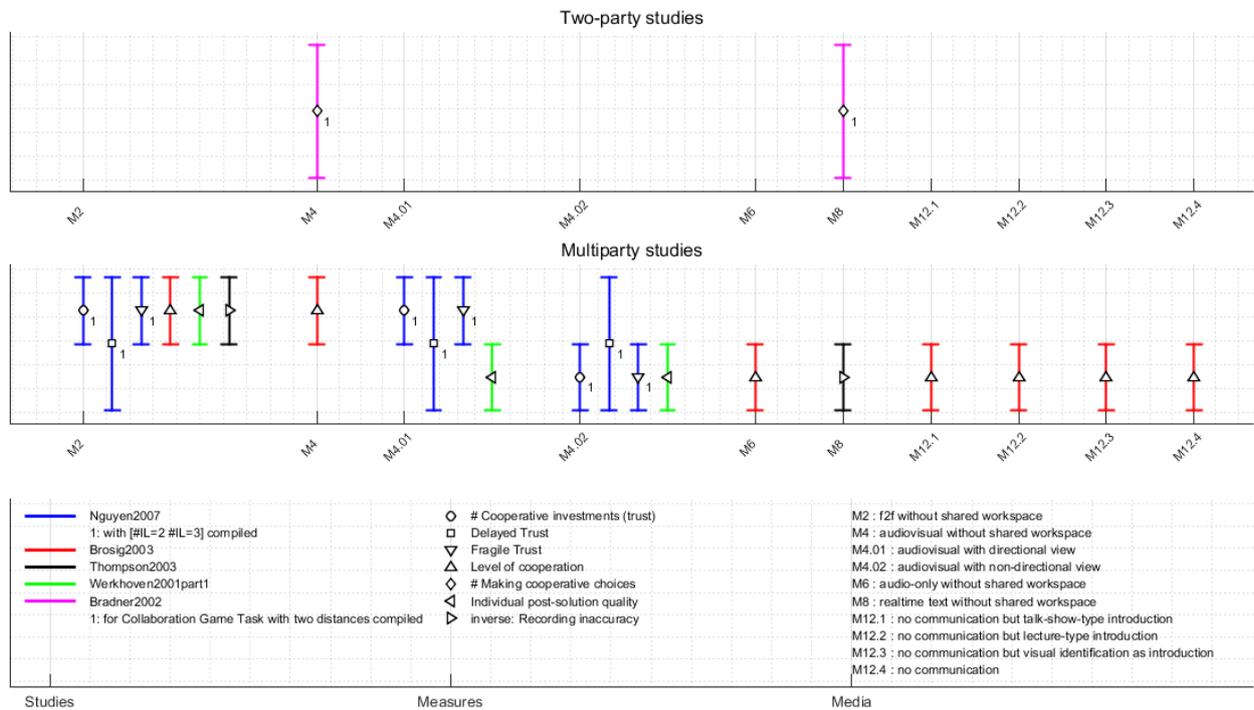
## *B*

# *Detailed Results: Deep Dive Analysis of Research on Group-Communication*

### *What this chapter is about*

The present appendix presents the detailed results of the literature review about research on group-communication. Each figure shows first a visualization of the effects of communication media on the different measures, using the graphical method described in Section 2.4. Then, each figure provides an explanation on how to interpret the visualization. Finally, each figure presents the main findings that can be extracted from the current visualization. The individual results are then compiled in the main text in Section 2.4.

### B.1 Analysis Results



**Interpretation Method:**

1. Look at the positions and sizes of the errorbars per medium  $M_x$ .
2. The more the errorbars for the medium  $M_x$  are clustered towards the upper end, the more likely  $M_x$  leads to an increased collaboration performance compared to other media. Example:  $M_2$  for the multiparty studies.
3. The more the errorbars are clustered toward the lower end, the more likely  $M_x$  leads to a decreased collaboration performance. Example:  $M_{4.02}$  for the multiparty studies.
4. The more centrally positioned and at the same time the smaller the errorbars are for the medium  $M_x$ , the more likely  $M_x$  has an average collaboration performance compared to other media. Example: none in this plot.
5. The more centrally positioned and at the same time the wider the errorbars are for the medium  $M_x$ , the more likely  $M_x$  does not differ in the collaboration performance compared to other media. Example: none in this plot.
6. The more the errorbars of individual studies are spread for medium  $M_x$ , the more likely the impact of  $M_x$  on collaboration performance depends on the specific characteristics of the corresponding studies. Example:  $M_{4.01}$  for the multiparty studies.

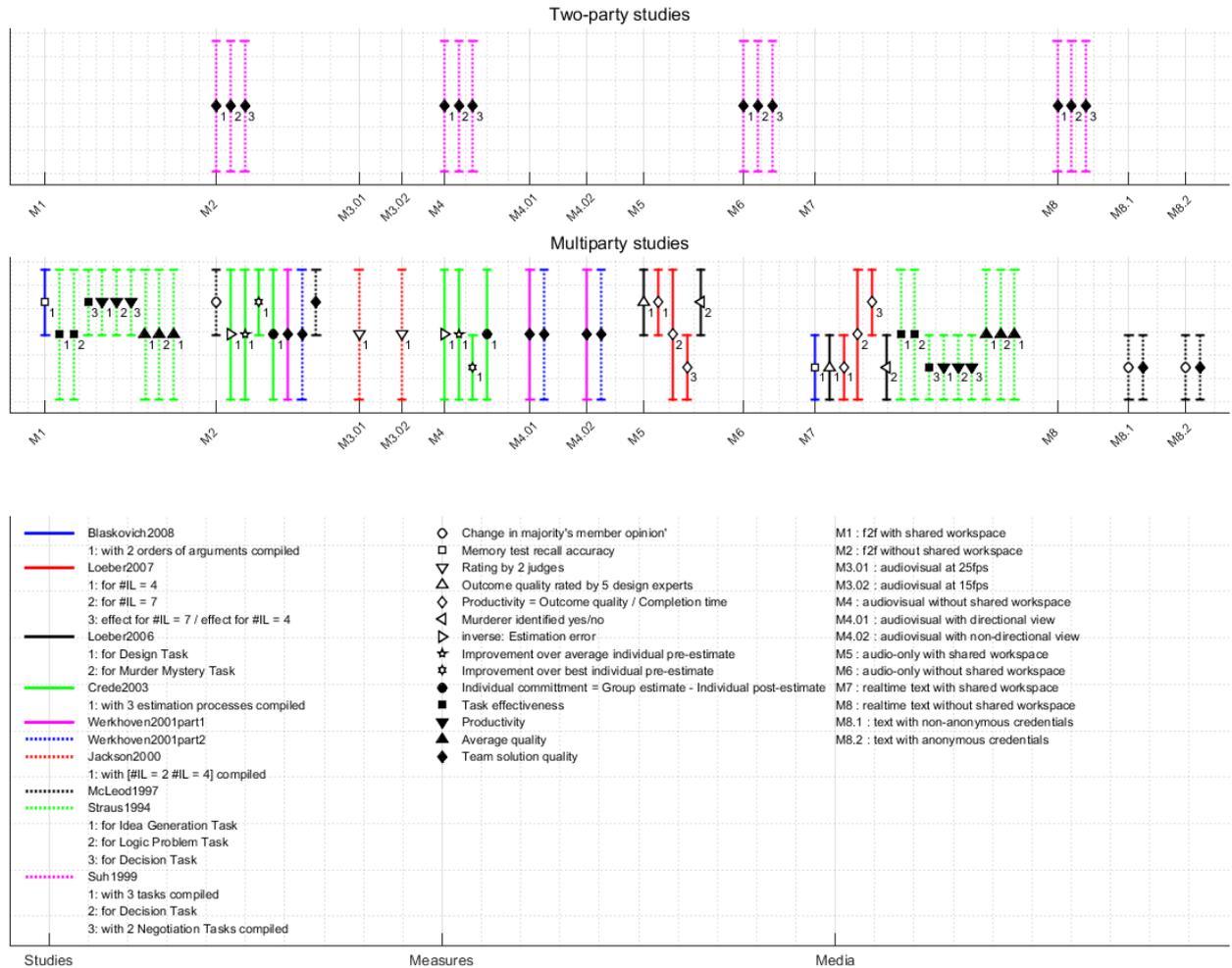
**Main finding for two-party studies:**

None since there is too little data, i.e. just one data point per medium, to draw proper conclusions.

**Main finding for multiparty studies:**

Extracting a systematic ranking is difficult since hardly any medium was used in more than one study. Nevertheless, there is a tendency that f2f ( $M_2$ ) leads to better collaboration performance than audiovisual, especially when there is no directional view of interlocutors provided ( $M_{4.02}$ ).

Figure B.1: Impact of communication media on measures of positive task outcome across studies with collaboration game and survival tasks: Higher values represent higher collaboration performance.



**Interpretation Method:**

1. Look at the positions and sizes of the errorbars per medium  $Mx$ .
2. The more the errorbars for the medium  $Mx$  are clustered towards the upper end, the more likely  $Mx$  leads to better task results compared to other media. Example:  $M1$  for multiparty studies, in tendency.
3. The more the errorbars are clustered toward the lower end, the more likely  $Mx$  leads to worse task results. Example:  $M7$  for multiparty studies, in tendency.
4. The more centrally positioned and at the same time the smaller the errorbars are for the medium  $Mx$ , the more likely  $Mx$  leads to average task results compared to other media. Example: none in this plot.
5. The more centrally positioned and at the same time the wider the errorbars are for the medium  $Mx$ , the more likely  $Mx$  does not differ in the task results compared to other media. Example:  $M2$  for multiparty studies, in tendency.
6. The more the errorbars of individual studies are spread for medium  $Mx$ , the more likely the impact of  $Mx$  on task results depends on the specific characteristics of the corresponding studies. Example:  $M5$  for multiparty studies, in tendency.

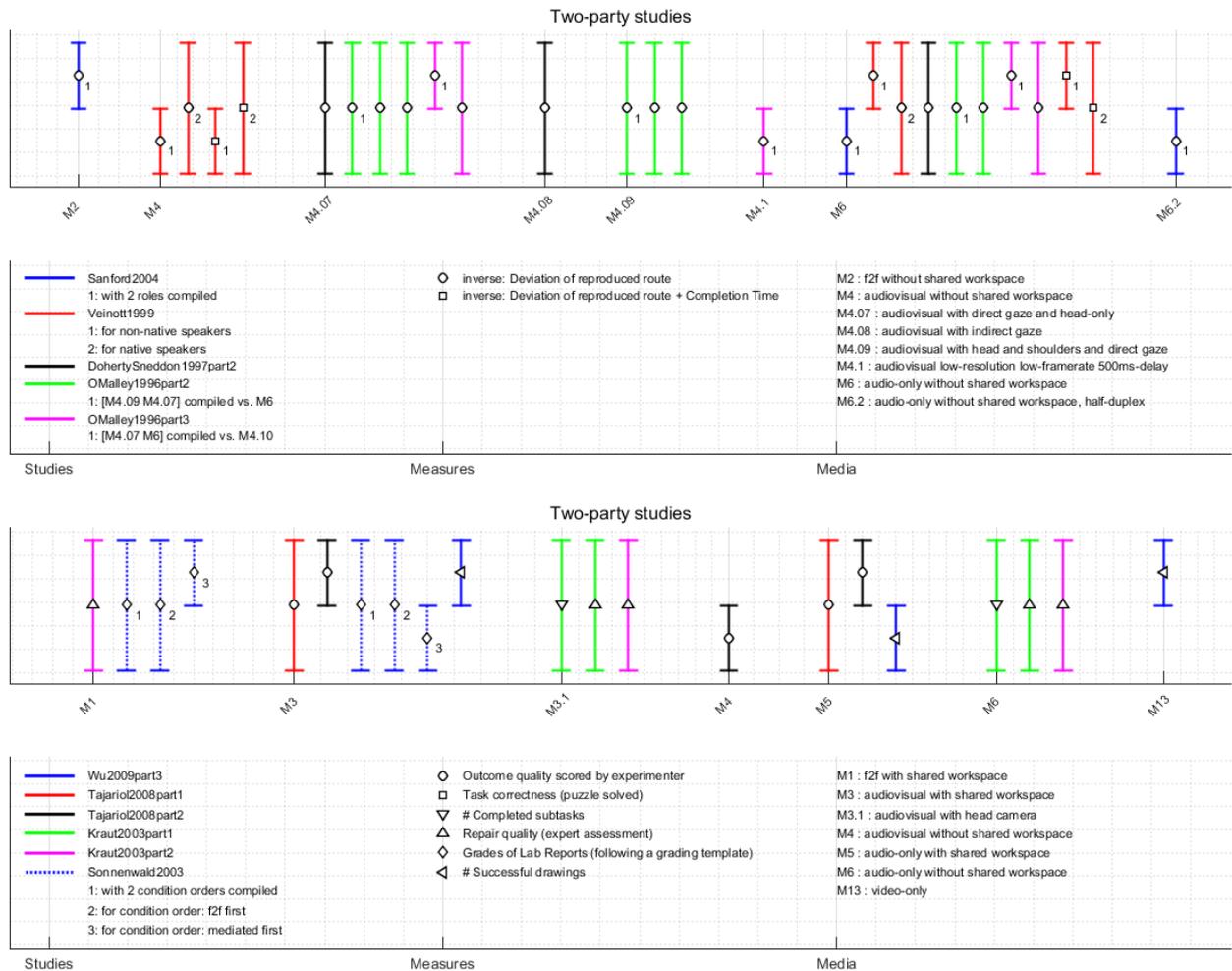
**Main finding for two-party studies:**

None since there is too little data, i.e. just one variable and study per medium, to draw proper conclusions.

**Main finding for multiparty studies:**

There is no simple ranking of media across studies in terms of completion time. Either no significant differences were found (large errorbars within a medium) or any significant rankings within certain studies were not found in other studies (mix of small and large errorbars within a medium). Filtering out non-significant differences (disregard all large errorbars), and grouping the media into rough categories, i.e. f2f ( $M1, M2$ ), audiovisual ( $M3.x, M4.x$ ), audio-only ( $M5.x, M6.x$ ), and text-only ( $M7.x, M8.x$ ), some trend is visible though: F2f leads to better task results than text-only. Audiovisual and audio-only can not be reasonably interpreted.

Figure B.2: Impact of communication media on measures of positive task outcome across studies with design, decision, minority opinion, mystery, and estimation tasks: Higher values represent better task results.



**Interpretation Method:**

1. Look at the positions and sizes of the errorbars per medium  $M_x$ .

2. The more the errorbars for the medium  $M_x$  are clustered towards the upper end, the more likely  $M_x$  leads to better task results compared to other media. Example: none in this plot.

3. The more the errorbars are clustered toward the lower end, the more likely  $M_x$  leads to worse task results. Example:  $M_4$  for map task studies (two top panels), in tendency.

4. The more centrally positioned and at the same time the smaller the errorbars are for the medium  $M_x$ , the more likely  $M_x$  leads to average task results compared to other media. Example: none in this plot.

5. The more centrally positioned and at the same time the wider the errorbars are for the medium  $M_x$ , the more likely  $M_x$  does not differ in the task results compared to other media. Example:  $M_{4.07}$  for map task studies (top two panels).

6. The more the errorbars of individual studies are spread for medium  $M_x$ , the more likely the impact of  $M_x$  on task results depends on the specific characteristics of the corresponding studies. Example: none in this plot.

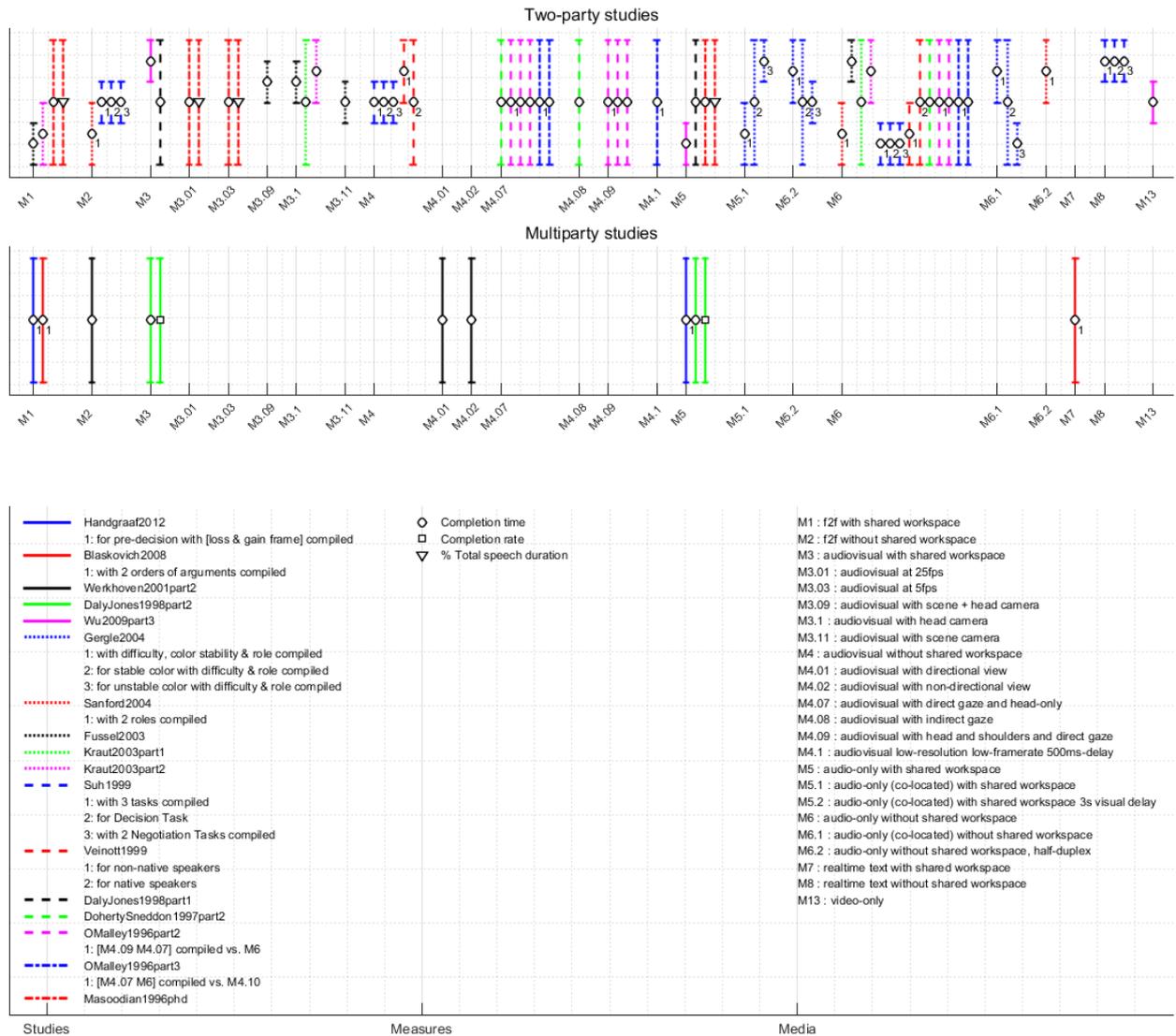
**Main finding for two-party studies:**

No clear tendency is visible. Either no significant differences were found (large errorbars within a medium), or any significant rankings within certain studies were not found in other studies or were contrary between studies (mix of small and large errorbars within a medium, and small errorbars are spread within a medium).

**Main finding for multiparty studies:**

None, as none of the reviewed multiparty study used the current variables.

Figure B.3: Impact of communication media on measures of positive task outcome across studies with map task (two top panels) and helper-worker & other tasks (two bottom panels): Higher values represent better task results.



**Interpretation Method:**

1. Look at the positions and sizes of the errorbars per medium  $M_x$ .
2. The more the errorbars for the medium  $M_x$  are clustered towards the upper end, the more likely  $M_x$  leads to an increased task completion time compared to other media. Example:  $M_8$  for two-party studies.
3. The more the errorbars are clustered toward the lower end, the more likely  $M_x$  leads to a decreased task completion time. Example:  $M_1$  for two-party studies, in tendency.
4. The more centrally positioned and at the same time the smaller the errorbars are for the medium  $M_x$ , the more likely  $M_x$  has an average task completion time compared to other media. Example:  $M_4$  for two-party studies, in tendency.
5. The more centrally positioned and at the same time the wider the errorbars are for the medium  $M_x$ , the more likely  $M_x$  does not differ in task completion time compared to other media. Example: all media for multiparty studies.
6. The more the errorbars of individual studies are spread for medium  $M_x$ , the more likely the impact of  $M_x$  on task completion time depends on the specific characteristics of the corresponding studies. Example:  $M_{5.1}$  for two-party studies.

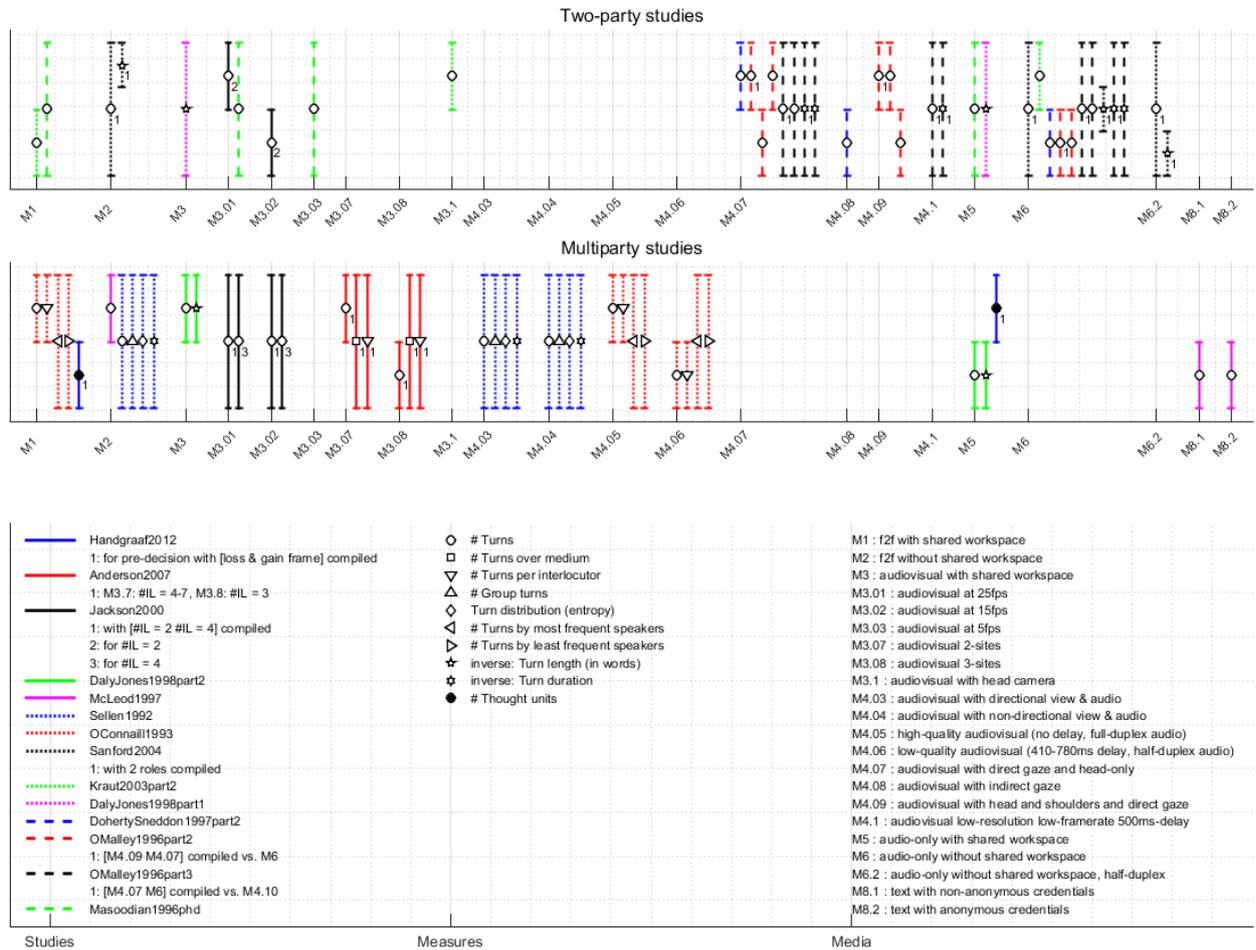
**Main finding for two-party studies:**

There is no simple ranking of media across studies in terms of completion time. Either no significant differences were found (large errorbars within a medium) or any significant rankings within certain studies were not found in other studies (mix of small and large errorbars within a medium). Filtering out non-significant differences (disregard all large errorbars), and grouping the media into rough categories, i.e. f2f ( $M_1, M_2$ ), audiovisual ( $M_{3.x}, M_{4.x}$ ), audio-only ( $M_{5.x}, M_{6.x}$ ), and text-only ( $M_7, M_8$ ), there is some tendency of a ranking: f2f ( $M_1, M_2$ ) < audiovisual ( $M_{3.x}, M_{4.x}$ ) < text-only ( $M_7, M_8$ ). However, audio-only ( $M_{5.x}, M_{6.x}$ ) does not show a systematic tendency.

**Main finding for multiparty studies:**

No differences of media in terms of completion time could be found.

Figure B.4: Impact of communication media on measures of task completion time across studies: Higher values represent higher task completion time.



Interpretation Method:

1. Look at the positions and sizes of the errorbars per medium  $M_x$ .
2. The more the errorbars for the medium  $M_x$  are clustered towards the upper end, the more likely  $M_x$  leads to an increased turn-taking rate compared to other media. Example:  $M3$  for multiparty studies.
3. The more the errorbars are clustered toward the lower end, the more likely  $M_x$  leads to a decreased turn-taking rate. Example:  $M8.x$  for multiparty studies.
4. The more centrally positioned and at the same time the smaller the errorbars are for the medium  $M_x$ , the more likely  $M_x$  has an average turn-taking rate compared to other media. Example: None in this plot.
5. The more centrally positioned and at the same time the wider the errorbars are for the medium  $M_x$ , the more likely  $M_x$  does not differ in turn-taking rate compared to other media. Example:  $M4.x$  for multiparty studies.
6. The more the errorbars of individual studies are spread for medium  $M_x$ , the more likely the impact of  $M_x$  on turn-taking rate depends on the specific characteristics of the corresponding studies. Example:  $M6$  for two-party studies.

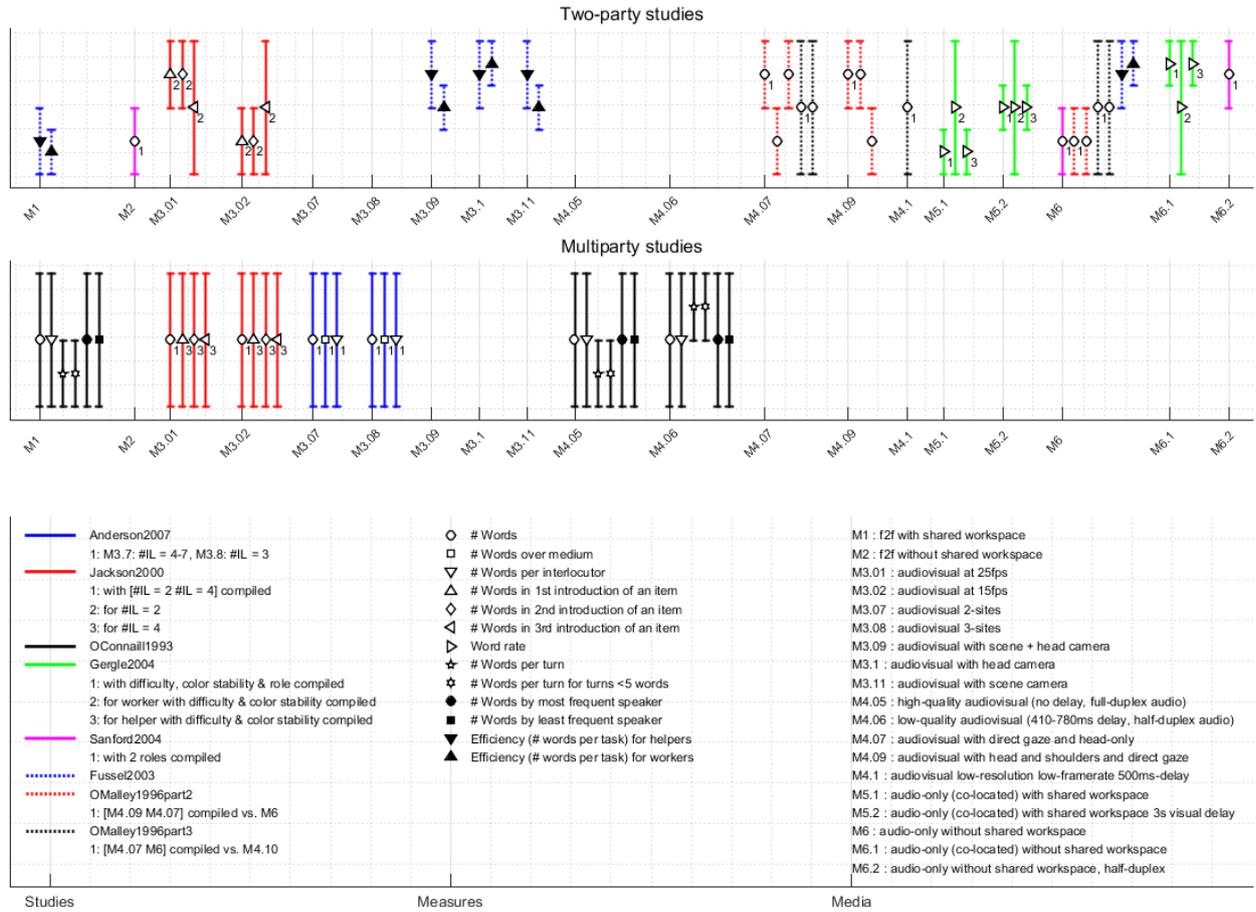
Main finding for two-party studies:

There is no simple ranking of media across studies in terms of turn-related measures. Either no significant differences were found (large errorbars within a medium), or any significant rankings within certain studies were not found in other studies (mix of small and large errorbars within a medium), or any significant rankings were contrary between studies (small errorbars are spread within a medium). Even filtering out non-significant differences (disregard all large errorbars), and grouping the media into rough categories, i.e. f2f ( $M1$ ,  $M2$ ), audiovisual ( $M3.x$ ,  $M4.x$ ), audio-only ( $M5.x$ ,  $M6.x$ ), and text-only ( $M8.x$ ), no systematic tendency of a ranking is visible.

Main finding for multiparty studies:

Same as main finding for two-party studies.

Figure B.5: Impact of communication media on turn-related measures across studies: Higher values represent a higher turn-taking rate, i.e. larger number of turns and shorter length of turns.



Interpretation Method:

1. Look at the positions and sizes of the errorbars per medium  $M_x$ .
2. The more the errorbars for the medium  $M_x$  are clustered towards the upper end, the more likely  $M_x$  leads to an increased relative word rate compared to other media. Example:  $M_{3.1}$  for two-party studies.
3. The more the errorbars are clustered toward the lower end, the more likely  $M_x$  leads to a decreased relative word rate. Example:  $M_1$  for two-party studies.
4. The more centrally positioned and at the same time the smaller the errorbars are for the medium  $M_x$ , the more likely  $M_x$  has an average relative word rate compared to other media. Example:  $M_{5.2}$  for two-party studies.
5. The more centrally positioned and at the same time the wider the errorbars are for the medium  $M_x$ , the more likely  $M_x$  does not differ in relative word rate compared to other media. Example:  $M_{3.x}$  for multiparty studies.
6. The more the errorbars of individual studies are spread for medium  $M_x$ , the more likely the impact of  $M_x$  on relative word rate depends on the specific characteristics of the corresponding studies. Example:  $M_6$  for two-party studies.

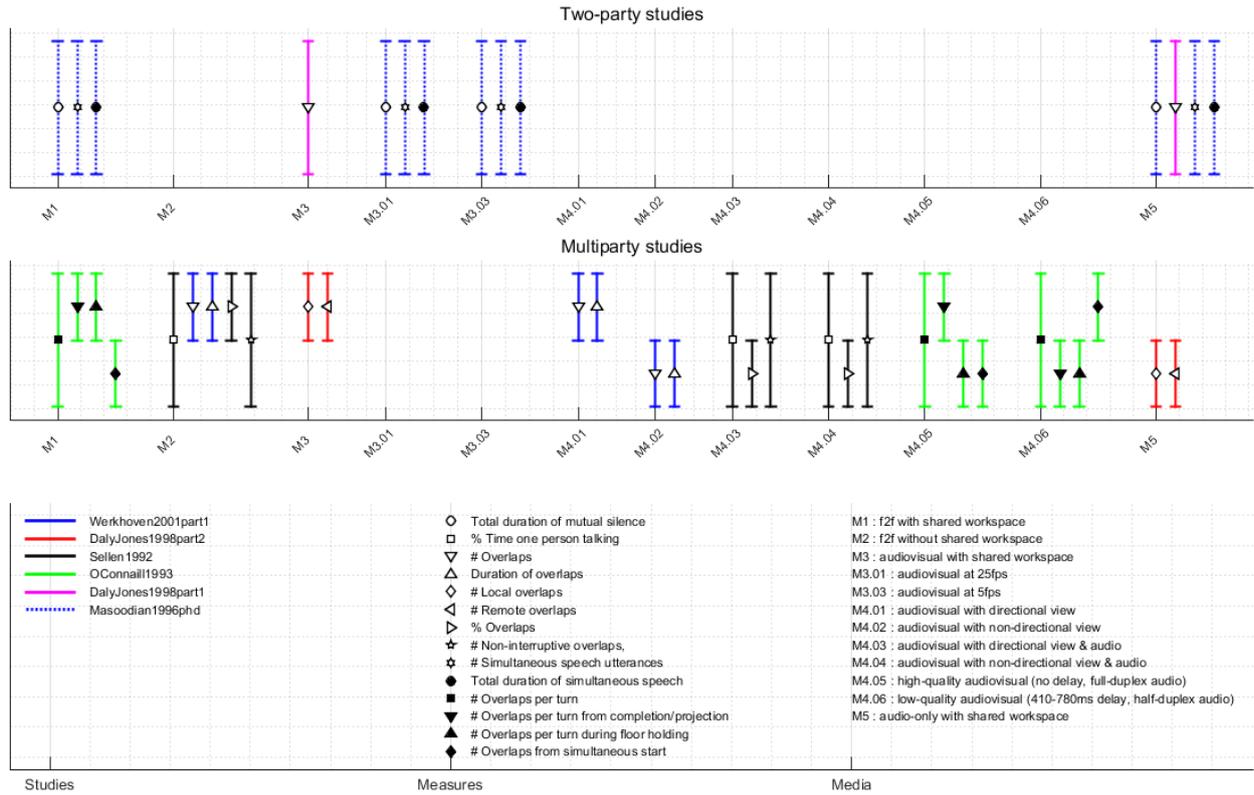
Main finding for two-party studies:

Extracting a systematic ranking is difficult since hardly any medium was used in more than one study. Grouping the media into rough categories, i.e. f2f ( $M_1$ ,  $M_2$ ), audiovisual ( $M_{3.x}$ ,  $M_{4.x}$ ), and audio-only ( $M_{5.x}$ ,  $M_{6.x}$ ), some trend is visible though: f2f leads to lower relative word rate than audiovisual (most errorbars, but not all, in higher range), while audio-only shows no systematic behavior (errorbars spread across whole range).

Main finding for multiparty studies:

Essentially no differences of media in terms of relative word rate could be found.

Figure B.6: Impact of communication media on word-related measures across studies: Higher values represent a higher relative word rate, i.e. larger number of words and higher values of word rate.



Interpretation Method:

1. Look at the positions and sizes of the errorbars per medium  $M_x$ .
2. The more the errorbars for the medium  $M_x$  are clustered towards the upper end, the more likely  $M_x$  leads to an increased amount of overlap compared to other media. Example:  $M_2$  for multiparty studies.
3. The more the errorbars are clustered toward the lower end, the more likely  $M_x$  leads to a decreased amount of overlap. Example:  $M_5$  for multiparty studies.
4. The more centrally positioned and at the same time the smaller the errorbars are for the medium  $M_x$ , the more likely  $M_x$  has an average amount of overlap compared to other media. Example: None in this plot.
5. The more centrally positioned and at the same time the wider the errorbars are for the medium  $M_x$ , the more likely  $M_x$  does not differ in amount of overlap compared to other media. Example: all media for two-party studies.
6. The more the errorbars of individual studies are spread for medium  $M_x$ , the more likely the impact of  $M_x$  on amount of overlap depends on the specific characteristics of the corresponding studies. Example: Example:  $M_{4.05}$  for multiparty studies.

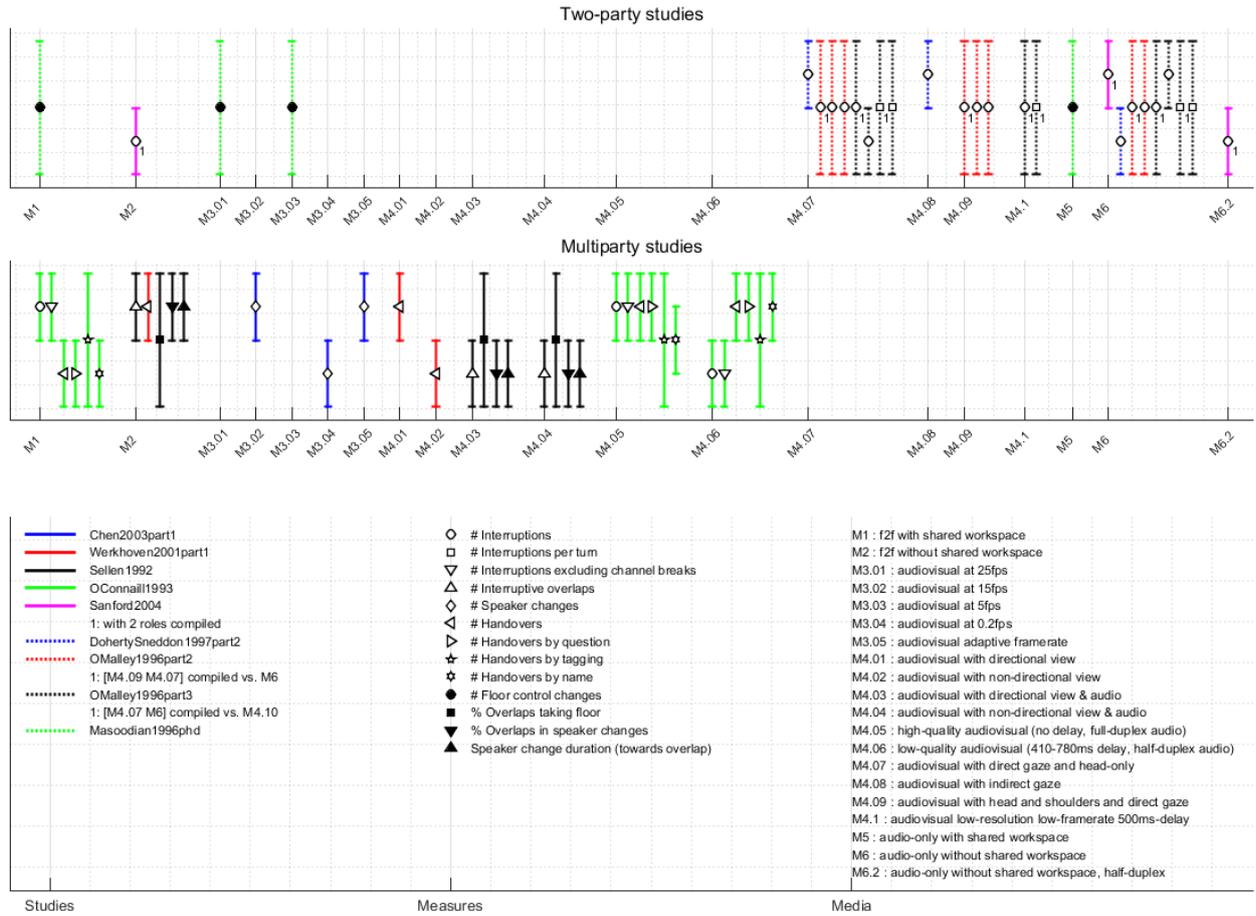
Main finding for two-party studies:

No differences of media in terms of amount of overlap could be found.

Main finding for multiparty studies:

Extracting a systematic ranking is difficult since hardly any medium was used in more than one study. Grouping the media into rough categories, i.e. f2f ( $M_1, M_2$ ), audiovisual ( $M_{3.x}, M_{4.x}$ ), and audio-only ( $M_{5.x}$ ), some trend is visible though: f2f leads to higher amount of overlap (most errorbars, but not all, in higher range) than audio-only (most errorbars, but not all, in lower range), while audiovisual shows no systematic behavior (errorbars spread across whole range).

Figure B.7: Impact of communication media on speaker-state-related measures across studies: Higher values represent higher speaker-state probabilities for silence, single-talk and multi-talk. Most measures refer to multi-talk, i.e. amount of overlap.



Interpretation Method:

1. Look at the positions and sizes of the errorbars per medium  $M_x$ .
2. The more the errorbars for the medium  $M_x$  are clustered towards the upper end, the more likely  $M_x$  leads to an increased amount of speaker changes compared to other media. Example:  $M_2$  for multiparty studies.
3. The more the errorbars are clustered toward the lower end, the more likely  $M_x$  leads to a decreased amount of speaker changes. Example:  $M_{4.03}$  for multiparty studies.
4. The more centrally positioned and at the same time the smaller the errorbars are for the medium  $M_x$ , the more likely  $M_x$  has an average amount of speaker changes compared to other media. Example: None in this plot.
5. The more centrally positioned and at the same time the wider the errorbars are for the medium  $M_x$ , the more likely  $M_x$  does not differ in amount of speaker changes compared to other media. Example: most media for two-party studies.
6. The more the errorbars of individual studies are spread for medium  $M_x$ , the more likely the impact of  $M_x$  on amount of speaker changes depends on the specific characteristics of the corresponding studies. Example: Example:  $M_6$  for two-party studies.

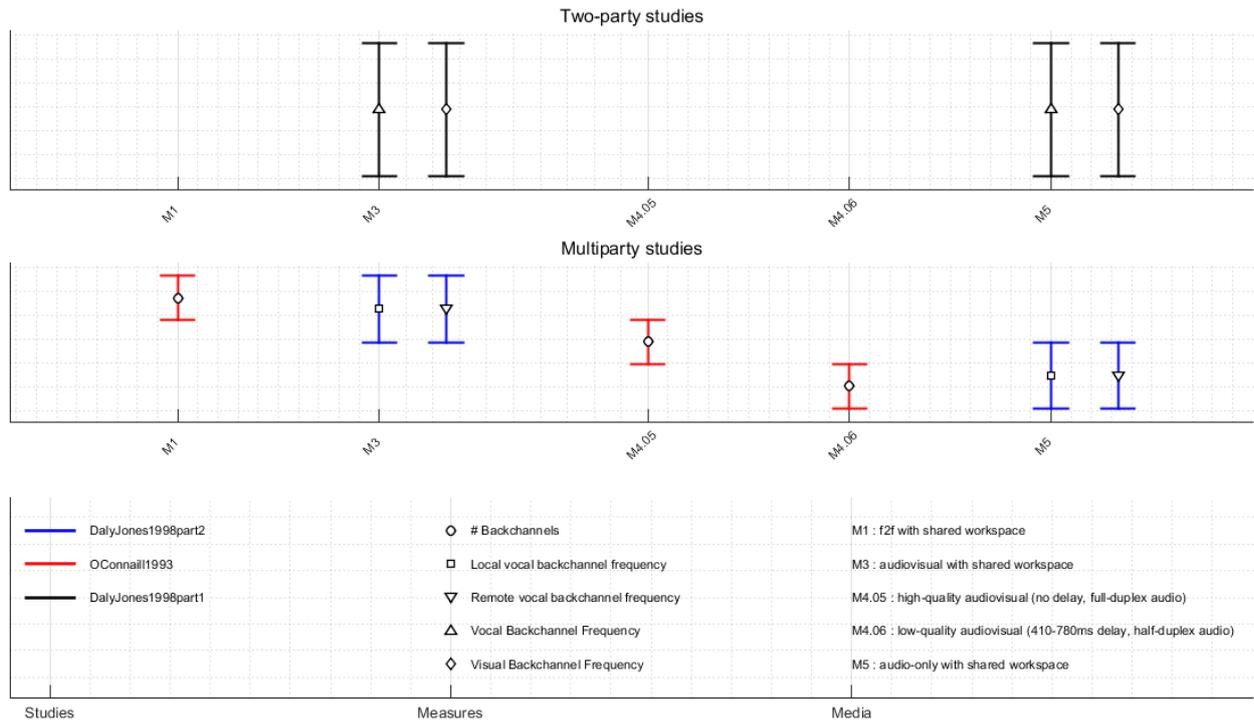
Main finding for two-party studies:

Essentially, no differences of media in terms of amount of speaker changes could be found.

Main finding for multiparty studies:

Extracting a systematic ranking is difficult since hardly any medium was used in more than one study. Even grouping the media into rough categories, i.e. f2f ( $M_1$ ,  $M_2$ ) and audiovisual ( $M_{3.x}$ ,  $M_{4.x}$ ), does not reveal a systematic trend: for both, f2f and audiovisual, errorbars are spread across whole range.

Figure B.8: Impact of communication media on speaker-change-related measures across studies: Higher values represent larger amount of speaker changes in terms of handovers and interruptions.



**Interpretation Method:**

1. Look at the positions and sizes of the errorbars per medium  $Mx$ .
2. The more the errorbars for the medium  $Mx$  are clustered towards the upper end, the more likely  $Mx$  leads to an increased amount of backchannels compared to other media. Example: none in this plot, too little data.
3. The more the errorbars are clustered toward the lower end, the more likely  $Mx$  leads to a decreased amount of backchannels. Example: none in this plot, too little data.
4. The more centrally positioned and at the same time the smaller the errorbars are for the medium  $Mx$ , the more likely  $Mx$  has an average amount of backchannels compared to other media. Example: none in this plot, too little data.
5. The more centrally positioned and at the same time the wider the errorbars are for the medium  $Mx$ , the more likely  $Mx$  does not differ in amount of backchannels compared to other media. Example: none in this plot, too little data.
6. The more the errorbars of individual studies are spread for medium  $Mx$ , the more likely the impact of  $Mx$  on amount of backchannels depends on the specific characteristics of the corresponding studies. Example: none in this plot, too little data.

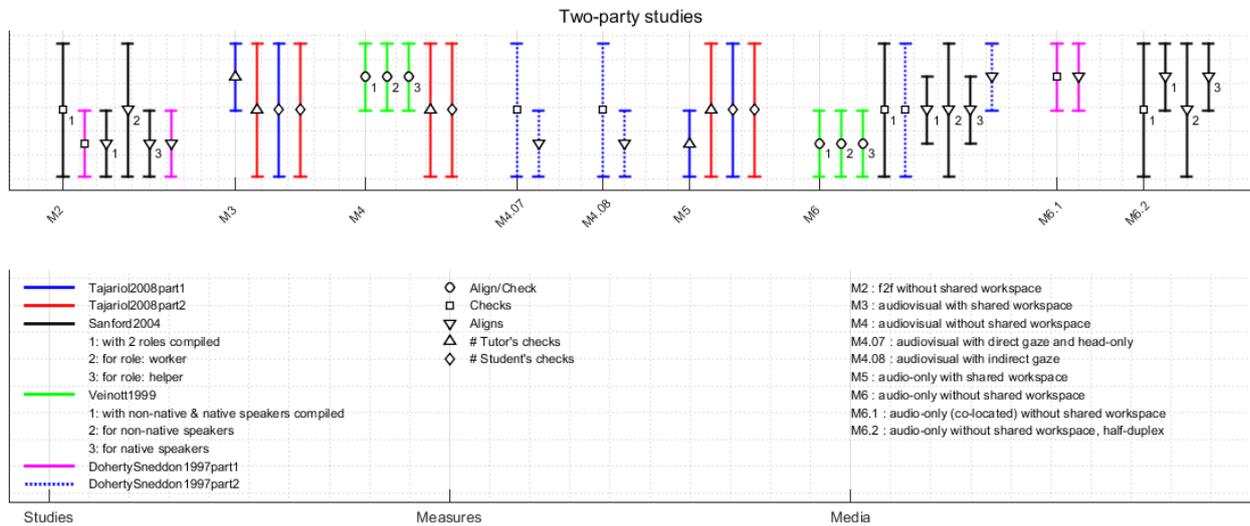
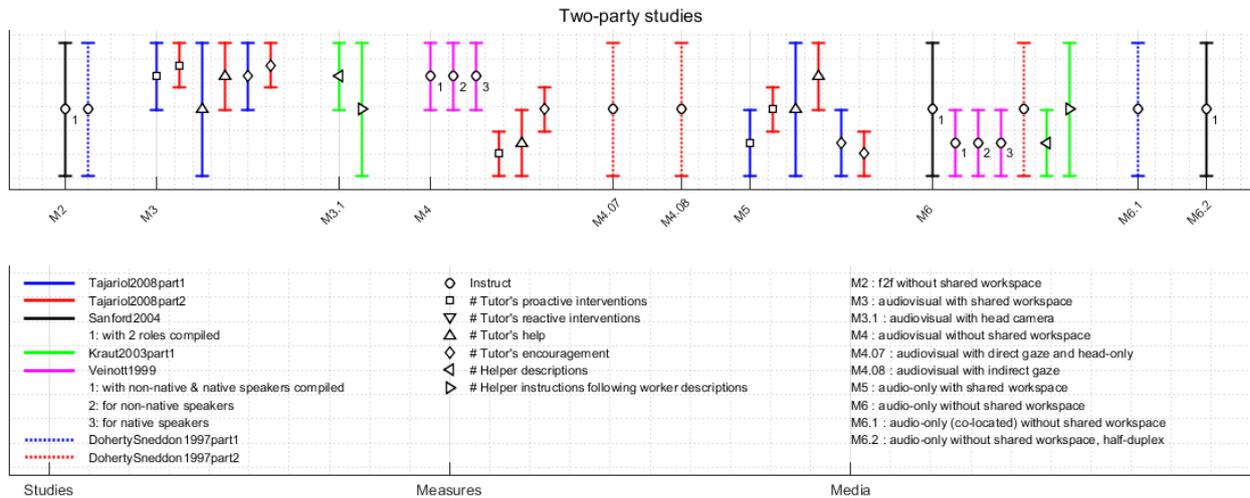
**Main finding for two-party studies:**

None since there is too little data, i.e. just one or two data points per medium, to draw proper conclusions.

**Main finding for multiparty studies:**

None since there is too little data, i.e. just one or two data points per medium, to draw proper conclusions. However, the few data points suggest the hypothesis that the higher the media richness is, the higher is the amount of backchannels (within each study a corresponding ranking is visible).

Figure B.9: Impact of communication media on backchannel-related measures across studies: Higher values represent higher amount of backchannels.



**Interpretation Method:**

1. Look at the positions and sizes of the errorbars per medium  $M_x$ .
2. The more the errorbars for the medium  $M_x$  are clustered towards the upper end, the more likely  $M_x$  leads to an increased amount of instructions compared to other media. Example:  $M_3$  for the studies concerning amount of instructions (two top panels).
3. The more the errorbars are clustered toward the lower end, the more likely  $M_x$  leads to a decreased amount of instructions. Example:  $M_6$  for the studies concerning amount of instructions (two top panels), in tendency.
4. The more centrally positioned and at the same time the smaller the errorbars are for the medium  $M_x$ , the more likely  $M_x$  has an average amount of instructions compared to other media. Example: none in this plot.
5. The more centrally positioned and at the same time the wider the errorbars are for the medium  $M_x$ , the more likely  $M_x$  does not differ in amount of instructions compared to other media. Example:  $M_3$  for the studies concerning amount of aligns and checks (two bottom panels).
6. The more the errorbars of individual studies are spread for medium  $M_x$ , the more likely the impact of  $M_x$  on amount of instructions depends on the specific characteristics of the corresponding studies. Example:  $M_6$  for the studies concerning amount of aligns and checks (two bottom panels).

**Main finding for two-party studies concerning amount of instructions (two top panels):**

Focusing on the media audiovisual with shared workspace ( $M_3$ ,  $M_{3.1}$ ), audiovisual without shared workspace ( $M_4$ ), audio-only with shared workspace ( $M_5$ ), and audio-only without share workspace ( $M_6$ ,  $M_{6.1}$ ,  $M_{6.2}$ ), some interaction between the communication modality (audiovisual vs. audio-only) and the availability of a shared workspace is visible. Audiovisual with shared workspace leads to more instructions than audio-only without shared workspace. For audiovisual without shared workspace and audio-only with shared workspace, the amount of instructions is spread.

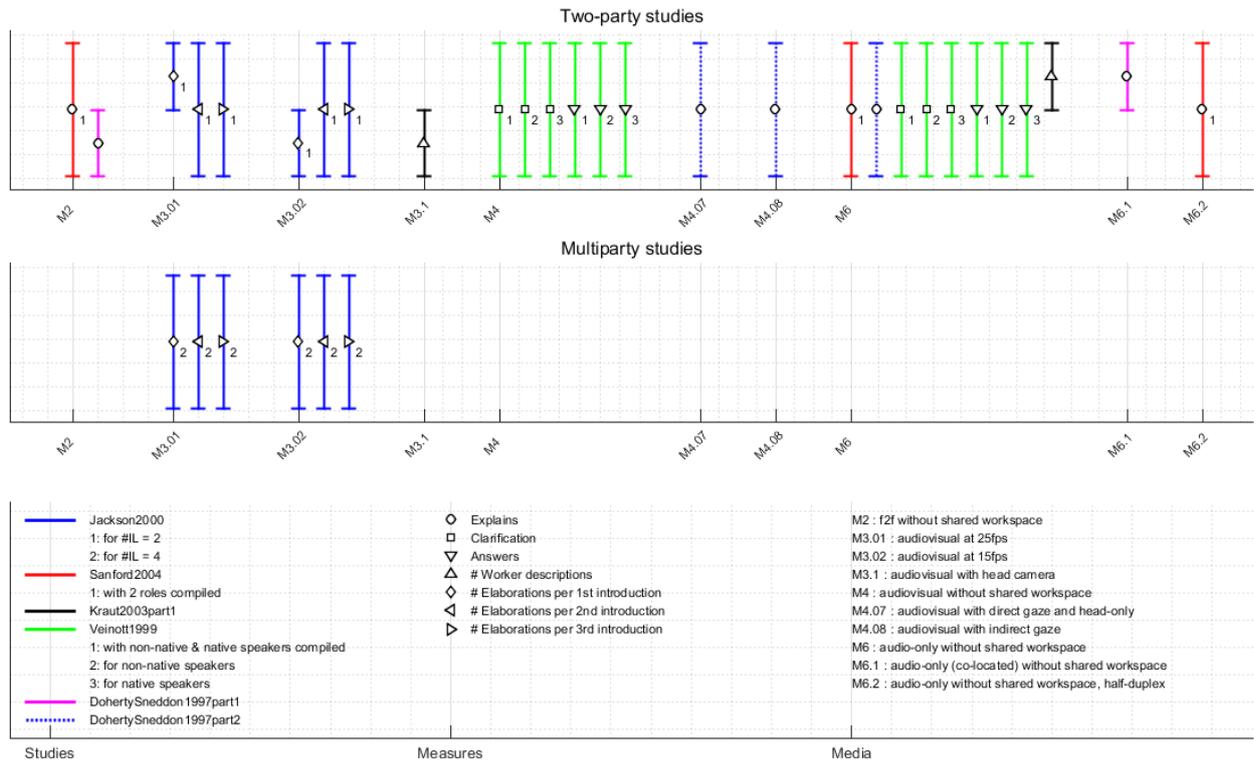
**Main finding for two-party studies concerning amount of aligns and checks (two bottom panels):**

There is no simple ranking of media across studies in terms of aligns and checks. Either any significant rankings within certain studies were not found in other studies (mix of small and large errorbars within a medium) or any significant rankings were contrary between studies (small errorbars are spread within a medium). Even filtering out non-significant differences (disregard all large errorbars), and grouping the media into rough categories, i.e. f2f ( $M_2$ ), audiovisual ( $M_{3.x}$ ,  $M_{4.x}$ ), and audio-only ( $M_{5.x}$ ,  $M_{6.x}$ ), no systematic tendency of a ranking is visible, except that f2f requires a smaller amount of aligns and checks, while the other media sometimes require smaller, sometimes larger amount of aligns and checks.

**Main finding for multiparty studies:**

None, as none of the reviewed multiparty study used the current variables.

Figure B.10: Impact of communication media on instruction measures (two top panels) and align and check measures (two bottom panels) across studies: Higher values represent higher larger amount of instructions, aligns and checks.



**Interpretation Method:**

1. Look at the positions and sizes of the errorbars per medium  $M_x$ .
2. The more the errorbars for the medium  $M_x$  are clustered towards the upper end, the more likely  $M_x$  leads to an increased amount of explanations compared to other media. Example: none in this plot.
3. The more the errorbars are clustered toward the lower end, the more likely  $M_x$  leads to a decreased amount of explanations. Example: none in this plot.
4. The more centrally positioned and at the same time the smaller the errorbars are for the medium  $M_x$ , the more likely  $M_x$  has an average amount of explanations compared to other media. Example: none in this plot.
5. The more centrally positioned and at the same time the wider the errorbars are for the medium  $M_x$ , the more likely  $M_x$  does not differ in amount of explanations compared to other media. Example: essentially all media in this plot.
6. The more the errorbars of individual studies are spread for medium  $M_x$ , the more likely the impact of  $M_x$  on amount of explanations depends on the specific characteristics of the corresponding studies. Example: none in this plot.

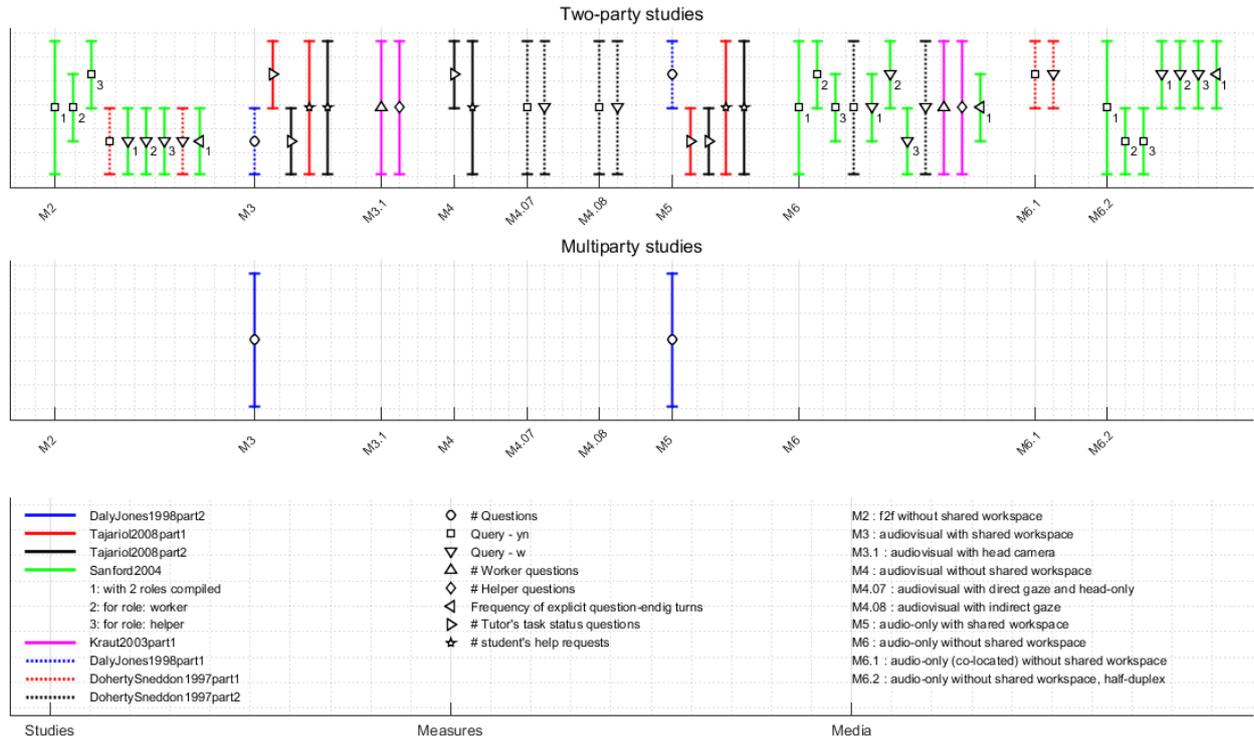
**Main finding for two-party studies:**

There is essentially no influence of the media on the amount of explanations.

**Main finding for multiparty studies:**

None since there is too little data, i.e. just one or two data points per medium, to draw proper conclusions.

Figure B.11: Impact of communication media on explanation measures across studies: Higher values represent larger amount of explanations.



**Interpretation Method:**

1. Look at the positions and sizes of the errorbars per medium  $Mx$ .
2. The more the errorbars for the medium  $Mx$  are clustered towards the upper end, the more likely  $Mx$  leads to more questions compared to other media. Example: none in this plot.
3. The more the errorbars are clustered toward the lower end, the more likely  $Mx$  leads to fewer questions. Example:  $M2$  for two-party studies.
4. The more centrally positioned and at the same time the smaller the errorbars are for the medium  $Mx$ , the more likely  $Mx$  has an average amount of questions compared to other media. Example: none in this plot.
5. The more centrally positioned and at the same time the wider the errorbars are for the medium  $Mx$ , the more likely  $Mx$  does not differ in amount of questions compared to other media. Example: none in this plot.
6. The more the errorbars of individual studies are spread for medium  $Mx$ , the more likely the impact of  $Mx$  on amount of questions depends on the specific characteristics of the corresponding studies. Example:  $M6$  for two-party studies.

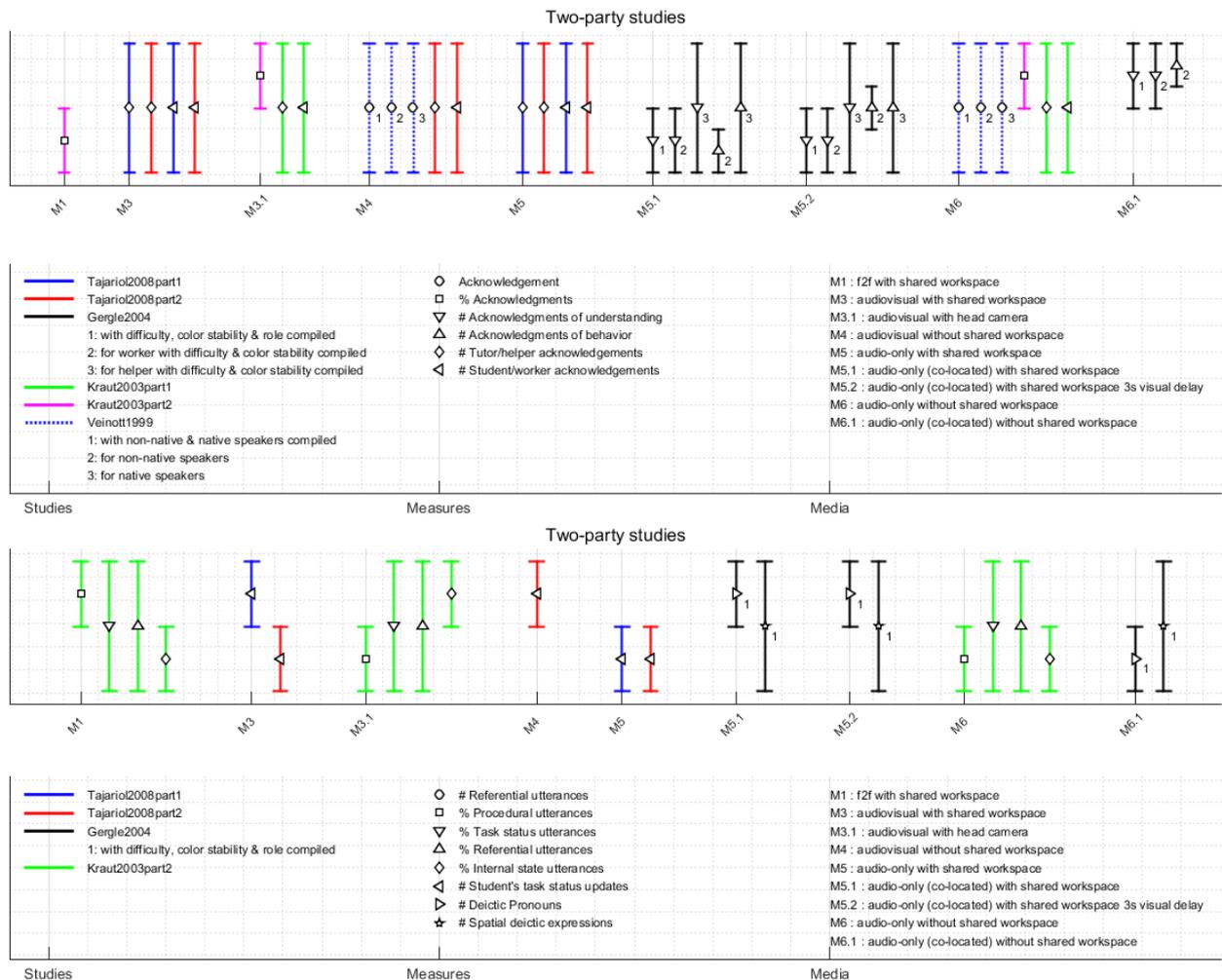
**Main finding for two-party studies:**

There is no simple ranking of media across studies in terms of amount of questions. Either no significant differences were found (large errorbars within a medium), or any significant rankings within certain studies were not found in other studies (mix of small and large errorbars within a medium), or any significant rankings were contrary between studies (small errorbars are spread within a medium). However, there is the tendency that f2f requires a smaller amount of questions, while the other media require sometimes a smaller, sometimes a larger amount of questions.

**Main finding for multiparty studies:**

None since there is too little data, i.e. just one or two data points per medium, to draw proper conclusions.

Figure B.12: Impact of communication media on question measures across studies: Higher values represent larger amount of questions.



**Interpretation Method:**

1. Look at the positions and sizes of the errorbars per medium  $M_x$ .
2. The more the errorbars for the medium  $M_x$  are clustered towards the upper end, the more likely  $M_x$  leads to an increased amount of acknowledgments compared to other media. Example:  $M_{6.1}$  for two-party studies.
3. The more the errorbars are clustered toward the lower end, the more likely  $M_x$  leads to a decreased amount of acknowledgments. Example:  $M_{5.1}$  for two-party studies, in tendency.
4. The more centrally positioned and at the same time the smaller the errorbars are for the medium  $M_x$ , the more likely  $M_x$  has an average amount of acknowledgments compared to other media. Example: none in this plot.
5. The more centrally positioned and at the same time the wider the errorbars are for the medium  $M_x$ , the more likely  $M_x$  does not differ in amount of acknowledgments compared to other media. Example:  $M_3$  for two-party studies.
6. The more the errorbars of individual studies are spread for medium  $M_x$ , the more likely the impact of  $M_x$  on amount of acknowledgments depends on the specific characteristics of the corresponding studies. Example:  $M_{5.2}$  for two-party studies, in tendency.

**Main finding for two-party studies concerning amount of acknowledgments:**

There is no simple ranking of media across studies. In most cases no significant differences were found (large errorbars within a medium), or any significant rankings within certain studies were not found in other studies (mix of small and large errorbars within a medium).

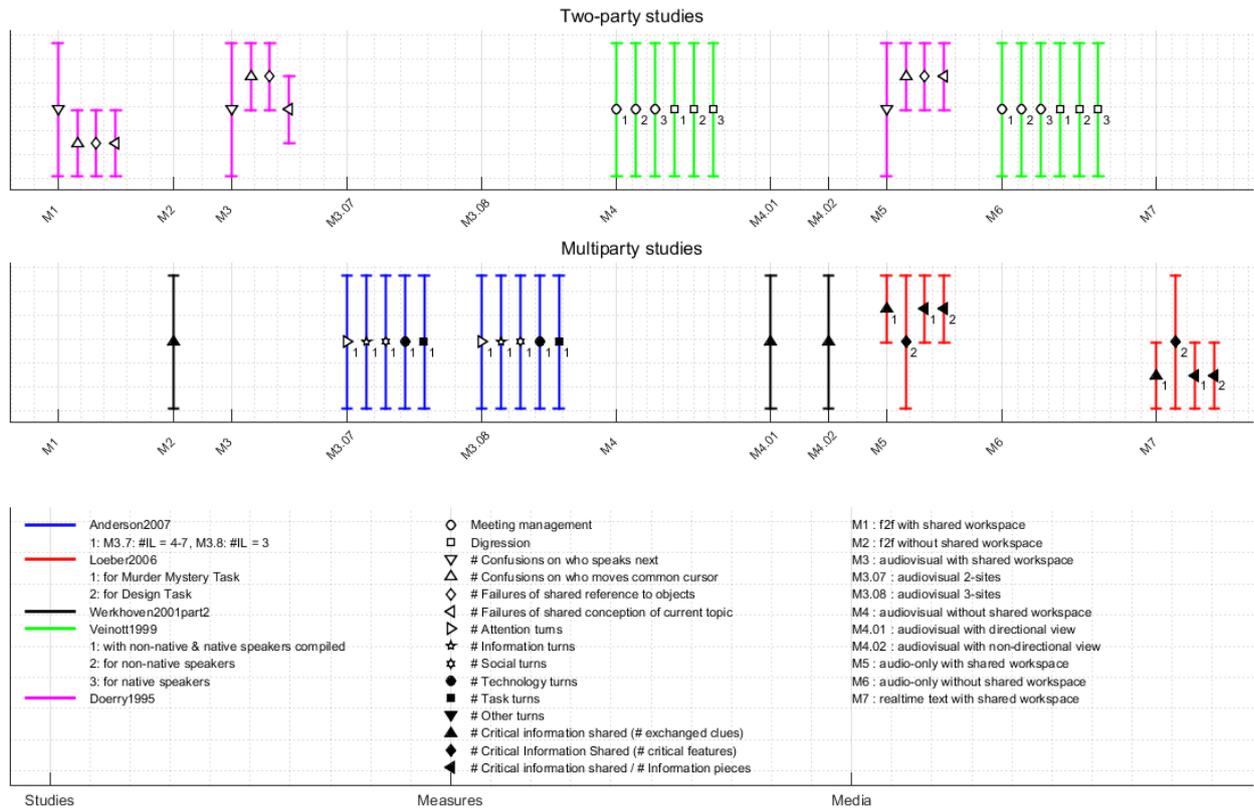
**Main finding for two-party studies concerning amount of object, process and status utterances:**

There is no simple ranking of media across studies, even after grouping into rough categories, i.e. f2f ( $M_1$ ), audiovisual ( $M_{3.x}$ ,  $M_{4.x}$ ), and audio-only ( $M_{5.x}$ ,  $M_{6.x}$ ): Either any significant rankings within certain studies were not found in other studies (mix of small and large errorbars within a medium), or any significant rankings were contrary between studies (small errorbars are spread within a medium).

**Main finding for multiparty studies:**

None, as none of the reviewed multiparty study used the current variables.

Figure B.13: Impact of communication media on acknowledgment measures (two top panels) and measures on references on object, process and status (two bottom panels) across studies: Higher values represent higher larger amount of acknowledgments and of object, process and status utterances.



Interpretation Method:

1. Look at the positions and sizes of the errorbars per medium  $M_x$ .
2. The more the errorbars for the medium  $M_x$  are clustered towards the upper end, the more likely  $M_x$  leads to an increased amount of conversation management & failures compared to other media. Example:  $M_5$  for the two-party and multiparty studies.
3. The more the errorbars are clustered toward the lower end, the more likely  $M_x$  leads to a decreased amount of conversation management & failures. Example:  $M_7$  for multiparty studies.
4. The more centrally positioned and at the same time the smaller the errorbars are for the medium  $M_x$ , the more likely  $M_x$  has an average amount of conversation management & failures compared to other media. Example: none in this plot.
5. The more centrally positioned and at the same time the wider the errorbars are for the medium  $M_x$ , the more likely  $M_x$  does not differ in amount of conversation management & failures compared to other media. Example:  $M_{3.x}$  for multiparty studies.
6. The more the errorbars of individual studies are spread for medium  $M_x$ , the more likely the impact of  $M_x$  on amount of conversation management & failures depends on the specific characteristics of the corresponding studies. Example: none in this plot.

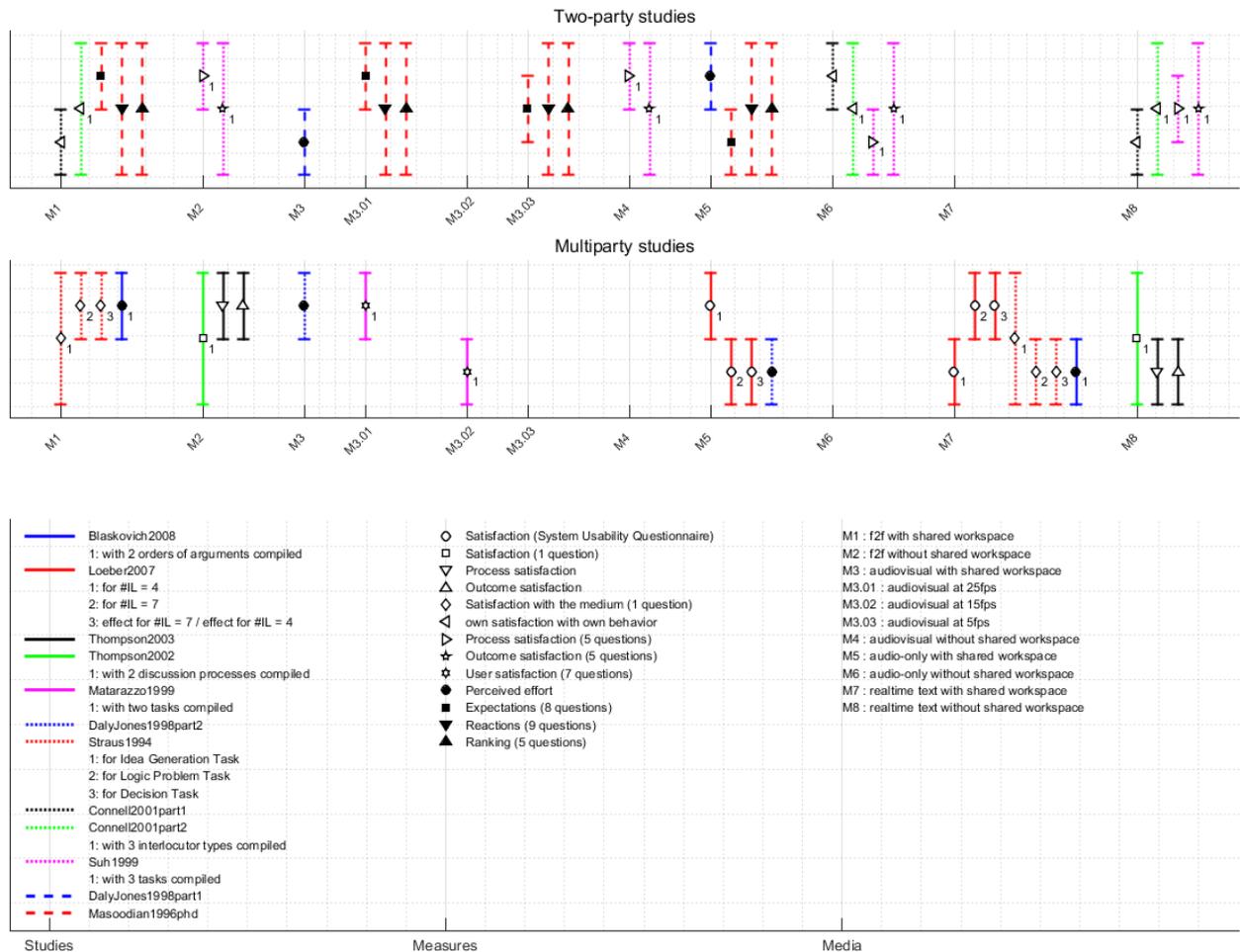
Main finding for two-party studies:

None since there is too little data, i.e. just one or two data points per medium, to draw proper conclusions.

Main finding for multiparty studies:

None since there is too little data, i.e. just one or two data points per medium, to draw proper conclusions.

Figure B.14: Impact of communication media on conversation management and conversation breakdown measures across studies: Higher values represent larger amount of attempts to manage the conversation or larger amount of conversation failures.



**Interpretation Method:**

1. Look at the positions and sizes of the errorbars per medium  $M_x$ .
2. The more the errorbars for the medium  $M_x$  are clustered towards the upper end, the more likely  $M_x$  leads to an increased satisfaction compared to other media. Example:  $M_1$  for multiparty studies.
3. The more the errorbars are clustered toward the lower end, the more likely  $M_x$  leads to a decreased satisfaction. Example:  $M_5$  for multiparty studies.
4. The more centrally positioned and at the same time the smaller the errorbars are for the medium  $M_x$ , the more likely  $M_x$  has an average satisfaction compared to other media. Example: none in this plot.
5. The more centrally positioned and at the same time the wider the errorbars are for the medium  $M_x$ , the more likely  $M_x$  does not differ in satisfaction compared to other media. Example:  $M_{3.03}$  for two-party studies.
6. The more the errorbars of individual studies are spread for medium  $M_x$ , the more likely the impact of  $M_x$  on satisfaction depends on the specific characteristics of the corresponding studies. Example:  $M_7$  for multiparty studies.

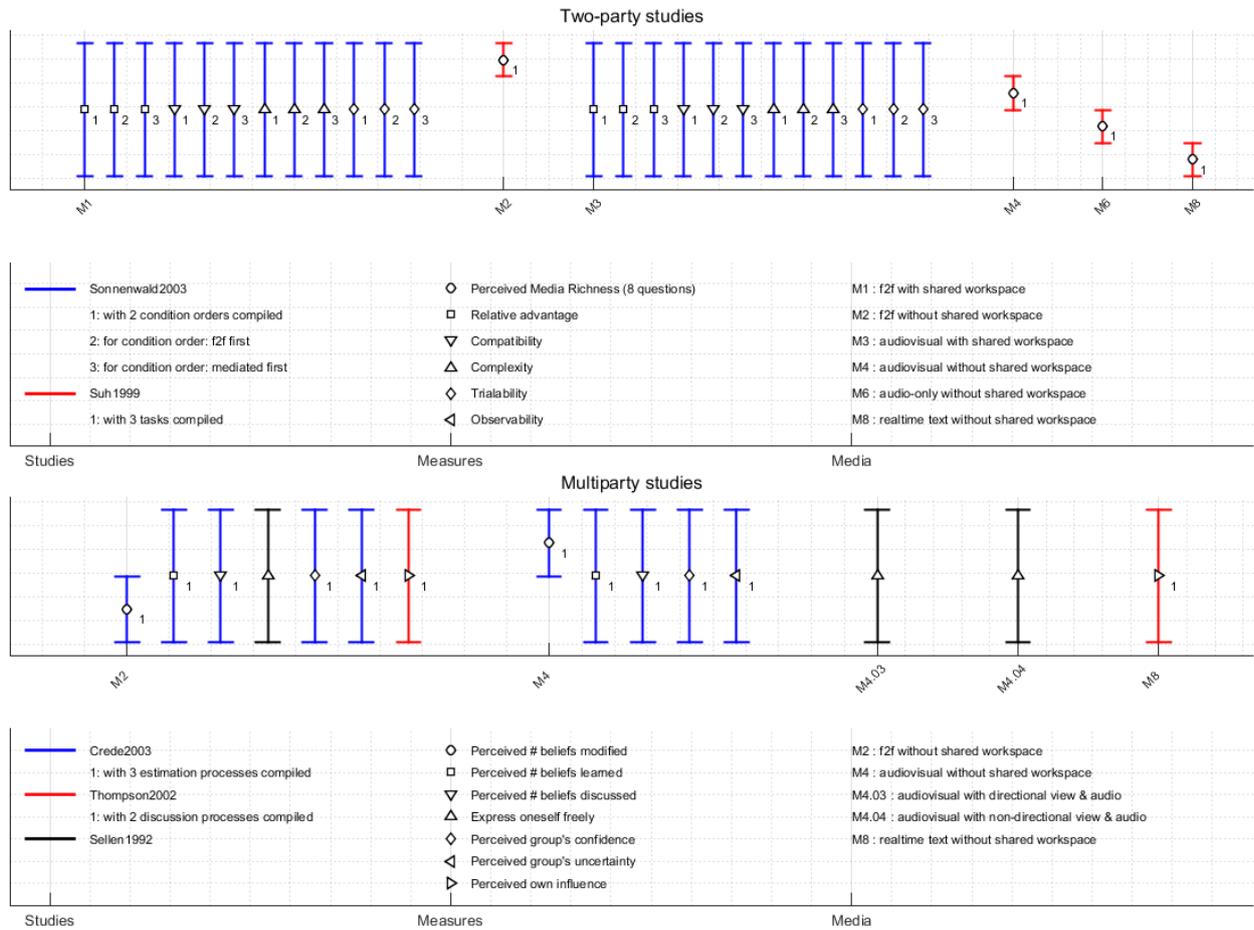
**Main finding for two-party studies:**

There is no simple ranking of media across studies, even after grouping into rough categories, i.e. f2f ( $M_1, M_2$ ), audiovisual ( $M_{3.x}, M_{4.x}$ ), audio-only ( $M_{5.x}, M_{6.x}$ ), and text-based ( $M_7, M_8$ ): Either no significant differences were found (large errorbars within a medium), or any significant rankings within certain studies were not found in other studies (mix of small and large errorbars within a medium), or any significant rankings were contrary between studies (small errorbars are spread within a medium).

**Main finding for multiparty studies:**

Although the amount of data is small, some trend is visible: f2f and audiovisual leads, with some deviations, to higher satisfaction than audio-only and text-only.

Figure B.15: Impact of communication media on satisfaction measures across studies: Higher values represent higher satisfaction.



**Interpretation Method:**

1. Look at the positions and sizes of the errorbars per medium  $Mx$ .
2. The more the errorbars for the medium  $Mx$  are clustered towards the upper end, the more likely  $Mx$  leads to an increased media richness and opinion exchange compared to other media. Example: none in this plot, too little data.
3. The more the errorbars are clustered toward the lower end, the more likely  $Mx$  leads to a decreased media richness and opinion exchange. Example: none in this plot, too little data.
4. The more centrally positioned and at the same time the smaller the errorbars are for the medium  $Mx$ , the more likely  $Mx$  has an average media richness and opinion exchange compared to other media. Example: none in this plot, too little data.
5. The more centrally positioned and at the same time the wider the errorbars are for the medium  $Mx$ , the more likely  $Mx$  does not differ in media richness and opinion exchange compared to other media. Example: essentially all media for the two-party and multiparty studies.
6. The more the errorbars of individual studies are spread for medium  $Mx$ , the more likely the impact of  $Mx$  on media richness and opinion exchange depends on the specific characteristics of the corresponding studies. Example: none in this plot, too little data.

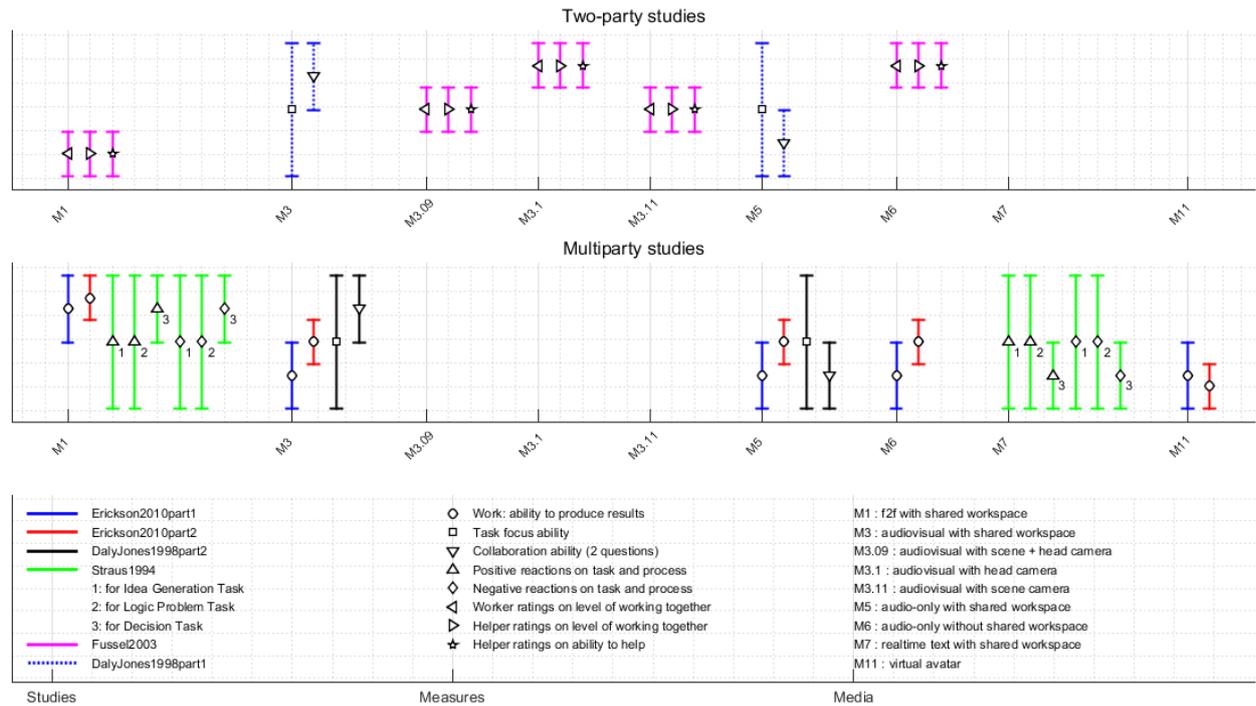
**Main finding for two-party studies concerning media richness:**

None since there is too little data, i.e. just one or two data points (studies) per medium and variable, to draw proper conclusions.

**Main finding for multiparty studies concerning opinion-related measures:**

There is very little data, i.e. three data points (studies) per medium or per rough category of media (f2f, audiovisual, text), but these suggest no effect of the communication medium.

Figure B.16: Impact of communication media on media richness measures (two top panels) and opinion-related measures (two bottom panels) across studies: Higher values represent higher media richness and higher change and exchange of opinions.



**Interpretation Method:**

1. Look at the positions and sizes of the errorbars per medium  $M_x$ .
2. The more the errorbars for the medium  $M_x$  are clustered towards the upper end, the more likely  $M_x$  leads to an increased collaboration compared to other media. Example:  $M1$  for multiparty studies, in tendency.
3. The more the errorbars are clustered toward the lower end, the more likely  $M_x$  leads to a decreased collaboration. Example:  $M7$  for multiparty studies, in tendency.
4. The more centrally positioned and at the same time the smaller the errorbars are for the medium  $M_x$ , the more likely  $M_x$  has an average collaboration compared to other media. Example: none in this plot.
5. The more centrally positioned and at the same time the wider the errorbars are for the medium  $M_x$ , the more likely  $M_x$  does not differ in collaboration compared to other media. Example: none in this plot.
6. The more the errorbars of individual studies are spread for medium  $M_x$ , the more likely the impact of  $M_x$  on collaboration depends on the specific characteristics of the corresponding studies. Example:  $M3$  for multiparty studies.

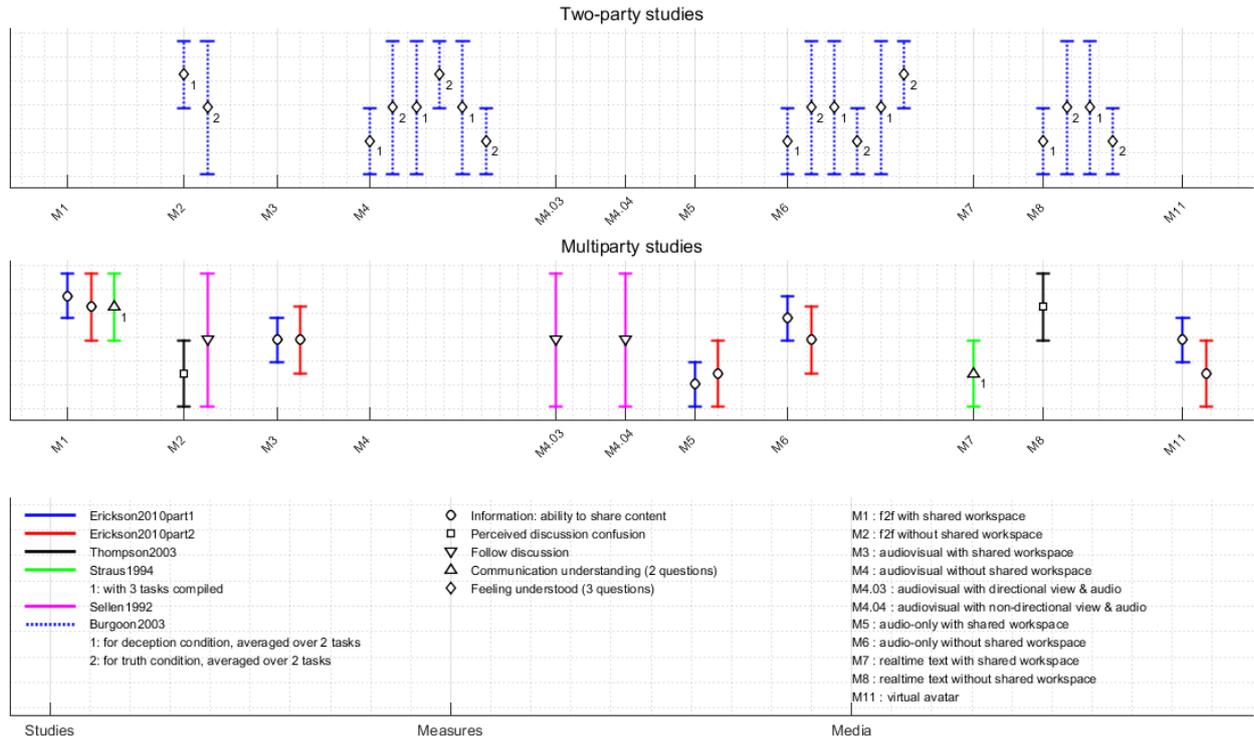
**Main finding for two-party studies:**

None since there is too little data, i.e. just one or two data points (studies) per medium and variable, to draw proper conclusions.

**Main finding for multiparty studies:**

There is very little data, i.e. three data points (studies) per medium or per rough category of media (f2f, audiovisual, audio-only, text, avatar). Furthermore, no clear ranking of media across studies is visible: Either any significant rankings within certain studies were not found in other studies (mix of small and large errorbars within a medium), or any significant rankings were contrary between studies (small errorbars are spread within a medium).

Figure B.17: Impact of communication media on work process measures across studies: Higher values represent more positive work processes and higher amount of collaboration.



**Interpretation Method:**

1. Look at the positions and sizes of the errorbars per medium  $M_x$ .
2. The more the errorbars for the medium  $M_x$  are clustered towards the upper end, the more likely  $M_x$  leads to an increased grounding compared to other media. Example:  $M1$  for multiparty studies.
3. The more the errorbars are clustered toward the lower end, the more likely  $M_x$  leads to a decreased grounding. Example:  $M5$  for multiparty studies.
4. The more centrally positioned and at the same time the smaller the errorbars are for the medium  $M_x$ , the more likely  $M_x$  has an average grounding compared to other media. Example:  $M3$  for multiparty studies.
5. The more centrally positioned and at the same time the wider the errorbars are for the medium  $M_x$ , the more likely  $M_x$  does not differ in grounding compared to other media. Example: none in this plot.
6. The more the errorbars of individual studies are spread for medium  $M_x$ , the more likely the impact of  $M_x$  on grounding depends on the specific characteristics of the corresponding studies. Example:  $M4$  for two-party studies.

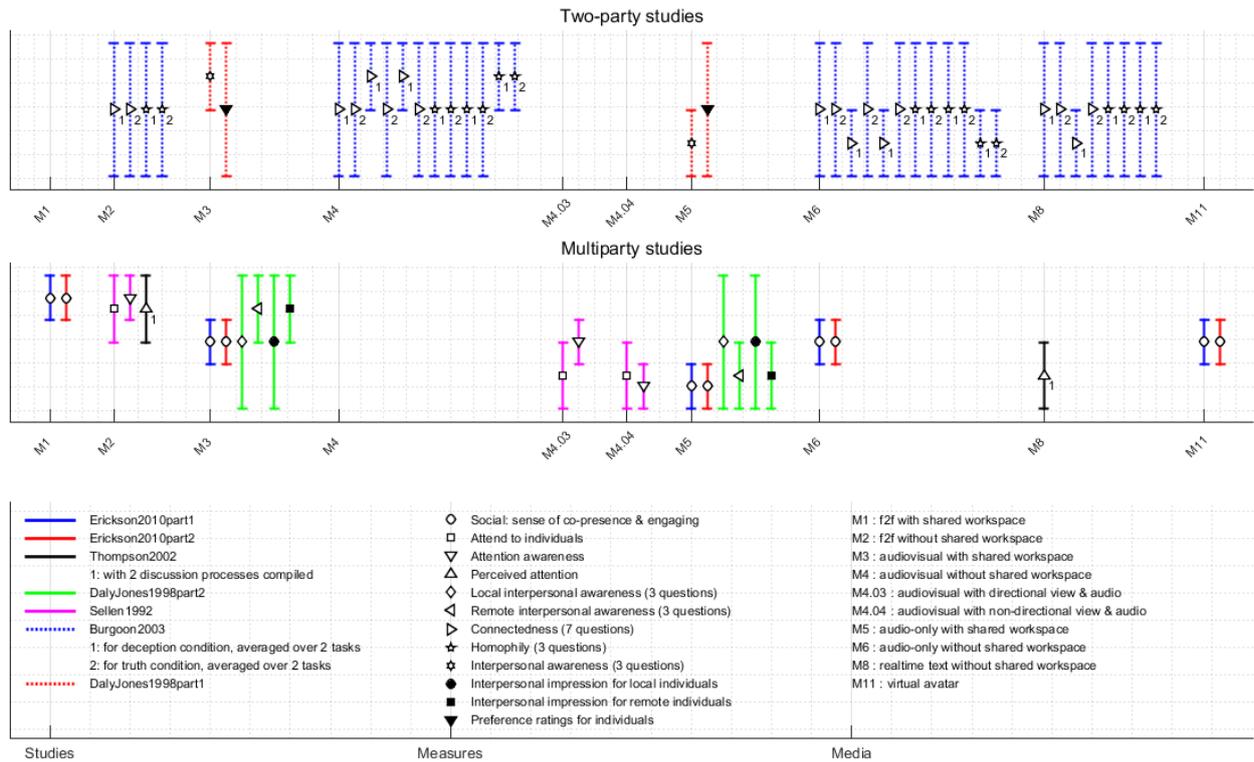
**Main finding for two-party studies:**

None since there is too little data, i.e. just one data point (study) to draw proper conclusions.

**Main finding for multiparty studies:**

None since there is very little data, i.e. just one to three data points (studies) per medium and no clear ranking of media across studies is visible.

Figure B.18: Impact of communication media on grounding-related measures across studies: Higher values represent positive aspects of grounding, except for one variable (*Perceived Discussion Confusion*).



**Interpretation Method:**

1. Look at the positions and sizes of the errorbars per medium  $M_x$ .
2. The more the errorbars for the medium  $M_x$  are clustered towards the upper end, the more likely  $M_x$  leads to an increased impression of the interlocutor compared to other media. Example:  $M_2$  for multiparty studies.
3. The more the errorbars are clustered toward the lower end, the more likely  $M_x$  leads to a decreased impression of the interlocutor. Example:  $M_5$  for multiparty studies, in tendency.
4. The more centrally positioned and at the same time the smaller the errorbars are for the medium  $M_x$ , the more likely  $M_x$  has an average impression of the interlocutor compared to other media. Example:  $M_6$  for multiparty studies.
5. The more centrally positioned and at the same time the wider the errorbars are for the medium  $M_x$ , the more likely  $M_x$  does not differ in impression of the interlocutor compared to other media. Example:  $M_2$  for two-party studies.
6. The more the errorbars of individual studies are spread for medium  $M_x$ , the more likely the impact of  $M_x$  on impression of the interlocutor depends on the specific characteristics of the corresponding studies. Example:  $M_3$  for multiparty studies, in tendency.

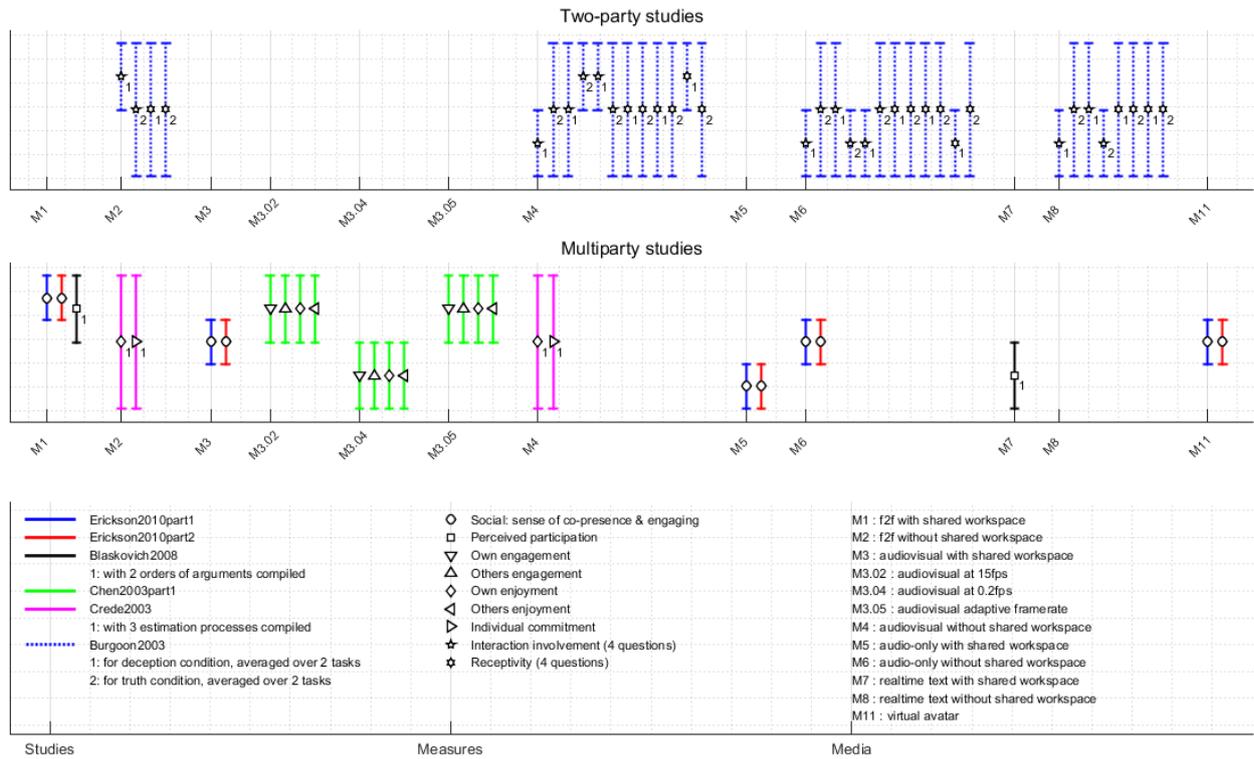
**Main finding for two-party studies:**

None since there is too little data, i.e. just one data point (study) to draw proper conclusions.

**Main finding for multiparty studies:**

There is some trend that the interlocutors are more positively perceived by study participants for f2f communication vs. audio-only and other media, with audiovisual communication in-between.

Figure B.19: Impact of communication media on measures of interlocutor behavior across studies: Higher values represent higher amount of attention, awareness and impression of the interlocutor.



Interpretation Method:

1. Look at the positions and sizes of the errorbars per medium  $M_x$ .
2. The more the errorbars for the medium  $M_x$  are clustered towards the upper end, the more likely  $M_x$  leads to an increased engagement compared to other media. Example:  $M_1$  for multiparty studies.
3. The more the errorbars are clustered toward the lower end, the more likely  $M_x$  leads to a decreased engagement. Example:  $M_5$  for multiparty studies.
4. The more centrally positioned and at the same time the smaller the errorbars are for the medium  $M_x$ , the more likely  $M_x$  has an average engagement compared to other media. Example:  $M_3$  for multiparty studies.
5. The more centrally positioned and at the same time the wider the errorbars are for the medium  $M_x$ , the more likely  $M_x$  does not differ in engagement compared to other media. Example:  $M_2$  for two-party and multiparty studies.
6. The more the errorbars of individual studies are spread for medium  $M_x$ , the more likely the impact of  $M_x$  on engagement depends on the specific characteristics of the corresponding studies. Example:  $M_4$  for two-party studies, in tendency.

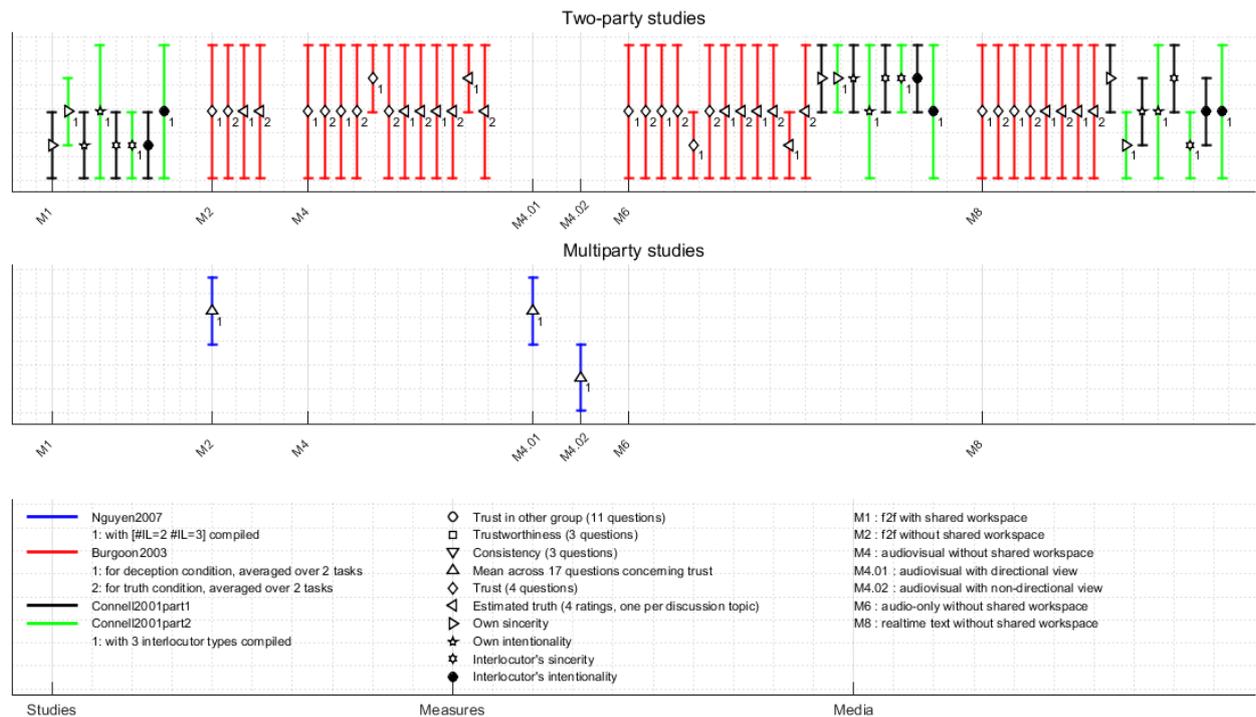
Main finding for two-party studies:

None since there is too little data, i.e. just one data point (study) to draw proper conclusions.

Main finding for multiparty studies:

None since there is very little data, i.e. just one to three data points (studies) per medium and no clear ranking of media across studies is visible.

Figure B.20: Impact of communication media on engagement measures across studies: Higher values represent engagement.



**Interpretation Method:**

1. Look at the positions and sizes of the errorbars per medium  $M_x$ .
2. The more the errorbars for the medium  $M_x$  are clustered towards the upper end, the more likely  $M_x$  leads to an increased trust compared to other media. Example: none in this plot.
3. The more the errorbars are clustered toward the lower end, the more likely  $M_x$  leads to a decreased trust. Example: M1 for two-party studies, in tendency.
4. The more centrally positioned and at the same time the smaller the errorbars are for the medium  $M_x$ , the more likely  $M_x$  has an average trust compared to other media. Example: none in this plot.
5. The more centrally positioned and at the same time the wider the errorbars are for the medium  $M_x$ , the more likely  $M_x$  does not differ in trust compared to other media. Example: M2 for two-party studies.
6. The more the errorbars of individual studies are spread for medium  $M_x$ , the more likely the impact of  $M_x$  on trust depends on the specific characteristics of the corresponding studies. Example: M8 for two-party studies.

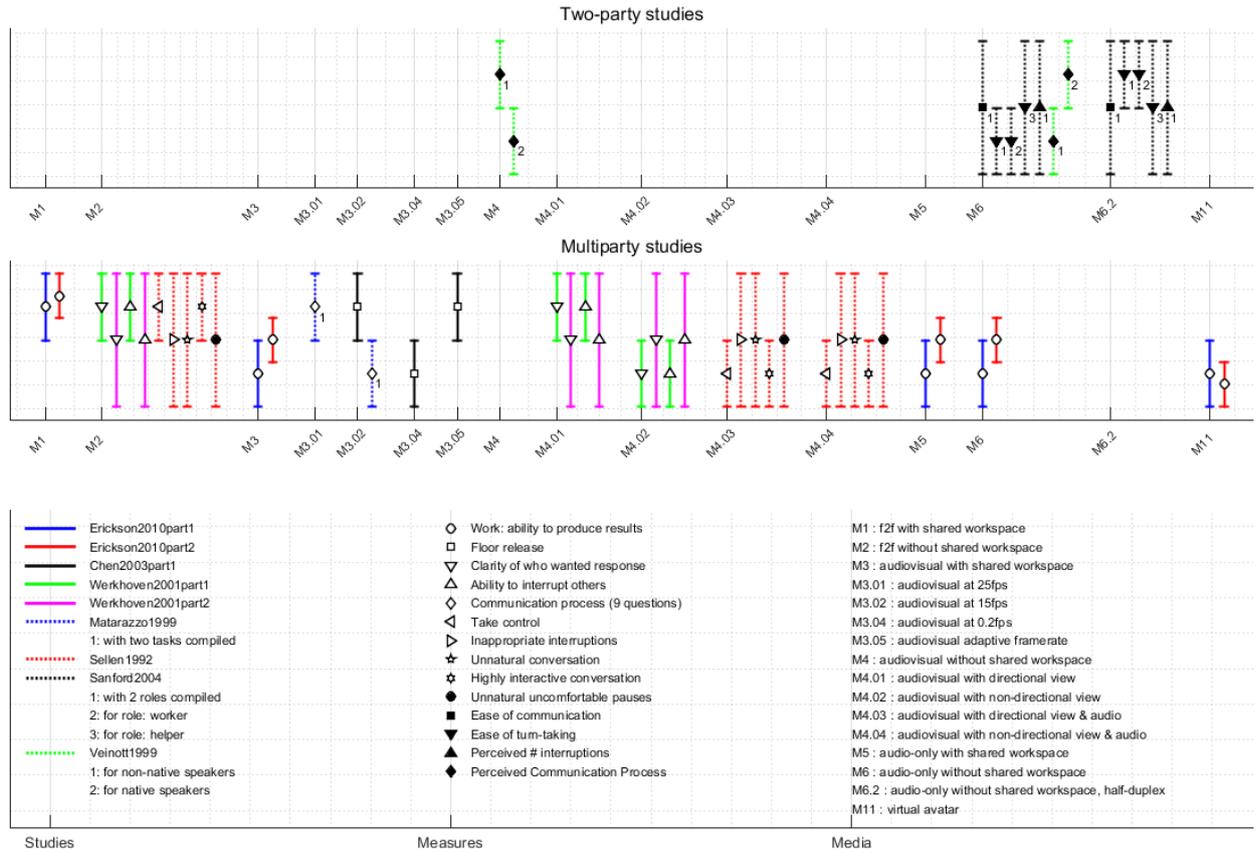
**Main finding for two-party studies:**

None since there is too little data, i.e. just one or two data points (studies) to draw proper conclusions.

**Main finding for multiparty studies:**

None since there is too little data, i.e. just one data point (study) to draw proper conclusions.

Figure B.21: Impact of communication media on trust-related measures across studies: Higher values represent higher amount of trust.



**Interpretation Method:**

1. Look at the positions and sizes of the errorbars per medium  $M_x$ .
2. The more the errorbars for the medium  $M_x$  are clustered towards the upper end, the more likely  $M_x$  leads to an increased turn-taking compared to other media. Example: M1 for multiparty studies.
3. The more the errorbars are clustered toward the lower end, the more likely  $M_x$  leads to a decreased turn-taking. Example: M11 for multiparty studies.
4. The more centrally positioned and at the same time the smaller the errorbars are for the medium  $M_x$ , the more likely  $M_x$  has an average turn-taking compared to other media. Example: none in this plot.
5. The more centrally positioned and at the same time the wider the errorbars are for the medium  $M_x$ , the more likely  $M_x$  does not differ in turn-taking compared to other media. Example: none in this plot.
6. The more the errorbars of individual studies are spread for medium  $M_x$ , the more likely the impact of  $M_x$  on turn-taking depends on the specific characteristics of the corresponding studies. Example: M6 for two-party studies.

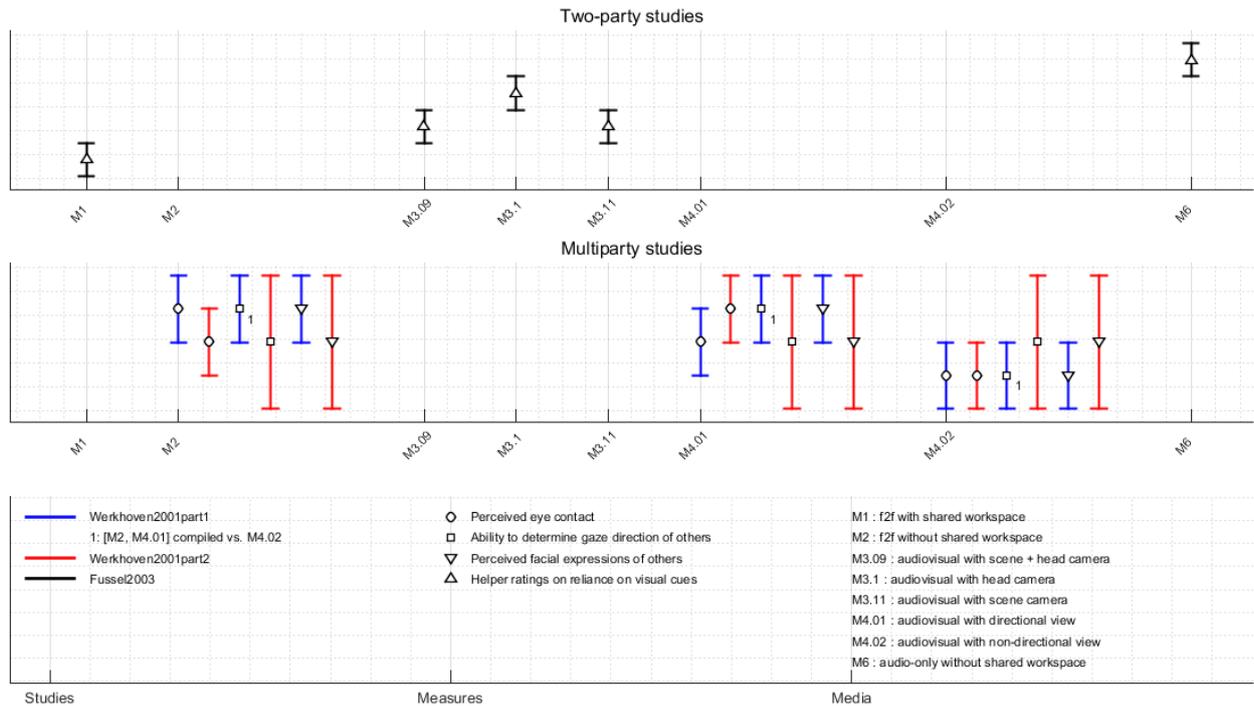
**Main finding for two-party studies:**

None since there is too little data, i.e. just one or two data points (studies) to draw proper conclusions.

**Main finding for multiparty studies:**

There is no simple ranking of media across studies, even after grouping into rough categories, i.e. f2f (M1, M2), audiovisual (M3.x, M4.x), audio-only (M5, M6), and other (M11): Either any significant rankings within certain studies were not found in other studies (mix of small and large errorbars within a medium), or any significant rankings were contrary between studies (small errorbars are spread within a medium).

Figure B.22: Impact of communication media on turn-taking-related measures across studies: Higher values represent more positive perception of turn-taking.



**Interpretation Method:**

1. Look at the positions and sizes of the errorbars per medium  $M_x$ .
2. The more the errorbars for the medium  $M_x$  are clustered towards the upper end, the more likely  $M_x$  leads to an increased benefit of visual cues compared to other media. Example:  $M4.01$  for multiparty studies, in tendency.
3. The more the errorbars are clustered toward the lower end, the more likely  $M_x$  leads to a decreased benefit of visual cues. Example:  $M4.02$  for multiparty studies, in tendency.
4. The more centrally positioned and at the same time the smaller the errorbars are for the medium  $M_x$ , the more likely  $M_x$  has an average benefit of visual cues compared to other media. Example: none in this plot.
5. The more centrally positioned and at the same time the wider the errorbars are for the medium  $M_x$ , the more likely  $M_x$  does not differ in benefit of visual cues compared to other media. Example: none in this plot.
6. The more the errorbars of individual studies are spread for medium  $M_x$ , the more likely the impact of  $M_x$  on benefit of visual cues depends on the specific characteristics of the corresponding studies. Example: none in this plot.

**Main finding for two-party studies:**

None since there is too little data, i.e. just one data point (study) to draw proper conclusions.

**Main finding for multiparty studies:**

None since there is too little data, i.e. just two data points (studies) to draw proper conclusions.

Figure B.23: Impact of communication media on measures concerning visual cues across studies: Higher values represent a larger perceived benefit of visual cues.

# C

## *Detailed Results: Study on Impact of Communication Complexity*

### *What this chapter is about*

The present appendix presents per employed measure the detailed results of the two experiments CC1 & CC2 on the impact of communication complexity on quality perception. First, each figure provides an explanation of the analysis steps conducted. Then, each figure shows the results for two analysis levels: a) the main effects (i.e. effect of *#IL* across all *SndRepr*, effect of *SndRepr* across all *#IL*), and (b) the individual effects (effect of *#IL* per *SndRepr*, effect of *SndRepr* per *#IL*). Finally, each figure discusses how the current results relate to the different hypotheses formulated in Section 7.3. The individual results are then compiled in the main text in Section 7.3.

### C.1 Experiment CC1

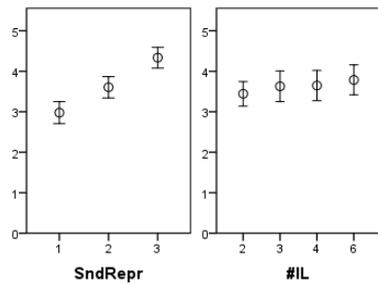
#### 1. Analysis Steps

The present figure shows the data analysis for the measure *Quality* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of *#IL* if all data over *SndRepr* is compiled and the effect of *SndRepr* if all data over *#IL* is compiled. Steps (d) to (f) concern the individual effect of *#IL* analyzed per *SndRepr*. Steps (g) to (i) concern the individual effect of *SndRepr* analyzed per *#IL*.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per *#IL* and *SndRepr*. Steps (b), (e), & (h): results of repeated-measures ANOVAs with *#IL* ∈ [2,3,4,6] and *SndRepr* ∈ [1 = NB/non-spatial, 2 = FB/non-spatial, 3 = FB/spatial] as input, expressed by the *F*-measure and significance level *p*, and using the Greenhouse-Geisser correction when the sphericity criterion for repeated-measures ANOVAs is violated. Steps (c), (f), & (i): results of PostHoc tests (estimated marginal means) with Sidak correction for pairwise comparisons between all pairs of *#IL* or *SndRepr*, expressed by the significance level *p*; Significant differences (*p* ≤ 0.05) are indicated with an asterisk (\*).

#### 2.1 Main Effects

(a) Errorbars



(b) ANOVA

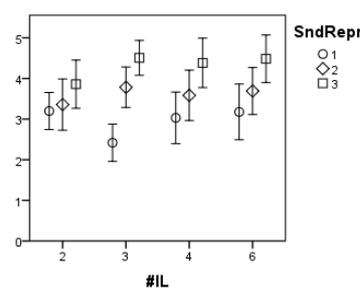
	<i>F</i>	<i>p</i>
<i>#IL</i>	0.147	0.790
<i>SndRepr</i>	22.592	0.000*
Interaction	2.125	0.067

(c) PostHoc tests

<i>SndRepr</i>	<i>p</i>	<i>#IL</i>	<i>p</i>
1 vs. 2	0.127	2 vs. 3	1.000
1 vs. 3	0.001*	2 vs. 4	1.000
2 vs. 3	0.004*	2 vs. 6	0.998
		3 vs. 4	1.000
		3 vs. 6	0.999
		4 vs. 6	0.788

#### 2.2 Individual Effect of *SndRepr* per *#IL*

(d) Errorbars



(e) ANOVA per *#IL*

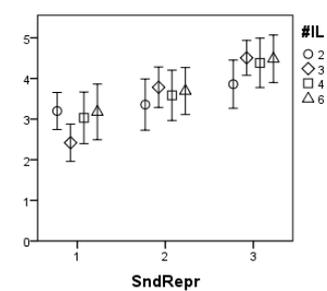
<i>#IL</i>	<i>F</i>	<i>p</i>
2	4.576	0.025*
3	34.681	0.000*
4	10.376	0.001*
6	10.628	0.001*

(f) PostHoc tests per *#IL*

<i>#IL</i>	<i>SndRepr</i>	<i>p</i>
2	1 vs. 2	0.951
	1 vs. 3	0.021*
	2 vs. 3	0.088
3	1 vs. 2	0.005*
	1 vs. 3	0.000*
	2 vs. 3	0.000*
4	1 vs. 2	0.448
	1 vs. 3	0.005*
	2 vs. 3	0.009*
6	1 vs. 2	0.172
	1 vs. 3	0.004*
	2 vs. 3	0.056

#### 2.3 Individual Effect of *#IL* per *SndRepr*

(g) Errorbars



(h) ANOVA per *SndRepr*

<i>SndRepr</i>	<i>F</i>	<i>p</i>
1	2.595	0.071
2	1.405	0.259
3	0.839	0.484

(i) PostHoc tests per *SndRepr*

<i>SndRepr</i>	<i>#IL</i>	<i>p</i>
1	2 vs. 3	0.104
	2 vs. 4	0.993
	2 vs. 6	0.999
	3 vs. 4	0.613
2	3 vs. 6	0.342
	4 vs. 6	0.997
	2 vs. 3	0.509
	2 vs. 4	0.998
	2 vs. 6	0.904
	3 vs. 4	0.336
3	3 vs. 6	0.999
	4 vs. 6	0.841
	2 vs. 3	0.281
	2 vs. 4	0.980
	2 vs. 6	0.709
	3 vs. 4	0.999
	3 vs. 6	1.000
	4 vs. 6	1.000

#### 3. Analysis Results:

3.1 Hypothesis H5 (Communication Complexity has an impact on Speech Communication Quality / Quality of Experience) is not supported, as *#IL* has no significant impact on *Quality*.

Rationale – main effect: Although values show a slight tendency to increase with increasing *#IL* (Step a), this unexpected effect is not significant (Steps b & c).

Rationale – effect per *SndRepr*: Although values show a slight tendency to increase with increasing *#IL* for all three *SndRepr* (Step g), this unexpected effect is not significant (Steps h & i).

3.2 Hypothesis H6 (Technical System Capability has an impact on Speech Communication Quality /Quality of Experience) is supported, as *SndRepr* has a significant impact on *Quality*.

Rationale – main effect: Values increase with increasing *SndRepr* (Step a). This effect is significant (Step b) and explained by significant differences between non-spatial and spatial sound reproduction (Step c).

Rationale – effect per *#IL*: Values increase for all four *#IL* (Step d). This is significant (Step e) explained by significant differences between almost all *SndRepr*, most often between non-spatial and spatial round reproduction (Step f).

Figure C.1: Results for the measure *Quality*.

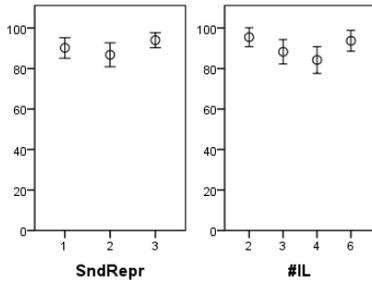
1. Analysis Steps

The present figure shows the data analysis for the measure *Topic Recognition Performance* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of #IL if all data over *SndRepr* is compiled and the effect of *SndRepr* if all data over #IL is compiled. Steps (d) to (f) concern the individual effect of #IL analyzed per *SndRepr*. Steps (g) to (i) concern the individual effect of *SndRepr* analyzed per #IL.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per #IL and *SndRepr*. Steps (b), (e), & (h): results of repeated-measures ANOVAs with #IL ∈ [2,3,4,6] and *SndRepr* ∈ [1 = NB/non-spatial, 2 = FB/non-spatial, 3 = FB/spatial] as input, expressed by the *F*-measure and significance level *p*, and using the Greenhouse-Geisser correction when the sphericity criterion for repeated-measures ANOVAs is violated. Steps (c), (f), & (i): results of PostHoc tests (estimated marginal means) with Sidak correction for pairwise comparisons between all pairs of #IL or *SndRepr*, expressed by the significance level *p*; Significant differences (*p* ≤ 0.05) are indicated with an asterisk (\*).

2.1 Main Effects

(a) Errorbars



(b) ANOVA

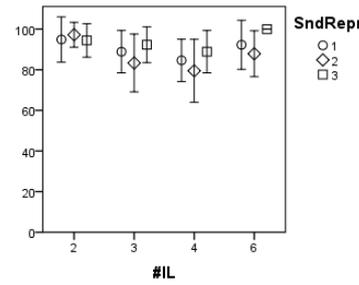
	<i>F</i>	<i>p</i>
#IL	2.862	0.081
<i>SndRepr</i>	2.995	0.046*
Interaction	0.738	0.492

(c) PostHoc tests

<i>SndRepr</i>	<i>p</i>	#IL	<i>p</i>
1 vs. 2	0.889	2 vs. 3	0.629
1 vs. 3	0.346	2 vs. 4	0.279
2 vs. 3	0.023*	2 vs. 6	0.985
		3 vs. 4	0.949
		3 vs. 6	0.687
		4 vs. 6	0.207

2.2 Individual Effect of *SndRepr* per #IL

(d) Errorbars



(e) ANOVA per #IL

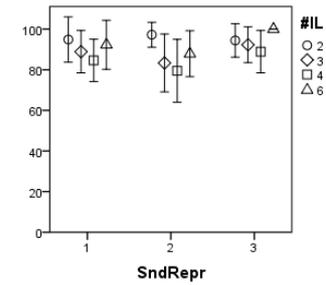
#IL	<i>F</i>	<i>p</i>
2	0.478	0.626
3	0.865	0.402
4	0.517	0.604
6	1.733	0.202

(f) PostHoc tests per #IL

#IL	<i>SndRepr</i>	<i>p</i>
2	1 vs. 2	0.711
	1 vs. 3	1.000
	2 vs. 3	0.711
3	1 vs. 2	0.711
	1 vs. 3	0.929
	2 vs. 3	0.711
4	1 vs. 2	0.823
	1 vs. 3	0.979
	2 vs. 3	0.814
6	1 vs. 2	0.979
	1 vs. 3	0.473
	2 vs. 3	0.110

2.3 Individual Effect of #IL per *SndRepr*

(g) Errorbars



(h) ANOVA per *SndRepr*

<i>SndRepr</i>	<i>F</i>	<i>p</i>
1	1.000	0.405
2	1.892	0.152
3	1.774	0.171

(i) PostHoc tests per *SndRepr*

<i>SndRepr</i>	#IL	<i>p</i>
1	2 vs. 3	0.916
	2 vs. 4	0.720
	2 vs. 6	0.916
	3 vs. 4	0.995
2	3 vs. 6	0.995
	4 vs. 6	0.916
	2 vs. 3	0.279
	2 vs. 4	0.400
3	2 vs. 6	0.722
	3 vs. 4	1.000
	3 vs. 6	0.985
	4 vs. 6	0.722
	2 vs. 3	0.995
	2 vs. 4	0.916
	2 vs. 6	0.664
	3 vs. 4	0.999
3 vs. 6	0.401	
4 vs. 6	0.211	

3. Analysis Results:

3.1 Hypothesis H1 (Communication Complexity has an impact on Cognitive Load) is not supported, as #IL has no significant impact on *Topic Recognition Performance*.

Rationale – main effect: With the exception of #IL = 6, values decrease with increasing #IL (Step a). This unexpected effect is not significant (Steps b & c).

Rationale – effect per *SndRepr*: With the exception of #IL = 6, values decrease with increasing #IL for all three *SndRepr* (Step g). However, this is not significant (Steps h & i).

3.2 Hypothesis H2 (Technical System Capability has an impact on Cognitive Load) is supported, as *SndRepr* has a significant impact on *Topic Recognition Performance*.

Rationale – main effect: Values for spatial sound reproduction are higher than non-spatial reproduction (Step a). This effect significant (Step b), but only between *SndRepr* = 2 & 3 (Step c).

Rationale – effect per #IL: No clear pattern is visible (Step d), and no significant differences can be found (Steps e & f).

Figure C.2: Results for the measure *Topic Recognition Performance*.

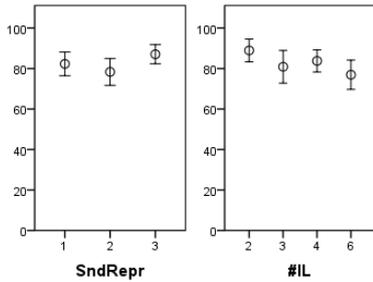
1. Analysis Steps

The present figure shows the data analysis for the measure *Topic Recognition Confidence* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of #IL if all data over *SndRepr* is compiled and the effect of *SndRepr* if all data over #IL is compiled. Steps (d) to (f) concern the individual effect of #IL analyzed per *SndRepr*. Steps (g) to (i) concern the individual effect of *SndRepr* analyzed per #IL.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per #IL and *SndRepr*. Steps (b), (e), & (h): results of repeated-measures ANOVAs with #IL ∈ [2,3,4,6] and *SndRepr* ∈ [1 = NB/non-spatial, 2 = FB/non-spatial, 3 = FB/spatial] as input, expressed by the *F*-measure and significance level *p*, and using the Greenhouse-Geisser correction when the sphericity criterion for repeated-measures ANOVAs is violated. Steps (c), (f), & (i): results of PostHoc tests (estimated marginal means) with Sidak correction for pairwise comparisons between all pairs of #IL or *SndRepr*, expressed by the significance level *p*; Significant differences (*p* ≤ 0.05) are indicated with an asterisk (\*).

2.1 Main Effects

(a) Errorbars



(b) ANOVA

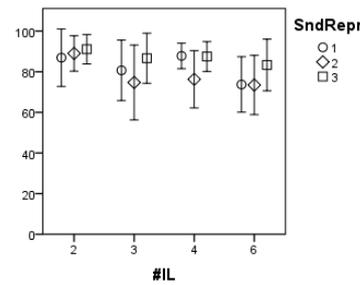
	<i>F</i>	<i>p</i>
#IL	3.114	0.039*
<i>SndRepr</i>	3.877	0.036*
Interaction	0.724	0.521

(c) PostHoc tests

<i>SndRepr</i>	<i>p</i>	#IL	<i>p</i>
1 vs. 2	0.680	2 vs. 3	0.344
1 vs. 3	0.310	2 vs. 4	0.639
2 vs. 3	0.018*	2 vs. 6	0.128
		3 vs. 4	0.993
		3 vs. 6	0.899
		4 vs. 6	0.675

2.2 Individual Effect of *SndRepr* per #IL

(d) Errorbars



(e) ANOVA per #IL

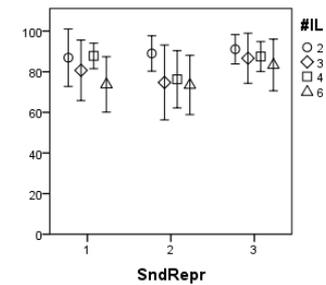
#IL	<i>F</i>	<i>p</i>
2	0.139	0.871
3	1.212	0.317
4	3.341	0.085
6	1.583	0.228

(f) PostHoc tests per #IL

#IL	<i>SndRepr</i>	<i>p</i>
2	1 vs. 2	1.000
	1 vs. 3	0.963
	2 vs. 3	0.876
3	1 vs. 2	0.771
	1 vs. 3	0.710
	2 vs. 3	0.575
4	1 vs. 2	0.200
	1 vs. 3	0.998
	2 vs. 3	0.264
6	1 vs. 2	1.000
	1 vs. 3	0.298
	2 vs. 3	0.409

2.3 Individual Effect of #IL per *SndRepr*

(g) Errorbars



(h) ANOVA per *SndRepr*

<i>SndRepr</i>	<i>F</i>	<i>p</i>
1	3.076	0.041*
2	1.814	0.164
3	0.430	0.733

(i) PostHoc tests per *SndRepr*

<i>SndRepr</i>	#IL	<i>p</i>
1	2 vs. 3	0.731
	2 vs. 4	1.000
	2 vs. 6	0.241
	3 vs. 4	0.884
2	3 vs. 6	0.355
	4 vs. 6	0.170
	2 vs. 3	0.407
	2 vs. 4	0.188
3	2 vs. 6	0.162
	3 vs. 4	1.000
	3 vs. 6	1.000
	4 vs. 6	1.000
	2 vs. 3	0.972
	2 vs. 4	0.774
	2 vs. 6	0.851
	3 vs. 4	1.000
3 vs. 6	1.000	
4 vs. 6	0.998	

3. Analysis Results:

3.1 Hypothesis H1 (Communication Complexity has an impact on Cognitive Load) is supported, as #IL has a significant impact on *Topic Recognition Confidence*.

Rationale – main effect: Values decrease with increasing #IL (Step a). This effect is significant (Step b), but not strong enough to be detected by the PostHoc tests (Step c).

Rationale – effect per *SndRepr*: Values decrease with increasing #IL for all three *SndRepr* (Step g). This is significant only for *SndRepr* = 1 (Step h), but it can not be detected by the PostHoc tests (Step i).

3.2 Hypothesis H2 (Technical System Capability has an impact on Cognitive Load) is supported, as *SndRepr* has a significant impact on *Topic Recognition Confidence*.

Rationale – main effect: Values for spatial sound reproduction are higher than non-spatial reproduction (Step a). This effect significant (Step b), but only between *SndRepr* = 2 & 3 (Step c).

Rationale – effect per #IL: No clear pattern is visible (Step d), and no significant differences can be found (Steps e & f).

Figure C.3: Results for the measure *Topic Recognition Confidence*.

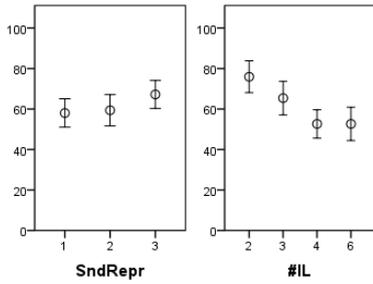
1. Analysis Steps

The present figure shows the data analysis for the measure *Speaker Recognition Performance* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of #IL if all data over *SndRepr* is compiled and the effect of *SndRepr* if all data over #IL is compiled. Steps (d) to (f) concern the individual effect of #IL analyzed per *SndRepr*. Steps (g) to (i) concern the individual effect of *SndRepr* analyzed per #IL.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per #IL and *SndRepr*. Steps (b), (e), & (h): results of repeated-measures ANOVAs with #IL ∈ [2,3,4,6] and *SndRepr* ∈ [1 = NB/non-spatial, 2 = FB/non-spatial, 3 = FB/spatial] as input, expressed by the *F*-measure and significance level *p*, and using the Greenhouse-Geisser correction when the sphericity criterion for repeated-measures ANOVAs is violated. Steps (c), (f), & (i): results of PostHoc tests (estimated marginal means) with Sidak correction for pairwise comparisons between all pairs of #IL or *SndRepr*, expressed by the significance level *p*; Significant differences (*p* ≤ 0.05) are indicated with an asterisk (\*).

2.1 Main Effects

(a) Errorbars



(b) ANOVA

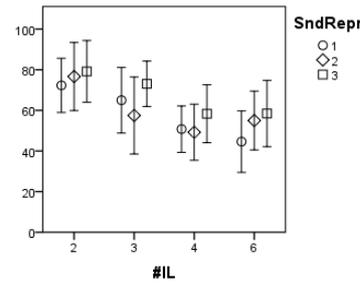
	<i>F</i>	<i>p</i>
#IL	27.750	0.000*
<i>SndRepr</i>	2.717	0.088
Interaction	0.842	0.542

(c) PostHoc tests

<i>SndRepr</i>	<i>p</i>	#IL	<i>p</i>
1 vs. 2	1.000	2 vs. 3	0.000*
1 vs. 3	0.033*	2 vs. 4	0.000*
2 vs. 3	0.320	2 vs. 6	0.000*
		3 vs. 4	0.025*
		3 vs. 6	0.003*
		4 vs. 6	1.000

2.2 Individual Effect of *SndRepr* per #IL

(d) Errorbars



(e) ANOVA per #IL

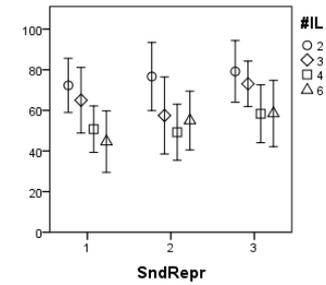
#IL	<i>F</i>	<i>p</i>
2	0.064	0.938
3	2.130	0.168
4	0.733	0.492
6	1.917	0.171

(f) PostHoc tests per #IL

#IL	<i>SndRepr</i>	<i>p</i>
2	1 vs. 2	0.999
	1 vs. 3	0.994
	2 vs. 3	0.982
3	1 vs. 2	0.915
	1 vs. 3	0.175
	2 vs. 3	0.140
4	1 vs. 2	0.998
	1 vs. 3	0.801
	2 vs. 3	0.567
6	1 vs. 2	0.694
	1 vs. 3	0.228
	2 vs. 3	0.741

2.3 Individual Effect of #IL per *SndRepr*

(g) Errorbars



(h) ANOVA per *SndRepr*

<i>SndRepr</i>	<i>F</i>	<i>p</i>
1	6.103	0.002*
2	6.455	0.001*
3	5.845	0.003*

(i) PostHoc tests per *SndRepr*

<i>SndRepr</i>	#IL	<i>p</i>
1	2 vs. 3	0.646
	2 vs. 4	0.003*
	2 vs. 6	0.013*
	3 vs. 4	0.612
	3 vs. 6	0.363
	4 vs. 6	0.970
2	2 vs. 3	0.233
	2 vs. 4	0.002*
	2 vs. 6	0.048*
	3 vs. 4	0.941
	3 vs. 6	0.999
	4 vs. 6	0.979
3	2 vs. 3	0.998
	2 vs. 4	0.140
	2 vs. 6	0.109
	3 vs. 4	0.039*
	3 vs. 6	0.055
	4 vs. 6	0.987

3. Analysis Results:

3.1 Hypothesis H1 (Communication Complexity has an impact on Cognitive Load) is supported, as #IL has a significant impact on *Speaker Recognition Performance*.

Rationale – main effect: Values decrease with increasing #IL (Step a). This effect is significant (Step b), and explained by significant differences between all expect one pair of #IL (Step c).

Rationale – effect per *SndRepr*: Values decrease with increasing #IL for all three *SndRepr* (Step g). This is significant (Step h), and is mainly explained by significant differences between #IL =2 and #IL > 3, i.e. two-party and medium-large multiparty party groups (Step i).

3.2 Hypothesis H2 (Technical System Capability has an impact on Cognitive Load) is not supported, as *SndRepr* has no significant impact on *Speaker Recognition Performance*.

Rationale – main effect: Although values for spatial sound reproduction are higher than non-spatial reproduction (Step a). This effect is overall not significant (Step b), even if the PostHoc tests found a single pair to be significantly different (Step c).

Rationale – effect per #IL: Some slight tendency of increasing values is visible (Step d), but no significant differences can be found (Steps e & f).

Figure C.4: Results for the measure *Speaker Recognition Performance*.

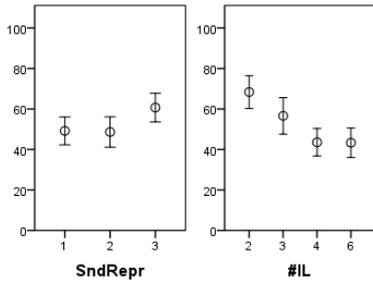
1. Analysis Steps

The present figure shows the data analysis for the measure *Speaker Recognition Confidence* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of #IL if all data over *SndRepr* is compiled and the effect of *SndRepr* if all data over #IL is compiled. Steps (d) to (f) concern the individual effect of #IL analyzed per *SndRepr*. Steps (g) to (i) concern the individual effect of *SndRepr* analyzed per #IL.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per #IL and *SndRepr*. Steps (b), (e), & (h): results of repeated-measures ANOVAs with #IL ∈ [2, 3, 4, 6] and *SndRepr* ∈ [1 = NB/non-spatial, 2 = FB/non-spatial, 3 = FB/spatial] as input, expressed by the *F*-measure and significance level *p*, and using the Greenhouse-Geisser correction when the sphericity criterion for repeated-measures ANOVAs is violated. Steps (c), (f), & (i): results of PostHoc tests (estimated marginal means) with Sidak correction for pairwise comparisons between all pairs of #IL or *SndRepr*, expressed by the significance level *p*; Significant differences (*p* ≤ 0.05) are indicated with an asterisk (\*).

2.1 Main Effects

(a) Errorbars



(b) ANOVA

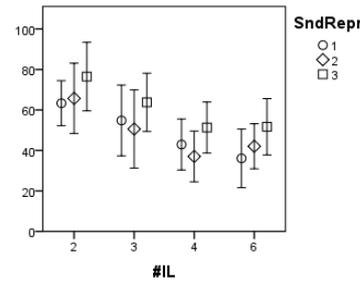
	<i>F</i>	<i>p</i>
#IL	54.634	0.000*
<i>SndRepr</i>	8.181	0.002*
Interaction	0.450	0.843

(c) PostHoc tests

<i>SndRepr</i>	<i>p</i>	#IL	<i>p</i>
1 vs. 2	0.948	2 vs. 3	0.001*
1 vs. 3	0.020*	2 vs. 4	0.000*
2 vs. 3	0.015*	2 vs. 6	0.000*
		3 vs. 4	0.007*
		3 vs. 6	0.002*
		4 vs. 6	1.000

2.2 Individual Effect of *SndRepr* per #IL

(d) Errorbars



(e) ANOVA per #IL

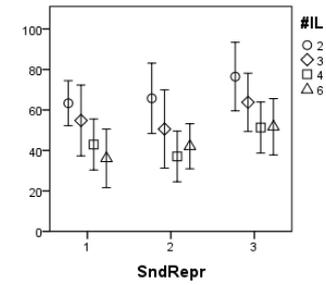
#IL	<i>F</i>	<i>p</i>
2	1.385	0.271
3	2.078	0.149
4	3.029	0.069
6	6.439	0.006*

(f) PostHoc tests per #IL

#IL	<i>SndRepr</i>	<i>p</i>
2	1 vs. 2	1.000
	1 vs. 3	0.513
	2 vs. 3	0.298
3	1 vs. 2	0.956
	1 vs. 3	0.374
	2 vs. 3	0.135
4	1 vs. 2	0.515
	1 vs. 3	0.685
	2 vs. 3	0.057
6	1 vs. 2	0.796
	1 vs. 3	0.020*
	2 vs. 3	0.046*

2.3 Individual Effect of #IL per *SndRepr*

(g) Errorbars



(h) ANOVA per *SndRepr*

<i>SndRepr</i>	<i>F</i>	<i>p</i>
1	7.237	0.001*
2	8.930	0.000*
3	13.401	0.000*

(i) PostHoc tests per *SndRepr*

<i>SndRepr</i>	#IL	<i>p</i>
1	2 vs. 3	0.691
	2 vs. 4	0.014*
	2 vs. 6	0.002*
	3 vs. 4	0.712
	3 vs. 6	0.231
	4 vs. 6	0.664
2	2 vs. 3	0.290
	2 vs. 4	0.002*
	2 vs. 6	0.011*
	3 vs. 4	0.152
	3 vs. 6	0.790
	4 vs. 6	0.977
3	2 vs. 3	0.227
	2 vs. 4	0.002*
	2 vs. 6	0.003*
	3 vs. 4	0.037*
	3 vs. 6	0.184
	4 vs. 6	0.899

3. Analysis Results:

3.1 Hypothesis H1 (Communication Complexity has an impact on Cognitive Load) is supported, as #IL has a significant impact on *Speaker Recognition Confidence*.

Rationale – main effect: Values decrease with increasing #IL (Step a). This effect is significant (Step b), and explained by significant differences between all expect one pair of #IL (Step c).

Rationale – effect per *SndRepr*: Values decrease with increasing #IL for all three *SndRepr* (Step g). This is significant (Step h), and is mainly explained by significant differences between #IL =2 and #IL > 3, i.e. two-party and medium-large multiparty party groups (Step i).

3.2 Hypothesis H2 (Technical System Capability has an impact on Cognitive Load) is supported, as *SndRepr*, more specifically spatial sound reproduction, has a significant impact on *Speaker Recognition Confidence*.

Rationale – main effect: Values for spatial sound reproduction are higher than non-spatial reproduction (Step a). This effect is significant (Step b), and confirmed by significant differences between the respective pairs of *SndRepr* (Step c).

Rationale – effect per #IL: Some slight tendency of increasing values is visible (Step d), but significant differences can be found only for #IL = 6, i.e. large groups (Step e), and there between non-spatial and spatial sound reproduction.

Figure C.5: Results for the measure *Speaker Recognition Confidence*.

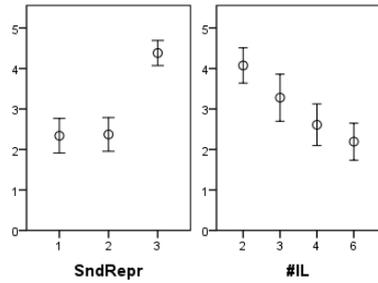
1. Analysis Steps

The present figure shows the data analysis for the measure *Speaker Recognition Effort* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of #IL if all data over *SndRepr* is compiled and the effect of *SndRepr* if all data over #IL is compiled. Steps (d) to (f) concern the individual effect of #IL analyzed per *SndRepr*. Steps (g) to (i) concern the individual effect of *SndRepr* analyzed per #IL.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per #IL and *SndRepr*. Steps (b), (e), & (h): results of repeated-measures ANOVAs with #IL ∈ [2, 3, 4, 6] and *SndRepr* ∈ [1 = NB/non-spatial, 2 = FB/non-spatial, 3 = FB/spatial] as input, expressed by the *F*-measure and significance level *p*, and using the Greenhouse-Geisser correction when the sphericity criterion for repeated-measures ANOVAs is violated. Steps (c), (f), & (i): results of PostHoc tests (estimated marginal means) with Sidak correction for pairwise comparisons between all pairs of #IL or *SndRepr*, expressed by the significance level *p*; Significant differences (*p* ≤ 0.05) are indicated with an asterisk (\*).

2.1 Main Effects

(a) Errorbars



(b) ANOVA

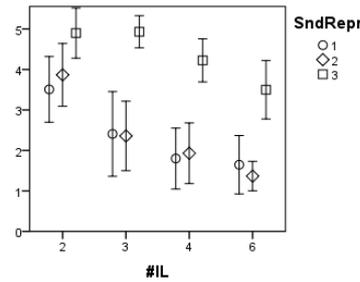
	<i>F</i>	<i>p</i>
#IL	25.332	0.000*
<i>SndRepr</i>	31.751	0.000*
Interaction	1.360	0.244

(c) PostHoc tests

<i>SndRepr</i>	<i>p</i>	#IL	<i>p</i>
1 vs. 2	0.988	2 vs. 3	0.003*
1 vs. 3	0.000*	2 vs. 4	0.001*
2 vs. 3	0.000*	2 vs. 6	0.000*
		3 vs. 4	0.041*
		3 vs. 6	0.009*
		4 vs. 6	0.506

2.2 Individual Effect of *SndRepr* per #IL

(d) Errorbars



(e) ANOVA per #IL

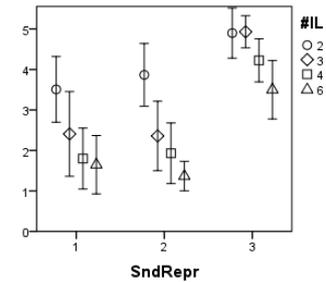
#IL	<i>F</i>	<i>p</i>
2	5.513	0.011*
3	13.504	0.000*
4	16.983	0.000*
6	13.462	0.000*

(f) PostHoc tests per #IL

#IL	<i>SndRepr</i>	<i>p</i>
2	1 vs. 2	0.816
	1 vs. 3	0.020*
	2 vs. 3	0.086
3	1 vs. 2	1.000
	1 vs. 3	0.001*
	2 vs. 3	0.000*
4	1 vs. 2	0.963
	1 vs. 3	0.000*
	2 vs. 3	0.002*
6	1 vs. 2	0.918
	1 vs. 3	0.014*
	2 vs. 3	0.001*

2.3 Individual Effect of #IL per *SndRepr*

(g) Errorbars



(h) ANOVA per *SndRepr*

<i>SndRepr</i>	<i>F</i>	<i>p</i>
1	4.924	0.006*
2	12.720	0.000*
3	10.318	0.000*

(i) PostHoc tests per *SndRepr*

<i>SndRepr</i>	#IL	<i>p</i>
1	2 vs. 3	0.353
	2 vs. 4	0.029*
	2 vs. 6	0.011*
	3 vs. 4	0.694
2	3 vs. 6	0.881
	4 vs. 6	1.000
	2 vs. 3	0.118
	2 vs. 4	0.009*
3	2 vs. 6	0.000*
	3 vs. 4	0.675
	3 vs. 6	0.220
	4 vs. 6	0.786
	2 vs. 3	1.000
	2 vs. 4	0.282
	2 vs. 6	0.008*
	3 vs. 4	0.060
3 vs. 6	0.008*	
4 vs. 6	0.205	

3. Analysis Results:

3.1 Hypothesis H1 (Communication Complexity has an impact on Cognitive Load) is supported, as #IL has a significant impact on *Speaker Recognition Effort*.

Rationale – main effect: Values decrease with increasing #IL (Step a). This effect is significant (Step b), and explained by significant differences between all expect one pair of #IL (Step c).

Rationale – effect per *SndRepr*: Values decrease with increasing #IL for all three *SndRepr* (Step g). This is significant (Step h), and is mainly explained by significant differences between #IL =2 and #IL > 3, i.e. two-party and medium-large multiparty party groups (Step i).

3.2 Hypothesis H2 (Technical System Capability has an impact on Cognitive Load) is supported, as *SndRepr*, more specifically spatial sound reproduction, has a significant impact on *Speaker Recognition Effort*.

Rationale – main effect: Values for spatial sound reproduction are higher than non-spatial reproduction (Step a). This effect is significant (Step b), and confirmed by significant differences between the respective pairs of *SndRepr* (Step c).

Rationale – effect per #IL: Values for spatial sound reproduction are higher than non-spatial reproduction for all #IL (Step c). This effect is significant (Step d), and mainly explained by significant differences between non-spatial and spatial sound reproduction (Step f).

Figure C.6: Results for the measure *Speaker Recognition Effort*.

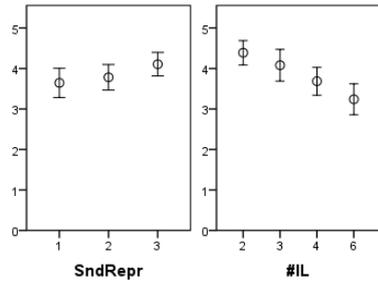
1. Analysis Steps

The present figure shows the data analysis for the measure *Topic Comprehension Effort* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of #IL if all data over *SndRepr* is compiled and the effect of *SndRepr* if all data over #IL is compiled. Steps (d) to (f) concern the individual effect of #IL analyzed per *SndRepr*. Steps (g) to (i) concern the individual effect of *SndRepr* analyzed per #IL.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per #IL and *SndRepr*. Steps (b), (e), & (h): results of repeated-measures ANOVAs with #IL ∈ [2, 3, 4, 6] and *SndRepr* ∈ [1 = NB/non-spatial, 2 = FB/non-spatial, 3 = FB/spatial] as input, expressed by the *F*-measure and significance level *p*, and using the Greenhouse-Geisser correction when the sphericity criterion for repeated-measures ANOVAs is violated. Steps (c), (f), & (i): results of PostHoc tests (estimated marginal means) with Sidak correction for pairwise comparisons between all pairs of #IL or *SndRepr*, expressed by the significance level *p*; Significant differences (*p* ≤ 0.05) are indicated with an asterisk (\*).

2.1 Main Effects

(a) Errorbars



b) ANOVA

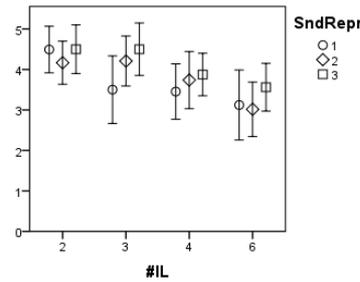
	<i>F</i>	<i>p</i>
#IL	7.718	0.000*
<i>SndRepr</i>	2.399	0.114
Interaction	3.194	0.008*

(c) PostHoc tests

<i>SndRepr</i>	<i>p</i>	#IL	<i>p</i>
1 vs. 2	0.972	2 vs. 3	0.705
1 vs. 3	0.193	2 vs. 4	0.052
2 vs. 3	0.114	2 vs. 6	0.011*
		3 vs. 4	0.653
		3 vs. 6	0.136
		4 vs. 6	0.265

2.2 Individual Effect of *SndRepr* per #IL

(d) Errorbars



(e) ANOVA per #IL

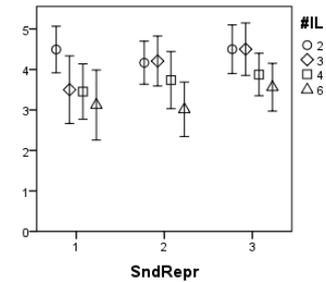
#IL	<i>F</i>	<i>p</i>
2	1.146	0.336
3	9.216	0.001*
4	0.332	0.721
6	3.568	0.045*

(f) PostHoc tests per #IL

#IL	<i>SndRepr</i>	<i>p</i>
2	1 vs. 2	0.595
	1 vs. 3	0.999
	2 vs. 3	0.391
3	1 vs. 2	0.084
	1 vs. 3	0.012*
	2 vs. 3	0.131
4	1 vs. 2	0.927
	1 vs. 3	0.841
	2 vs. 3	1.000
6	1 vs. 2	0.886
	1 vs. 3	0.304
	2 vs. 3	0.015*

2.3 Individual Effect of #IL per *SndRepr*

(g) Errorbars



(h) ANOVA per *SndRepr*

<i>SndRepr</i>	<i>F</i>	<i>p</i>
1	3.934	0.017*
2	6.576	0.001*
3	12.030	0.000*

(i) PostHoc tests per *SndRepr*

<i>SndRepr</i>	#IL	<i>p</i>
1	2 vs. 3	0.060
	2 vs. 4	0.078
	2 vs. 6	0.095
	3 vs. 4	1.000
	3 vs. 6	0.996
	4 vs. 6	0.962
2	2 vs. 3	1.000
	2 vs. 4	0.912
	2 vs. 6	0.010*
	3 vs. 4	0.964
	3 vs. 6	0.048*
	4 vs. 6	0.047*
3	2 vs. 3	0.907
	2 vs. 4	0.043*
	2 vs. 6	0.010*
	3 vs. 4	0.008*
	3 vs. 6	0.003*
	4 vs. 6	0.941

3. Analysis Results:

3.1 Hypothesis H1 (Communication Complexity has an impact on Cognitive Load) is supported, as #IL has a significant impact on *Topic Comprehension Effort*.

Rationale – main effect: Values decrease with increasing #IL (Step a). This effect is significant (Step b), and explained by significant difference between #IL = 2 & 6, i.e. two-party vs. large multiparty groups (Step c).

Rationale – effect per *SndRepr*: Values decrease with increasing #IL for all three *SndRepr* (Step g). This is significant (Step h) for all three *SndRepr* (Step i). The PostHoc tests can identify an increasing number of respective significant pairs of #IL for an increasing *SndRepr*.

3.2 Hypothesis H2 (Technical System Capability has an impact on Cognitive Load) is not supported, as *SndRepr* has no significant impact on *Topic Comprehension Effort*, except in few subsets of the data.

Rationale – main effect: Values for spatial sound reproduction are higher than non-spatial reproduction (Step a). However, this effect is not significant (Steps b&c).

Rationale – effect per #IL: No consistent tendency is visible (Step c). There are significant effects for #IL = 3 & 6 (Step d), explained by significant differences between some single pairs of *SndRepr* (Step f).

Figure C.7: Results for the measure *Topic Comprehension Effort*.

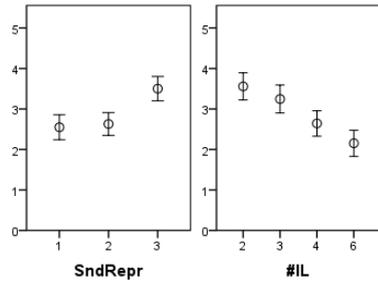
1. Analysis Steps

The present figure shows the data analysis for the measure *Concentration Effort* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of #IL if all data over *SndRepr* is compiled and the effect of *SndRepr* if all data over #IL is compiled. Steps (d) to (f) concern the individual effect of #IL analyzed per *SndRepr*. Steps (g) to (i) concern the individual effect of *SndRepr* analyzed per #IL.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per #IL and *SndRepr*. Steps (b), (e), & (h): results of repeated-measures ANOVAs with #IL ∈ [2, 3, 4, 6] and *SndRepr* ∈ [1 = NB/non-spatial, 2 = FB/non-spatial, 3 = FB/spatial] as input, expressed by the *F*-measure and significance level *p*, and using the Greenhouse-Geisser correction when the sphericity criterion for repeated-measures ANOVAs is violated. Steps (c), (f), & (i): results of PostHoc tests (estimated marginal means) with Sidak correction for pairwise comparisons between all pairs of #IL or *SndRepr*, expressed by the significance level *p*; Significant differences (*p* ≤ 0.05) are indicated with an asterisk (\*).

2.1 Main Effects

(a) Errorbars



(b) ANOVA

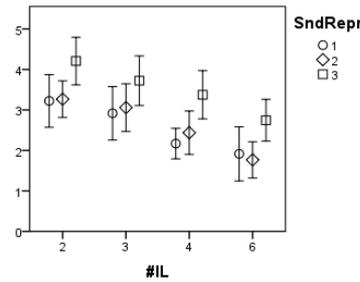
	<i>F</i>	<i>p</i>
#IL	24.601	0.000*
<i>SndRepr</i>	15.067	0.000*
Interaction	0.279	0.945

(c) PostHoc tests

<i>SndRepr</i>	<i>p</i>	#IL	<i>p</i>
1 vs. 2	0.944	2 vs. 3	0.605
1 vs. 3	0.006*	2 vs. 4	0.001*
2 vs. 3	0.002*	2 vs. 6	0.000*
		3 vs. 4	0.038*
		3 vs. 6	0.000*
		4 vs. 6	0.093

2.2 Individual Effect of *SndRepr* per #IL

(d) Errorbars



(e) ANOVA per #IL

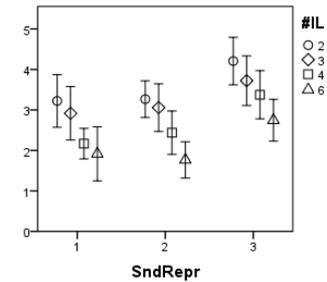
#IL	<i>F</i>	<i>p</i>
2	7.548	0.003*
3	4.682	0.020*
4	6.491	0.006*
6	9.447	0.001*

(f) PostHoc tests per #IL

#IL	<i>SndRepr</i>	<i>p</i>
2	1 vs. 2	0.999
	1 vs. 3	0.043*
	2 vs. 3	0.016*
3	1 vs. 2	0.969
	1 vs. 3	0.094
	2 vs. 3	0.032*
4	1 vs. 2	0.727
	1 vs. 3	0.028*
	2 vs. 3	0.075
6	1 vs. 2	0.922
	1 vs. 3	0.068
	2 vs. 3	0.000*

2.3 Individual Effect of #IL per *SndRepr*

(g) Errorbars



(h) ANOVA per *SndRepr*

<i>SndRepr</i>	<i>F</i>	<i>p</i>
1	7.430	0.001*
2	15.709	0.000*
3	9.510	0.000*

(i) PostHoc tests per *SndRepr*

<i>SndRepr</i>	#IL	<i>p</i>
1	2 vs. 3	0.925
	2 vs. 4	0.027*
	2 vs. 6	0.023*
	3 vs. 4	0.088
	3 vs. 6	0.073
	4 vs. 6	0.964
2	2 vs. 3	0.945
	2 vs. 4	0.026*
	2 vs. 6	0.000*
	3 vs. 4	0.416
	3 vs. 6	0.002*
	4 vs. 6	0.118
3	2 vs. 3	0.880
	2 vs. 4	0.153
	2 vs. 6	0.008*
	3 vs. 4	0.231
	3 vs. 6	0.001*
	4 vs. 6	0.324

3. Analysis Results:

3.1 Hypothesis H1 (Communication Complexity has an impact on Cognitive Load) is supported, as #IL has a significant impact on *Concentration Effort*.

Rationale – main effect: Values decrease with increasing #IL (Step a). This effect is significant (Step b), and explained by significant differences between almost all pairs of #IL (Step c).

Rationale – effect per *SndRepr*: Values decrease with increasing #IL for all three *SndRepr* (Step g). This is significant (Step h), and is mainly explained by significant differences between #IL = 2 & 3 vs. #IL = 4 & 6, i.e. groups sizes need to differ substantially (Step i).

3.2 Hypothesis H2 (Technical System Capability has an impact on Cognitive Load) is supported, as *SndRepr*, more specifically spatial sound reproduction, has a significant impact on *Concentration Effort*.

Rationale – main effect: Values for spatial sound reproduction are higher than non-spatial reproduction (Step a). This effect is significant (Step b), and confirmed by significant differences between the respective pairs of *SndRepr* (Step c).

Rationale – effect per #IL: Values for spatial sound reproduction are higher than non-spatial reproduction for all #IL (Step c). This effect is significant (Step d), and mainly explained by significant differences between non-spatial and spatial sound reproduction (Step f).

Figure C.8: Results for the measure *Concentration Effort*.

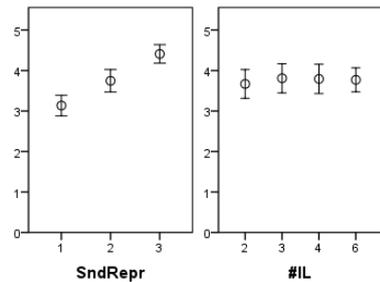
1. Analysis Steps

The present figure shows the data analysis for the measure *Speech Intelligibility* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of #IL if all data over *SndRepr* is compiled and the effect of *SndRepr* if all data over #IL is compiled. Steps (d) to (f) concern the individual effect of #IL analyzed per *SndRepr*. Steps (g) to (i) concern the individual effect of *SndRepr* analyzed per #IL.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per #IL and *SndRepr*. Steps (b), (e), & (h): results of repeated-measures ANOVAs with #IL ∈ [2,3,4,6] and *SndRepr* ∈ [1 = NB/non-spatial, 2 = FB/non-spatial, 3 = FB/spatial] as input, expressed by the *F*-measure and significance level *p*, and using the Greenhouse-Geisser correction when the sphericity criterion for repeated-measures ANOVAs is violated. Steps (c), (f), & (i): results of PostHoc tests (estimated marginal means) with Sidak correction for pairwise comparisons between all pairs of #IL or *SndRepr*, expressed by the significance level *p*; Significant differences (*p* ≤ 0.05) are indicated with an asterisk (\*).

2.1 Main Effects

(a) Errorbars



(b) ANOVA

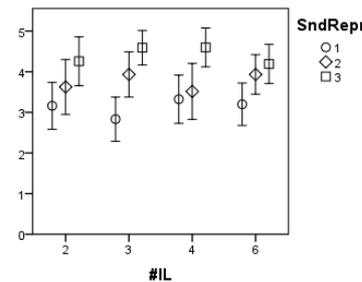
	<i>F</i>	<i>p</i>
#IL	0.304	0.822
<i>SndRepr</i>	25.807	0.000*
Interaction	2.696	0.060

(c) PostHoc tests

<i>SndRepr</i>	<i>p</i>	#IL	<i>p</i>
1 vs. 2	0.028*	2 vs. 3	0.920
1 vs. 3	0.000*	2 vs. 4	0.996
2 vs. 3	0.004*	2 vs. 6	0.998
		3 vs. 4	1.000
		3 vs. 6	0.999
		4 vs. 6	1.000

2.2 Individual Effect of *SndRepr* per #IL

(d) Errorbars



(e) ANOVA per #IL

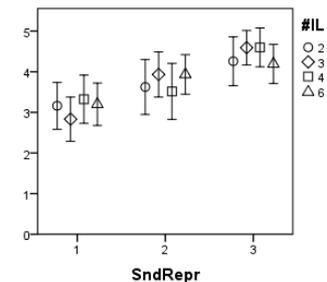
#IL	<i>F</i>	<i>p</i>
2	10.202	0.001*
3	20.956	0.000*
4	11.392	0.000*
6	9.004	0.001*

(f) PostHoc tests per #IL

#IL	<i>SndRepr</i>	<i>p</i>
2	1 vs. 2	0.202
	1 vs. 3	0.000*
	2 vs. 3	0.155
3	1 vs. 2	0.008*
	1 vs. 3	0.000*
	2 vs. 3	0.094
4	1 vs. 2	0.983
	1 vs. 3	0.003*
	2 vs. 3	0.001*
6	1 vs. 2	0.048*
	1 vs. 3	0.005*
	2 vs. 3	0.786

2.3 Individual Effect of #IL per *SndRepr*

(g) Errorbars



(h) ANOVA per *SndRepr*

<i>SndRepr</i>	<i>F</i>	<i>p</i>
1	1.031	0.392
2	2.707	0.061
3	2.360	0.118

(i) PostHoc tests per *SndRepr*

<i>SndRepr</i>	#IL	<i>p</i>
1	2 vs. 3	0.921
	2 vs. 4	0.994
	2 vs. 6	1.000
	3 vs. 4	0.731
	3 vs. 6	0.795
	4 vs. 6	0.991
2	2 vs. 3	0.743
	2 vs. 4	0.829
	2 vs. 6	0.845
	3 vs. 4	0.152
	3 vs. 6	1.000
	4 vs. 6	0.374
3	2 vs. 3	0.850
	2 vs. 4	0.793
	2 vs. 6	0.976
	3 vs. 4	1.000
	3 vs. 6	0.362
	4 vs. 6	0.036*

3. Analysis Results:

3.1 Hypothesis H3.1 (Communication Complexity has an impact on Speech Communication Quality with focus on Telecommunication Component) is not supported, as #IL has no significant impact on *Speech Intelligibility*.

Rationale – main effect: Values are relatively constant with increasing #IL (Step a), and no significant effects can be found (Steps b & c).

Rationale – effect per *SndRepr*: Values are relatively constant with increasing #IL for all three *SndRepr* (Step g), and no significant effects can be found (Step h), except for one single pair (Step i).

3.2 Hypothesis H4.1 (Technical System Capability has an impact on Speech Communication Quality with focus on Telecommunication Component) is supported, as *SndRepr* has a significant impact on *Speech Intelligibility*.

Rationale – main effect: Values increase with increasing *SndRepr* (Step a). This effect is significant (Step b), and confirmed by significant differences between all pairs of *SndRepr* (Step c).

Rationale – effect per #IL: Values increase with increasing *SndRepr* for all #IL (Step d). This effect is significant (Step e), and confirmed by significant differences between almost all pairs of *SndRepr* (Step f).

Figure C.9: Results for the measure *Speech Intelligibility*.

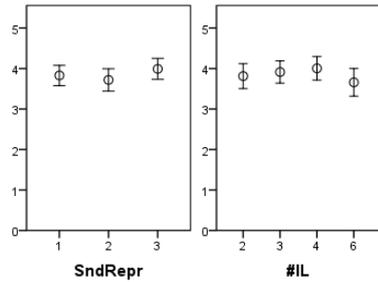
1. Analysis Steps

The present figure shows the data analysis for the measure *Conversation Naturalness* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of #IL if all data over *SndRepr* is compiled and the effect of *SndRepr* if all data over #IL is compiled. Steps (d) to (f) concern the individual effect of #IL analyzed per *SndRepr*. Steps (g) to (i) concern the individual effect of *SndRepr* analyzed per #IL.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per #IL and *SndRepr*. Steps (b), (e), & (h): results of repeated-measures ANOVAs with #IL ∈ [2,3,4,6] and *SndRepr* ∈ [1 = NB/non-spatial, 2 = FB/non-spatial, 3 = FB/spatial] as input, expressed by the *F*-measure and significance level *p*, and using the Greenhouse-Geisser correction when the sphericity criterion for repeated-measures ANOVAs is violated. Steps (c), (f), & (i): results of PostHoc tests (estimated marginal means) with Sidak correction for pairwise comparisons between all pairs of #IL or *SndRepr*, expressed by the significance level *p*; Significant differences (*p* ≤ 0.05) are indicated with an asterisk (\*).

2.1 Main Effects

(a) Errorbars



(b) ANOVA

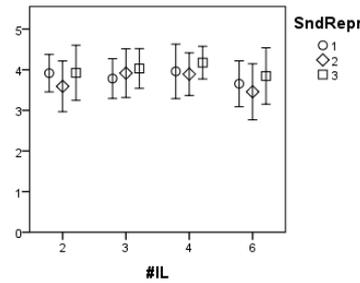
	<i>F</i>	<i>p</i>
#IL	1.865	0.190
<i>SndRepr</i>	1.903	0.173
Interaction	0.272	0.948

(c) PostHoc tests

<i>SndRepr</i>	<i>p</i>	#IL	<i>p</i>
1 vs. 2	0.925	2 vs. 3	0.994
1 vs. 3	0.557	2 vs. 4	0.857
2 vs. 3	0.171	2 vs. 6	0.977
		3 vs. 4	0.864
		3 vs. 6	0.339
		4 vs. 6	0.145

2.2 Individual Effect of *SndRepr* per #IL

(d) Errorbars



(e) ANOVA per #IL

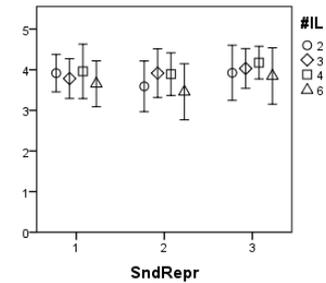
#IL	<i>F</i>	<i>p</i>
2	0.689	0.513
3	0.533	0.594
4	0.680	0.517
6	0.964	0.397

(f) PostHoc tests per #IL

#IL	<i>SndRepr</i>	<i>p</i>
2	1 vs. 2	0.771
	1 vs. 3	1.000
	2 vs. 3	0.377
3	1 vs. 2	0.899
	1 vs. 3	0.752
	2 vs. 3	0.955
4	1 vs. 2	1.000
	1 vs. 3	0.613
	2 vs. 3	0.526
6	1 vs. 2	0.910
	1 vs. 3	0.754
	2 vs. 3	0.632

2.3 Individual Effect of #IL per *SndRepr*

(g) Errorbars



(h) ANOVA per *SndRepr*

<i>SndRepr</i>	<i>F</i>	<i>p</i>
1	0.674	0.574
2	1.258	0.305
3	0.684	0.568

(i) PostHoc tests per *SndRepr*

<i>SndRepr</i>	#IL	<i>p</i>
1	2 vs. 3	0.997
	2 vs. 4	1.000
	2 vs. 6	0.692
	3 vs. 4	0.999
2	3 vs. 6	0.948
	4 vs. 6	0.809
	2 vs. 3	0.737
	2 vs. 4	0.812
3	2 vs. 6	0.999
	3 vs. 4	1.000
	3 vs. 6	0.671
	4 vs. 6	0.641
	2 vs. 3	1.000
	2 vs. 4	0.961
	2 vs. 6	0.999
	3 vs. 4	0.959
3 vs. 6	0.899	
4 vs. 6	0.490	

3. Analysis Results:

3.1 Hypothesis H3.2 (Communication Complexity has an impact on Speech Communication Quality with focus on Group-Communication Component) is not supported, as #IL has no significant impact on *Conversation Naturalness*.

Rationale – main effect: Values are relatively constant with increasing #IL (Step a), and no significant effects can be found (Steps b & c).

Rationale – effect per *SndRepr*: Values are relatively constant with increasing #IL for all three *SndRepr* (Step g), and no significant effects can be found (Steps h & i).

3.2 Hypothesis H4.2 (Technical System Capability has an impact on Speech Communication Quality with focus on Group-Communication Component) is not supported, as *SndRepr* has no significant impact on *Conversation Naturalness*.

Rationale – main effect: Values are relatively constant with increasing *SndRepr* (Step a), and no significant effects can be found (Steps b & c).

Rationale – effect per #IL: Values are relatively constant with increasing *SndRepr* for all four #IL (Step g), and no significant effects can be found (Steps e & f).

Figure C.10: Results for the measure *Conversation Naturalness*.

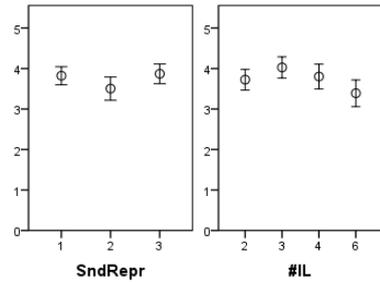
1. Analysis Steps

The present figure shows the data analysis for the measure *Conversation Efficiency* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of #IL if all data over *SndRepr* is compiled and the effect of *SndRepr* if all data over #IL is compiled. Steps (d) to (f) concern the individual effect of #IL analyzed per *SndRepr*. Steps (g) to (i) concern the individual effect of *SndRepr* analyzed per #IL.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per #IL and *SndRepr*. Steps (b), (e), & (h): results of repeated-measures ANOVAs with #IL ∈ [2,3,4,6] and *SndRepr* ∈ [1 = NB/non-spatial, 2 = FB/non-spatial, 3 = FB/spatial] as input, expressed by the *F*-measure and significance level *p*, and using the Greenhouse-Geisser correction when the sphericity criterion for repeated-measures ANOVAs is violated. Steps (c), (f), & (i): results of PostHoc tests (estimated marginal means) with Sidak correction for pairwise comparisons between all pairs of #IL or *SndRepr*, expressed by the significance level *p*; Significant differences (*p* ≤ 0.05) are indicated with an asterisk (\*).

2.1 Main Effects

(a) Errorbars



(b) ANOVA

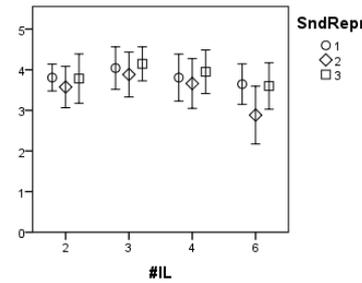
	<i>F</i>	<i>p</i>
#IL	5.659	0.016*
<i>SndRepr</i>	5.506	0.012*
Interaction	0.581	0.596

(c) PostHoc tests

<i>SndRepr</i>	<i>p</i>	#IL	<i>p</i>
1 vs. 2	0.168	2 vs. 3	0.284
1 vs. 3	0.939	2 vs. 4	0.997
2 vs. 3	0.012*	2 vs. 6	0.107
		3 vs. 4	0.191
		3 vs. 6	0.061
		4 vs. 6	0.300

2.2 Individual Effect of *SndRepr* per #IL

(d) Errorbars



(e) ANOVA per #IL

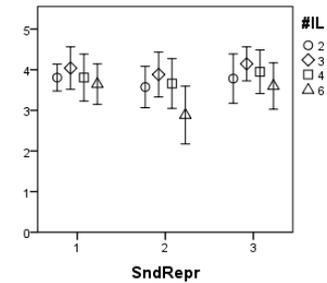
#IL	<i>F</i>	<i>p</i>
2	0.369	0.696
3	0.616	0.549
4	0.620	0.547
6	2.888	0.110

(f) PostHoc tests per #IL

#IL	<i>SndRepr</i>	<i>p</i>
2	1 vs. 2	0.853
	1 vs. 3	1.000
	2 vs. 3	0.906
3	1 vs. 2	0.716
	1 vs. 3	0.996
	2 vs. 3	0.797
4	1 vs. 2	0.913
	1 vs. 3	0.943
	2 vs. 3	0.641
6	1 vs. 2	0.321
	1 vs. 3	0.994
	2 vs. 3	0.264

2.3 Individual Effect of #IL per *SndRepr*

(g) Errorbars



(h) ANOVA per *SndRepr*

<i>SndRepr</i>	<i>F</i>	<i>p</i>
1	1.134	0.349
2	3.443	0.028*
3	0.789	0.441

(i) PostHoc tests per *SndRepr*

<i>SndRepr</i>	#IL	<i>p</i>
1	2 vs. 3	0.856
	2 vs. 4	1.000
	2 vs. 6	0.976
	3 vs. 4	0.731
2	3 vs. 6	0.483
	4 vs. 6	0.987
	2 vs. 3	0.741
	2 vs. 4	1.000
3	2 vs. 6	0.548
	3 vs. 4	0.786
	3 vs. 6	0.190
	4 vs. 6	0.385
	2 vs. 3	0.930
	2 vs. 4	0.999
	2 vs. 6	0.988
	3 vs. 4	0.963
3 vs. 6	0.720	
4 vs. 6	0.968	

3. Analysis Results:

3.1 Hypothesis H3.2 (Communication Complexity has an impact on Speech Communication Quality with focus on Group-Communication Component) is weakly supported, as #IL has a significant impact on *Conversation Efficiency*, but can not be detected by the PostHoc tests.

Rationale – main effect: With the expectation of #IL = 2, values are decreasing with increasing #IL (Step a). This effect is significant (Step b), but respective significantly different pairs can not be detected by the PostHoc tests (Step c).

Rationale – effect per *SndRepr*: With the expectation of #IL = 2, values tend to decrease with increasing #IL for all three *SndRepr*(Step g). This effect, however, is only significant for *SndRepr* = 2 (Step h), but respective significantly different pairs can not be detected by the PostHoc tests (Step i).

3.2 Hypothesis H4.2 (Technical System Capability has an impact on Speech Communication Quality with focus on Group-Communication Component) is weakly supported, as *SndRepr* has a significant impact on *Conversation Efficiency*, but only in one subset of the data.

Rationale – main effect: Values for wideband non-spatial sound reproduction are lower than for the other *SndRepr* (Step a). This effect is significant (Step b), but respective significantly different pairs can not be detected by the PostHoc tests (Step c).

Rationale – effect per #IL: No consistent trend is visible (Step g), and no significant effects can be found (Steps e & f).

Figure C.11: Results for the measure *Conversation Efficiency*.

## C.2 Experiment CC2

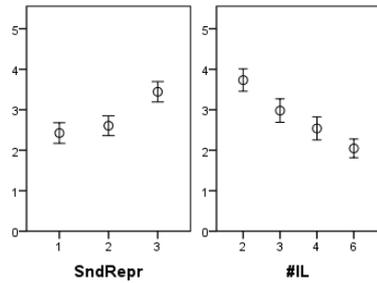
### 1. Analysis Steps

The present figure shows the data analysis for the measure *Concentration Effort* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of #IL if all data over *SndRepr* is compiled and the effect of *SndRepr* if all data over #IL is compiled. Steps (d) to (f) concern the individual effect of #IL analyzed per *SndRepr*. Steps (g) to (i) concern the individual effect of *SndRepr* analyzed per #IL.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per #IL and *SndRepr*. Steps (b), (e), & (h): results of repeated-measures ANOVAs with #IL ∈ [2,3,4,6] and *SndRepr* ∈ [1 = NB/non-spatial, 2 = FB/non-spatial, 3 = FB/spatial] as input, expressed by the *F*-measure and significance level *p*, and using the Greenhouse-Geisser correction when the sphericity criterion for repeated-measures ANOVAs is violated. Steps (c), (f), & (i): results of PostHoc tests (estimated marginal means) with Sidak correction for pairwise comparisons between all pairs of #IL or *SndRepr*, expressed by the significance level *p*; Significant differences (*p* ≤ 0.05) are indicated with an asterisk (\*).

#### 2.1 Main Effects

(a) Errorbars



(b) ANOVA

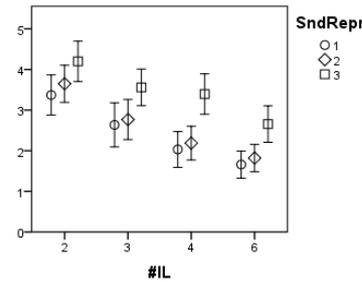
	<i>F</i>	<i>p</i>
<i>SndRepr</i>	17.712	0.000*
#IL	33.039	0.000*
Interaction	1.038	0.394

(c) PostHoc tests

<i>SndRepr</i>	<i>p</i>	#IL	<i>p</i>
1 vs. 2	0.375	2 vs. 3	0.000*
1 vs. 3	0.000*	2 vs. 4	0.000*
2 vs. 3	0.002*	2 vs. 6	0.000*
		3 vs. 4	0.102
		3 vs. 6	0.001*
		4 vs. 6	0.024*

#### 2.2 Individual Effect of *SndRepr* per #IL

(d) Errorbars



(e) ANOVA per #IL

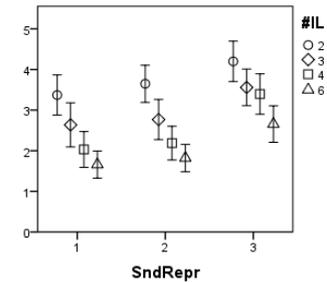
#IL	<i>F</i>	<i>p</i>
2	7.373	0.002*
3	7.933	0.004*
4	15.950	0.000*
6	9.893	0.003*

(f) PostHoc tests per #IL

#IL	<i>SndRepr</i>	<i>p</i>
2	1 vs. 2	0.637
	1 vs. 3	0.008*
	2 vs. 3	0.023*
3	1 vs. 2	0.853
	1 vs. 3	0.012*
	2 vs. 3	0.030*
4	1 vs. 2	0.639
	1 vs. 3	0.001*
	2 vs. 3	0.001*
6	1 vs. 2	0.405
	1 vs. 3	0.002*
	2 vs. 3	0.035*

#### 2.3 Individual Effect of #IL per *SndRepr*

(g) Errorbars



(h) ANOVA per *SndRepr*

<i>SndRepr</i>	<i>F</i>	<i>p</i>
1	20.887	0.000*
2	25.896	0.000*
3	12.750	0.000*

(i) PostHoc tests per *SndRepr*

<i>SndRepr</i>	#IL	<i>p</i>
1	2 vs. 3	0.005*
	2 vs. 4	0.000*
	2 vs. 6	0.000*
	3 vs. 4	0.070*
2	3 vs. 6	0.006*
	4 vs. 6	0.186
	2 vs. 3	0.012*
	2 vs. 4	0.000*
	2 vs. 6	0.000*
	3 vs. 4	0.031*
	3 vs. 6	0.000*
	4 vs. 6	0.241
3	2 vs. 3	0.028*
	2 vs. 4	0.007*
	2 vs. 6	0.000*
	3 vs. 4	1.000
	3 vs. 6	0.011*
	4 vs. 6	0.090

### 3. Analysis Results:

**3.1 Hypothesis H1** (Communication Complexity has an impact on Cognitive Load) is supported, as #IL has a significant impact on *Concentration Effort*.

Rationale – main effect: Values decrease (i.e. increasing *Concentration Effort*) with increasing #IL (Step a). This effect is significant (Step b) and explained by significant differences between all except one pair of #IL (Step c).

Rationale – effect per *SndRepr*: Values decrease for all three *SndRepr* (Step g). This is significant (Step h) and almost all steps between #IL are large enough to be significant as well (Step i).

**3.2 Hypothesis H2** (Technical System Capability has an impact on Cognitive Load) is supported, as *SndRepr*, more specifically the spatial sound reproduction, has a significant impact on *Concentration Effort*.

Rationale – main effect: Values increase (i.e. decreasing *Concentration Effort*) with increasing *SndRepr* (Step a). This effect is significant (Step b) and explained by significant differences between spatial and non-spatial sound reproduction, but not between the two non-spatial conditions (Step c).

Rationale – effect per #IL: Values increase for all four #IL (Step d). This is significant (Step e) and is always explained by significant differences between spatial and non-spatial sound reproduction, but not between the two non-spatial conditions (Step f).

Figure C.12: Results for the measure *Concentration Effort*.

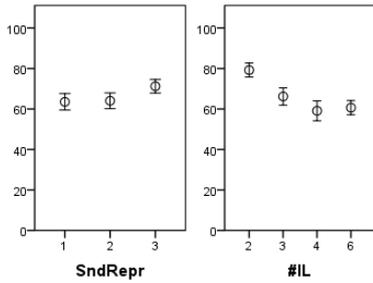
1. Analysis Steps

The present figure shows the data analysis for the measure *Speaker Recognition Performance* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of #IL if all data over *SndRepr* is compiled and the effect of *SndRepr* if all data over #IL is compiled. Steps (d) to (f) concern the individual effect of #IL analyzed per *SndRepr*. Steps (g) to (i) concern the individual effect of *SndRepr* analyzed per #IL.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per #IL and *SndRepr*. Steps (b), (e), & (h): results of repeated-measures ANOVAs with #IL ∈ [2, 3, 4, 6] and *SndRepr* ∈ [1 = NB/non-spatial, 2 = FB/non-spatial, 3 = FB/spatial] as input, expressed by the *F*-measure and significance level *p*, and using the Greenhouse-Geisser correction when the sphericity criterion for repeated-measures ANOVAs is violated. Steps (c), (f), & (i): results of PostHoc tests (estimated marginal means) with Sidak correction for pairwise comparisons between all pairs of #IL or *SndRepr*, expressed by the significance level *p*; Significant differences (*p* ≤ 0.05) are indicated with an asterisk (\*).

2.1 Main Effects

(a) Errorbars



(b) ANOVA

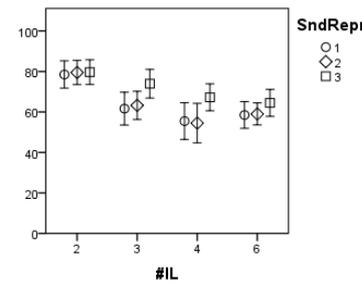
	<i>F</i>	<i>p</i>
<i>SndRepr</i>	7.713	0.001*
#IL	23.755	0.000*
Interaction	1.179	0.321

(c) PostHoc tests

<i>SndRepr</i>	<i>p</i>	#IL	<i>p</i>
1 vs. 2	1.000	2 vs. 3	0.000*
1 vs. 3	0.026*	2 vs. 4	0.000*
2 vs. 3	0.003*	2 vs. 6	0.000*
		3 vs. 4	0.194
		3 vs. 6	0.351
		4 vs. 6	1.000

2.2 Individual Effect of *SndRepr* per #IL

(d) Errorbars



(e) ANOVA per #IL

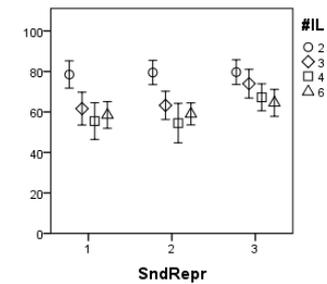
#IL	<i>F</i>	<i>p</i>
2	0.043	0.958
3	3.598	0.035*
4	5.630	0.006*
6	1.944	0.154

(f) PostHoc tests per #IL

#IL	<i>SndRepr</i>	<i>p</i>
2	1 vs. 2	1.000
	1 vs. 3	0.990
	2 vs. 3	0.994
3	1 vs. 2	0.999
	1 vs. 3	0.056
	2 vs. 3	0.141
4	1 vs. 2	0.998
	1 vs. 3	0.041*
	2 vs. 3	0.039*
6	1 vs. 2	0.998
	1 vs. 3	0.287
	2 vs. 3	0.304

2.3 Individual Effect of #IL per *SndRepr*

(g) Errorbars



(h) ANOVA per *SndRepr*

<i>SndRepr</i>	<i>F</i>	<i>p</i>
1	13.084	0.000*
2	11.577	0.000*
3	6.536	0.001*

(i) PostHoc tests per *SndRepr*

<i>SndRepr</i>	#IL	<i>p</i>
1	2 vs. 3	0.001*
	2 vs. 4	0.000*
	2 vs. 6	0.000*
	3 vs. 4	0.628
	3 vs. 6	0.976
	4 vs. 6	0.944
2	2 vs. 3	0.015*
	2 vs. 4	0.001*
	2 vs. 6	0.000*
	3 vs. 4	0.229
	3 vs. 6	0.897
	4 vs. 6	0.902
3	2 vs. 3	0.483
	2 vs. 4	0.062
	2 vs. 6	0.000*
	3 vs. 4	0.768
	3 vs. 6	0.296
	4 vs. 6	0.661

3. Analysis Results:

3.1 Hypothesis H1 (Communication Complexity has an impact on Cognitive Load) is supported, as #IL has a significant impact on *Speaker Recognition Performance*, whereas this effect is predominantly visible between two-party and multiparty settings for non-spatial conditions.

Rationale – main effect: Values decrease with increasing #IL (Step a). This effect is significant (Step b) and explained by significant differences between two-party (#IL = 2) and multiparty (#IL > 2) groups (Step c).

Rationale – effect per *SndRepr*: Values decrease for all three *SndRepr* (Step g). This is significant (Step h). For the non-spatial conditions, the differences are explained between two-party and multiparty; for the spatial condition between most extreme cases 2 vs. 6 (Step i).

3.2 Hypothesis H2 (Technical System Capability has an impact on Cognitive Load) is supported, as *SndRepr*, more specifically the spatial sound reproduction, has a significant impact on *Speaker Recognition Performance*.

Rationale – main effect: Values increase with increasing *SndRepr* (Step a). This effect is significant (Step b) and explained by significant differences between spatial and non-spatial sound reproduction, but not between the two non-spatial conditions (Step c).

Rationale – effect per #IL: Values increase for small to medium groups #IL = 3 & 4 (Step d). This is significant (Step e) and is explained by significant differences between spatial and non-spatial sound reproduction, but not between the two non-spatial conditions (Step f).

Figure C.13: Results for the measure *Speaker Recognition Performance*.

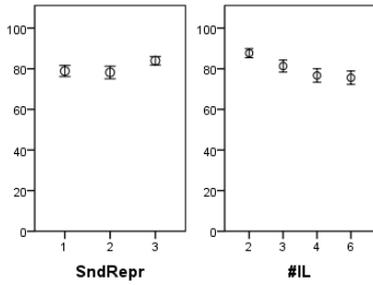
1. Analysis Steps

The present figure shows the data analysis for the measure *Speaker Recognition Confidence* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of #IL if all data over *SndRepr* is compiled and the effect of *SndRepr* if all data over #IL is compiled. Steps (d) to (f) concern the individual effect of #IL analyzed per *SndRepr*. Steps (g) to (i) concern the individual effect of *SndRepr* analyzed per #IL.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per #IL and *SndRepr*. Steps (b), (e), & (h): results of repeated-measures ANOVAs with #IL ∈ [2, 3, 4, 6] and *SndRepr* ∈ [1 = NB/non-spatial, 2 = FB/non-spatial, 3 = FB/spatial] as input, expressed by the *F*-measure and significance level *p*, and using the Greenhouse-Geisser correction when the sphericity criterion for repeated-measures ANOVAs is violated. Steps (c), (f), & (i): results of PostHoc tests (estimated marginal means) with Sidak correction for pairwise comparisons between all pairs of #IL or *SndRepr*, expressed by the significance level *p*; Significant differences (*p* ≤ 0.05) are indicated with an asterisk (\*).

2.1 Main Effects

(a) Errorbars



(b) ANOVA

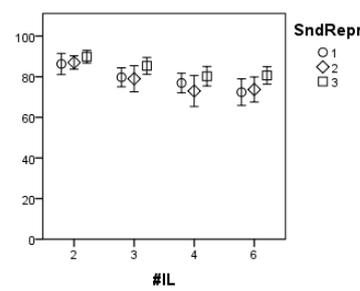
	<i>F</i>	<i>p</i>
<i>SndRepr</i>	7.463	0.002*
#IL	20.448	0.000*
Interaction	0.497	0.809

(c) PostHoc tests

<i>SndRepr</i>	<i>p</i>	#IL	<i>p</i>
1 vs. 2	0.985	2 vs. 3	0.005*
1 vs. 3	0.005*	2 vs. 4	0.000*
2 vs. 3	0.026*	2 vs. 6	0.000*
		3 vs. 4	0.066
		3 vs. 6	0.012*
		4 vs. 6	0.775

2.2 Individual Effect of *SndRepr* per #IL

(d) Errorbars



(e) ANOVA per #IL

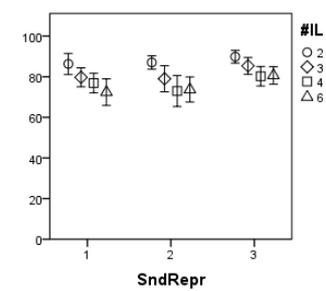
#IL	<i>F</i>	<i>p</i>
2	0.741	0.482
3	2.088	0.136
4	4.074	0.034*
6	7.667	0.001*

(f) PostHoc tests per #IL

#IL	<i>SndRepr</i>	<i>p</i>
2	1 vs. 2	0.999
	1 vs. 3	0.706
	2 vs. 3	0.483
3	1 vs. 2	1.000
	1 vs. 3	0.238
	2 vs. 3	0.341
4	1 vs. 2	0.695
	1 vs. 3	0.059
	2 vs. 3	0.071
6	1 vs. 2	0.867
	1 vs. 3	0.006*
	2 vs. 3	0.030*

2.3 Individual Effect of #IL per *SndRepr*

(g) Errorbars



(h) ANOVA per *SndRepr*

<i>SndRepr</i>	<i>F</i>	<i>p</i>
1	8.464	0.000*
2	10.804	0.000*
3	7.141	0.000*

(i) PostHoc tests per *SndRepr*

<i>SndRepr</i>	#IL	<i>p</i>
1	2 vs. 3	0.301
	2 vs. 4	0.017*
	2 vs. 6	0.001*
	3 vs. 4	0.752
	3 vs. 6	0.123
	4 vs. 6	0.322
2	2 vs. 3	0.045*
	2 vs. 4	0.001*
	2 vs. 6	0.000*
	3 vs. 4	0.069
	3 vs. 6	0.404
	4 vs. 6	1.000
3	2 vs. 3	0.619
	2 vs. 4	0.005*
	2 vs. 6	0.000*
	3 vs. 4	0.415
	3 vs. 6	0.174
	4 vs. 6	0.998

3. Analysis Results:

3.1 Hypothesis H1 (Communication Complexity has an impact on Cognitive Load) is supported, as #IL has a significant impact on *Speaker Recognition Confidence*, whereas this effect is predominantly visible between two-party and multiparty settings.

Rationale – main effect: Values decrease with increasing #IL (Step a). This effect is significant (Step b) and explained mainly by significant differences between two-party (#IL = 2) and multiparty (#IL > 2) groups (Step c).

Rationale – effect per *SndRepr*: Values decrease for all three *SndRepr* (Step g). This is significant (Step h). The differences are mainly explained between two-party and larger multiparty groups, i.e. 2 vs. 4 & 6 (Step i).

3.2 Hypothesis H2 (Technical System Capability has an impact on Cognitive Load) is supported, as *SndRepr*, more specifically the spatial sound reproduction, has a significant impact on *Concentration Effort*.

Rationale – main effect: Values increase with increasing *SndRepr* (Step a). This effect is significant (Step b) and explained by significant differences between spatial and non-spatial sound reproduction, but not between the two non-spatial conditions (Step c).

Rationale – effect per #IL: Values increase for medium to large groups #IL = 4 & 6 (Step d). This is significant (Step e) and is explained by significant differences between spatial and non-spatial sound reproduction, whereas the effect is only for large groups (#IL = 6) strong enough to be detected by the PostHoc tests (Step f).

Figure C.14: Results for the measure *Speaker Recognition Confidence*.

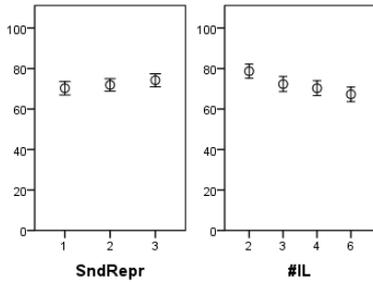
1. Analysis Steps

The present figure shows the data analysis for the measure *Focal Assurance* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of #IL if all data over *SndRepr* is compiled and the effect of *SndRepr* if all data over #IL is compiled. Steps (d) to (f) concern the individual effect of #IL analyzed per *SndRepr*. Steps (g) to (i) concern the individual effect of *SndRepr* analyzed per #IL.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per #IL and *SndRepr*. Steps (b), (e), & (h): results of repeated-measures ANOVAs with #IL ∈ [2,3,4,6] and *SndRepr* ∈ [1 = NB/non-spatial, 2 = FB/non-spatial, 3 = FB/spatial] as input, expressed by the *F*-measure and significance level *p*, and using the Greenhouse-Geisser correction when the sphericity criterion for repeated-measures ANOVAs is violated. Steps (c), (f), & (i): results of PostHoc tests (estimated marginal means) with Sidak correction for pairwise comparisons between all pairs of #IL or *SndRepr*, expressed by the significance level *p*; Significant differences (*p* ≤ 0.05) are indicated with an asterisk (\*).

2.1 Main Effects

(a) Errorbars



(b) ANOVA

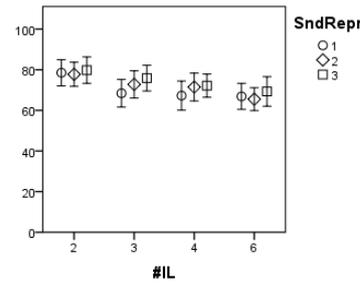
	<i>F</i>	<i>p</i>
<i>SndRepr</i>	2.691	0.079
#IL	8.779	0.000*
Interaction	0.855	0.530

(c) PostHoc tests

<i>SndRepr</i>	<i>p</i>	#IL	<i>p</i>
1 vs. 2	0.645	2 vs. 3	0.109
1 vs. 3	0.133	2 vs. 4	0.000*
2 vs. 3	0.395	2 vs. 6	0.000*
		3 vs. 4	0.967
		3 vs. 6	0.363
		4 vs. 6	0.427

2.2 Individual Effect of *SndRepr* per #IL

(d) Errorbars



(e) ANOVA per #IL

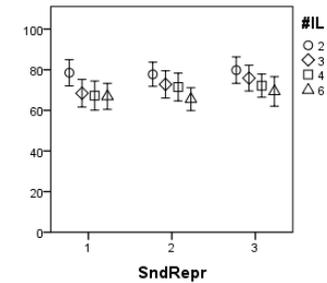
#IL	<i>F</i>	<i>p</i>
2	0.036	0.964
3	2.019	0.145
4	1.398	0.257
6	1.174	0.307

(f) PostHoc tests per #IL

#IL	<i>SndRepr</i>	<i>p</i>
2	1 vs. 2	0.999
	1 vs. 3	0.999
	2 vs. 3	0.987
3	1 vs. 2	0.376
	1 vs. 3	0.303
	2 vs. 3	0.894
4	1 vs. 2	0.456
	1 vs. 3	0.416
	2 vs. 3	0.995
6	1 vs. 2	0.919
	1 vs. 3	0.833
	2 vs. 3	0.144

2.3 Individual Effect of #IL per *SndRepr*

(g) Errorbars



(h) ANOVA per *SndRepr*

<i>SndRepr</i>	<i>F</i>	<i>p</i>
1	5.008	0.008*
2	6.177	0.001*
3	2.923	0.041*

(i) PostHoc tests per *SndRepr*

<i>SndRepr</i>	#IL	<i>p</i>
1	2 vs. 3	0.027*
	2 vs. 4	0.011*
	2 vs. 6	0.013*
	3 vs. 4	1.000
2	3 vs. 6	0.999
	4 vs. 6	1.000
	2 vs. 3	0.532
	2 vs. 4	0.117
3	2 vs. 6	0.000*
	3 vs. 4	0.999
	3 vs. 6	0.166
	4 vs. 6	0.236
	2 vs. 3	0.959
	2 vs. 4	0.305
	2 vs. 6	0.138
	3 vs. 4	0.825
3 vs. 6	0.180	
4 vs. 6	0.963	

3. Analysis Results:

3.1 Hypothesis H1 (Communication Complexity has an impact on Cognitive Load) is supported, as #IL has a significant impact on *Focal Assurance*, whereas this effect is predominantly visible between two-party and larger multiparty groups, and decreases with improving sound reproduction method.

Rationale – main effect: Values decrease with increasing #IL (Step a). This effect is significant (Step b) and explained by significant differences between two-party (#IL = 2) and medium to large multiparty groups (#IL = 4 & 6) (Step c).

Rationale – effect per *SndRepr*: Values decrease for all three *SndRepr* (Step g). This is significant (Step h). The number of pairs per condition that explain the differences is decreasing with increasing *SndRepr* (Step i).

3.2 Hypothesis H2 (Technical System Capability has an impact on Cognitive Load) is not supported, as *SndRepr* has no significant impact on *Concentration Effort*.

Rationale – main effect: Although values increase with increasing *SndRepr* (Step a), this effect is not significant (Steps b & c).

Rationale – effect per #IL: Although values tend to increase with increasing *SndRepr* for all #IL (Step d), this effect is not significant (Steps e & f).

Figure C.15: Results for the measure *Focal Assurance*.

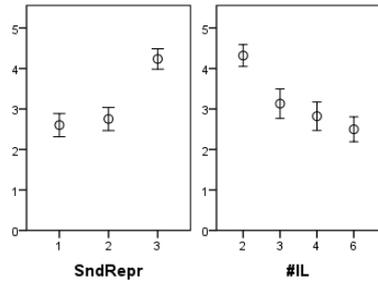
1. Analysis Steps

The present figure shows the data analysis for the measure *Speaker Recognition Effort* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of #IL if all data over *SndRepr* is compiled and the effect of *SndRepr* if all data over #IL is compiled. Steps (d) to (f) concern the individual effect of #IL analyzed per *SndRepr*. Steps (g) to (i) concern the individual effect of *SndRepr* analyzed per #IL.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per #IL and *SndRepr*. Steps (b), (e), & (h): results of repeated-measures ANOVAs with #IL ∈ [2, 3, 4, 6] and *SndRepr* ∈ [1 = NB/non-spatial, 2 = FB/non-spatial, 3 = FB/spatial] as input, expressed by the *F*-measure and significance level *p*, and using the Greenhouse-Geisser correction when the sphericity criterion for repeated-measures ANOVAs is violated. Steps (c), (f), & (i): results of PostHoc tests (estimated marginal means) with Sidak correction for pairwise comparisons between all pairs of #IL or *SndRepr*, expressed by the significance level *p*; Significant differences (*p* ≤ 0.05) are indicated with an asterisk (\*).

2.1 Main Effects

(a) Errorbars



(b) ANOVA

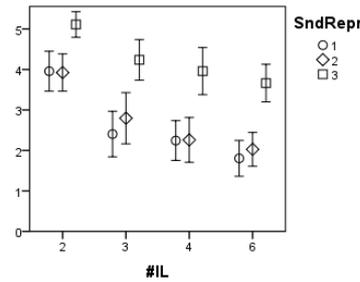
	<i>F</i>	<i>p</i>
<i>SndRepr</i>	28.399	0.000*
#IL	30.949	0.000*
Interaction	1.191	0.315

(c) PostHoc tests

<i>SndRepr</i>	<i>p</i>	#IL	<i>p</i>
1 vs. 2	0.841	2 vs. 3	0.000*
1 vs. 3	0.000*	2 vs. 4	0.000*
2 vs. 3	0.000*	2 vs. 6	0.000*
		3 vs. 4	0.903
		3 vs. 6	0.059
		4 vs. 6	0.177

2.2 Individual Effect of *SndRepr* per #IL

(d) Errorbars



(e) ANOVA per #IL

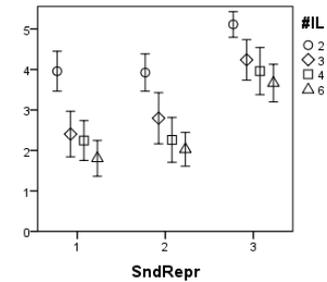
#IL	<i>F</i>	<i>p</i>
2	18.192	0.000*
3	12.897	0.000*
4	16.289	0.000*
6	23.418	0.000*

(f) PostHoc tests per #IL

#IL	<i>SndRepr</i>	<i>p</i>
2	1 vs. 2	0.952
	1 vs. 3	0.000*
	2 vs. 3	0.000*
3	1 vs. 2	0.532
	1 vs. 3	0.000*
	2 vs. 3	0.011*
4	1 vs. 2	1.000
	1 vs. 3	0.001*
	2 vs. 3	0.001*
6	1 vs. 2	0.430
	1 vs. 3	0.000*
	2 vs. 3	0.000*

2.3 Individual Effect of #IL per *SndRepr*

(g) Errorbars



(h) ANOVA per *SndRepr*

<i>SndRepr</i>	<i>F</i>	<i>p</i>
1	18.616	0.000*
2	17.229	0.000*
3	11.584	0.000*

(i) PostHoc tests per *SndRepr*

<i>SndRepr</i>	#IL	<i>p</i>
1	2 vs. 3	0.000*
	2 vs. 4	0.000*
	2 vs. 6	0.000*
	3 vs. 4	0.998
2	3 vs. 6	0.300
	4 vs. 6	0.279
	2 vs. 3	0.024*
	2 vs. 4	0.000*
3	2 vs. 6	0.000*
	3 vs. 4	0.250
	3 vs. 6	0.034*
	4 vs. 6	0.824
	2 vs. 3	0.007*
	2 vs. 4	0.001*
	2 vs. 6	0.000*
	3 vs. 4	0.998
3 vs. 6	0.549	
	4 vs. 6	0.662

3. Analysis Results:

3.1 Hypothesis H1 (Communication Complexity has an impact on Cognitive Load) is supported, as #IL has a significant impact on *Speaker Recognition Effort*, whereas this effect is visible between two-party and multiparty settings.

Rationale – main effect: Values decrease with increasing #IL (Step a). This effect is significant (Step b) and explained by significant differences between two-party (#IL = 2) and multiparty (#IL > 2) groups (Step c).

Rationale – effect per *SndRepr*: Values decrease for all three *SndRepr* (Step g). This is significant (Step h) and explained between two-party and multiparty groups (Step i).

3.2 Hypothesis H2 (Technical System Capability has an impact on Cognitive Load) is supported, as *SndRepr*, more specifically the spatial sound reproduction, has a significant impact on *Speaker Recognition Effort*.

Rationale – main effect: Values increase with increasing *SndRepr* (Step a). This effect is significant (Step b) and explained by significant differences between spatial and non-spatial sound reproduction, but not between the two non-spatial conditions (Step c).

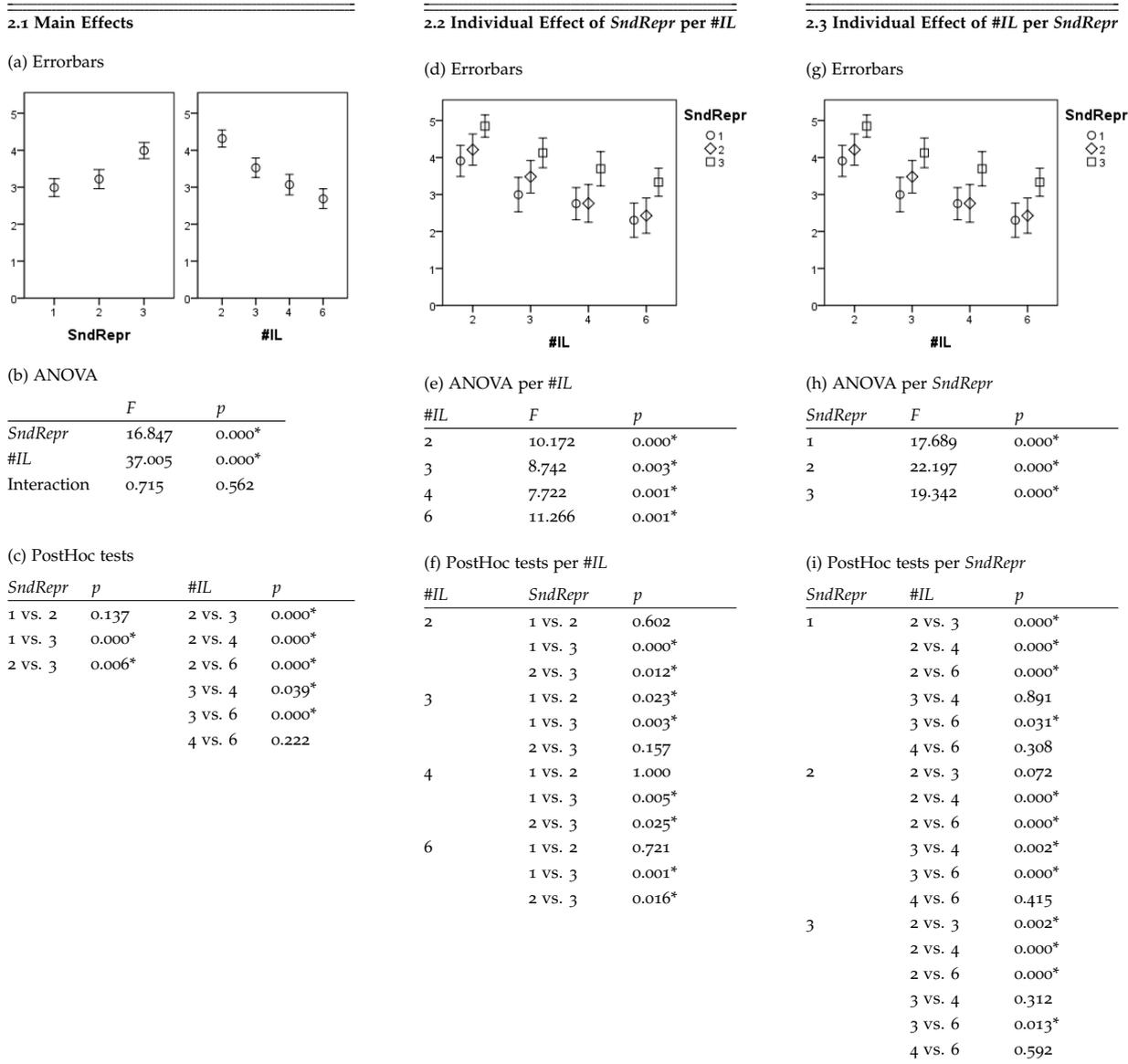
Rationale – effect per #IL: Values increase for all #IL (Step d). This is significant (Step e) and is explained by significant differences between spatial and non-spatial sound reproduction, but not between the two non-spatial conditions (Step f).

Figure C.16: Results for the measure *Speaker Recognition Effort*.

1. Analysis Steps

The present figure shows the data analysis for the measure *Topic Comprehension Effort* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of #IL if all data over *SndRepr* is compiled and the effect of *SndRepr* if all data over #IL is compiled. Steps (d) to (f) concern the individual effect of #IL analyzed per *SndRepr*. Steps (g) to (i) concern the individual effect of *SndRepr* analyzed per #IL.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per #IL and *SndRepr*. Steps (b), (e), & (h): results of repeated-measures ANOVAs with #IL ∈ [2, 3, 4, 6] and *SndRepr* ∈ [1 = NB/non-spatial, 2 = FB/non-spatial, 3 = FB/spatial] as input, expressed by the *F*-measure and significance level *p*, and using the Greenhouse-Geisser correction when the sphericity criterion for repeated-measures ANOVAs is violated. Steps (c), (f), & (i): results of PostHoc tests (estimated marginal means) with Sidak correction for pairwise comparisons between all pairs of #IL or *SndRepr*, expressed by the significance level *p*; Significant differences (*p* ≤ 0.05) are indicated with an asterisk (\*).



3. Analysis Results:

3.1 Hypothesis H1 (Communication Complexity has an impact on Cognitive Load) is supported, as #IL has a significant impact on *Topic Comprehension Effort*.

Rationale – main effect: Values decrease with increasing #IL (Step a). This effect is significant (Step b) and explained by significant differences between all expect one pair of #IL (Step c).

Rationale – effect per *SndRepr*: Values decrease for all three *SndRepr* (Step g). This is significant (Step h) and almost all steps between #IL are large enough to be significant as well (Step i).

3.2 Hypothesis H2 (Technical System Capability has an impact on Cognitive Load) is supported, as *SndRepr*, more specifically the spatial sound reproduction, has a significant impact on *Topic Comprehension Effort*.

Rationale – main effect: Values increase with increasing *SndRepr* (Step a). This effect is significant (Step b) and explained by significant differences between spatial and non-spatial sound reproduction, but not between the two non-spatial conditions (Step c).

Rationale – effect per #IL: Values increase for all four #IL (Step d). This is significant (Step e) and is in most cases explained by significant differences between spatial and non-spatial sound reproduction, but not between the two non-spatial conditions (Step f).

Figure C.17: Results for the measure *Topic Comprehension Effort*.

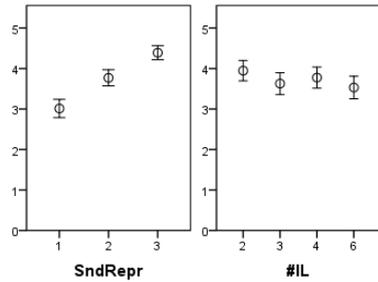
1. Analysis Steps

The present figure shows the data analysis for the measure *Connection Quality* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of *#IL* if all data over *SndRepr* is compiled and the effect of *SndRepr* if all data over *#IL* is compiled. Steps (d) to (f) concern the individual effect of *#IL* analyzed per *SndRepr*. Steps (g) to (i) concern the individual effect of *SndRepr* analyzed per *#IL*.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per *#IL* and *SndRepr*. Steps (b), (e), & (h): results of repeated-measures ANOVAs with *#IL* ∈ [2,3,4,6] and *SndRepr* ∈ [1 = NB/non-spatial, 2 = FB/non-spatial, 3 = FB/spatial] as input, expressed by the *F*-measure and significance level *p*, and using the Greenhouse-Geisser correction when the sphericity criterion for repeated-measures ANOVAs is violated. Steps (c), (f), & (i): results of PostHoc tests (estimated marginal means) with Sidak correction for pairwise comparisons between all pairs of *#IL* or *SndRepr*, expressed by the significance level *p*; Significant differences (*p* ≤ 0.05) are indicated with an asterisk (\*).

2.1 Main Effects

(a) Errorbars



(b) ANOVA

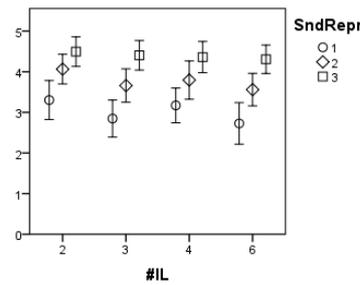
	<i>F</i>	<i>p</i>
<i>SndRepr</i>	24.731	0.000*
<i>#IL</i>	2.553	0.063
Interaction	0.468	0.831

(c) PostHoc tests

<i>SndRepr</i>	<i>p</i>	<i>#IL</i>	<i>p</i>
1 vs. 2	0.003*	2 vs. 3	0.300
1 vs. 3	0.000*	2 vs. 4	0.631
2 vs. 3	0.002*	2 vs. 6	0.119
		3 vs. 4	0.955
		3 vs. 6	0.998
		4 vs. 6	0.845

2.2 Individual Effect of *SndRepr* per *#IL*

(d) Errorbars



(e) ANOVA per *#IL*

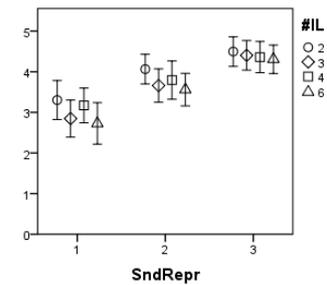
<i>#IL</i>	<i>F</i>	<i>p</i>
2	12.244	0.000*
3	22.546	0.000*
4	9.136	0.000*
6	17.043	0.000*

(f) PostHoc tests per *#IL*

<i>#IL</i>	<i>SndRepr</i>	<i>p</i>
2	1 vs. 2	0.045*
	1 vs. 3	0.001*
	2 vs. 3	0.038*
3	1 vs. 2	0.001*
	1 vs. 3	0.000*
	2 vs. 3	0.009*
4	1 vs. 2	0.188
	1 vs. 3	0.001*
	2 vs. 3	0.067
6	1 vs. 2	0.017*
	1 vs. 3	0.000*
	2 vs. 3	0.016*

2.3 Individual Effect of *#IL* per *SndRepr*

(g) Errorbars



(h) ANOVA per *SndRepr*

<i>SndRepr</i>	<i>F</i>	<i>p</i>
1	2.533	0.064
2	2.082	0.110
3	0.294	0.829

(i) PostHoc tests per *SndRepr*

<i>SndRepr</i>	<i>#IL</i>	<i>p</i>
1	2 vs. 3	0.065
	2 vs. 4	0.998
	2 vs. 6	0.073
	3 vs. 4	0.826
2	3 vs. 6	0.995
	4 vs. 6	0.416
	2 vs. 3	0.411
	2 vs. 4	0.789
3	2 vs. 6	0.157
	3 vs. 4	0.974
	3 vs. 6	0.998
	4 vs. 6	0.886
	2 vs. 3	0.982
	2 vs. 4	0.922
	2 vs. 6	0.970
	3 vs. 4	1.000
3 vs. 6	1.000	
4 vs. 6	1.000	

3. Analysis Results:

3.1 Hypothesis H<sub>3</sub> (Communication Complexity has an impact on Speech Communication Quality) is not supported, as *#IL* has no significant impact on *Connection Quality*.

Rationale – main effect: Although values decrease with increasing *#IL* (Step a), this effect is not significant (Steps b & c).

Rationale – effect per *SndRepr*: Although values decrease with increasing *#IL* for all three *SndRepr* (Step g), this effect is not significant (Steps h & i).

3.2 Hypothesis H<sub>4</sub> (Technical System Capability has an impact on Speech Communication Quality) is supported, as *SndRepr* has a significant impact on *Connection Quality*.

Rationale – main effect: Values increase with increasing *SndRepr* (Step a). This effect is significant (Step b) and explained by significant differences between all sound reproduction methods (Step c).

Rationale – effect per *#IL*: Values increase for all four *#IL* (Step d). This is significant (Step e) and all steps except one between *SndRepr* are large enough to be significant as well (Step f).

Figure C.18: Results for the measure *Connection Quality*.

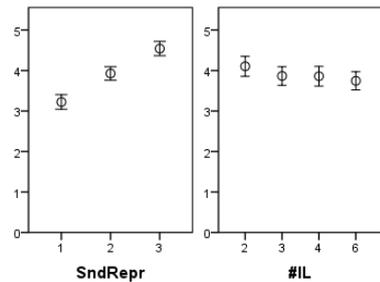
1. Analysis Steps

The present figure shows the data analysis for the measure *Speech Intelligibility* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of #IL if all data over *SndRepr* is compiled and the effect of *SndRepr* if all data over #IL is compiled. Steps (d) to (f) concern the individual effect of #IL analyzed per *SndRepr*. Steps (g) to (i) concern the individual effect of *SndRepr* analyzed per #IL.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per #IL and *SndRepr*. Steps (b), (e), & (h): results of repeated-measures ANOVAs with #IL ∈ [2,3,4,6] and *SndRepr* ∈ [1 = NB/non-spatial, 2 = FB/non-spatial, 3 = FB/spatial] as input, expressed by the *F*-measure and significance level *p*, and using the Greenhouse-Geisser correction when the sphericity criterion for repeated-measures ANOVAs is violated. Steps (c), (f), & (i): results of PostHoc tests (estimated marginal means) with Sidak correction for pairwise comparisons between all pairs of #IL or *SndRepr*, expressed by the significance level *p*; Significant differences (*p* ≤ 0.05) are indicated with an asterisk (\*).

2.1 Main Effects

(a) Errorbars



(b) ANOVA

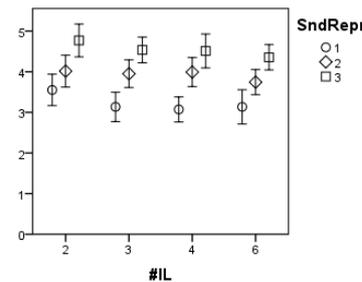
	<i>F</i>	<i>p</i>
<i>SndRepr</i>	34.656	0.000*
#IL	2.716	0.952
Interaction	1.130	0.348

(c) PostHoc tests

<i>SndRepr</i>	<i>p</i>	#IL	<i>p</i>
1 vs. 2	0.000*	2 vs. 3	0.562
1 vs. 3	0.000*	2 vs. 4	0.167
2 vs. 3	0.001*	2 vs. 6	0.077
		3 vs. 4	1.000
		3 vs. 6	0.850
		4 vs. 6	0.970

2.2 Individual Effect of *SndRepr* per #IL

(d) Errorbars



(e) ANOVA per #IL

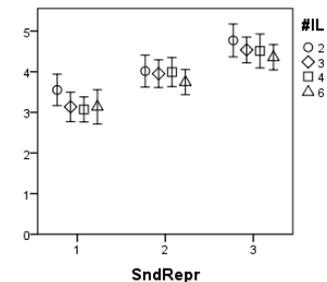
#IL	<i>F</i>	<i>p</i>
2	15.357	0.000*
3	25.790	0.000*
4	24.479	0.000*
6	14.623	0.000*

(f) PostHoc tests per #IL

#IL	<i>SndRepr</i>	<i>p</i>
2	1 vs. 2	0.326
	1 vs. 3	0.000*
	2 vs. 3	0.000*
3	1 vs. 2	0.000*
	1 vs. 3	0.000*
	2 vs. 3	0.072
4	1 vs. 2	0.000*
	1 vs. 3	0.000*
	2 vs. 3	0.085
6	1 vs. 2	0.036*
	1 vs. 3	0.000*
	2 vs. 3	0.010*

2.3 Individual Effect of #IL per *SndRepr*

(g) Errorbars



(h) ANOVA per *SndRepr*

<i>SndRepr</i>	<i>F</i>	<i>p</i>
1	2.316	0.103
2	0.862	0.465
3	2.011	0.138

(i) PostHoc tests per *SndRepr*

<i>SndRepr</i>	#IL	<i>p</i>
1	2 vs. 3	0.118
	2 vs. 4	0.179
	2 vs. 6	0.366
	3 vs. 4	1.000
2	3 vs. 6	1.000
	4 vs. 6	0.999
	2 vs. 3	1.000
	2 vs. 4	1.000
3	2 vs. 6	0.726
	3 vs. 4	1.000
	3 vs. 6	0.856
	4 vs. 6	0.741
	2 vs. 3	0.704
	2 vs. 4	0.302
	2 vs. 6	0.197
	3 vs. 4	1.000
3 vs. 6	0.941	
4 vs. 6	0.932	

3. Analysis Results:

3.1 Hypothesis H<sub>3</sub> (Communication Complexity has an impact on Speech Communication Quality) is not supported, as #IL has no significant impact on *Speech Intelligibility*.

Rationale – main effect: Although values decrease with increasing #IL (Step a), this effect is not significant (Steps b & c).

Rationale – effect per *SndRepr*: Although values decrease with increasing #IL for all three *SndRepr* (Step g), this effect is not significant (Steps h & i).

3.2 Hypothesis H<sub>4</sub> (Technical System Capability has an impact on Speech Communication Quality) is supported, as *SndRepr* has a significant impact on *Speech Intelligibility*.

Rationale – main effect: Values increase with increasing *SndRepr* (Step a). This effect is significant (Step b) and explained by significant differences between all sound reproduction methods (Step c).

Rationale – effect per #IL: Values increase for all four #IL (Step d). This is significant (Step e) and almost all steps between *SndRepr* are large enough to be significant as well (Step f).

Figure C.19: Results for the measure *Speech Intelligibility*.

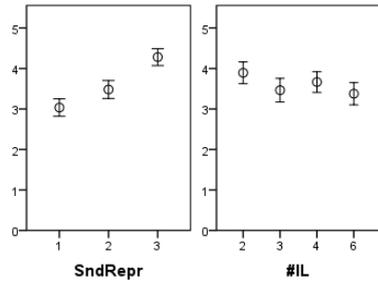
1. Analysis Steps

The present figure shows the data analysis for the measure *Overall Quality* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of #IL if all data over *SndRepr* is compiled and the effect of *SndRepr* if all data over #IL is compiled. Steps (d) to (f) concern the individual effect of #IL analyzed per *SndRepr*. Steps (g) to (i) concern the individual effect of *SndRepr* analyzed per #IL.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per #IL and *SndRepr*. Steps (b), (e), & (h): results of repeated-measures ANOVAs with #IL ∈ [2, 3, 4, 6] and *SndRepr* ∈ [1 = NB/non-spatial, 2 = FB/non-spatial, 3 = FB/spatial] as input, expressed by the *F*-measure and significance level *p*, and using the Greenhouse-Geisser correction when the sphericity criterion for repeated-measures ANOVAs is violated. Steps (c), (f), & (i): results of PostHoc tests (estimated marginal means) with Sidak correction for pairwise comparisons between all pairs of #IL or *SndRepr*, expressed by the significance level *p*; Significant differences (*p* ≤ 0.05) are indicated with an asterisk (\*).

2.1 Main Effects

(a) Errorbars



(b) ANOVA

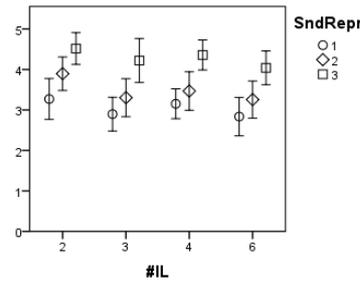
	<i>F</i>	<i>p</i>
<i>SndRepr</i>	13.474	0.000*
#IL	3.853	0.014*
Interaction	0.115	0.994

(c) PostHoc tests

<i>SndRepr</i>	<i>p</i>	#IL	<i>p</i>
1 vs. 2	0.154	2 vs. 3	0.184
1 vs. 3	0.001*	2 vs. 4	0.178
2 vs. 3	0.018*	2 vs. 6	0.029*
		3 vs. 4	0.842
		3 vs. 6	1.000
		4 vs. 6	0.806

2.2 Individual Effect of *SndRepr* per #IL

(d) Errorbars



(e) ANOVA per #IL

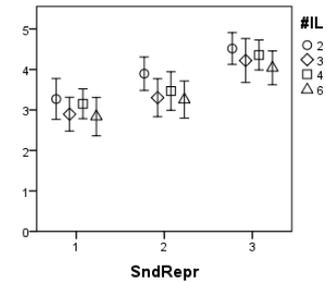
#IL	<i>F</i>	<i>p</i>
2	10.400	0.000*
3	9.715	0.000*
4	13.527	0.000*
6	9.654	0.000*

(f) PostHoc tests per #IL

#IL	<i>SndRepr</i>	<i>p</i>
2	1 vs. 2	0.230
	1 vs. 3	0.002*
	2 vs. 3	0.007*
3	1 vs. 2	0.314
	1 vs. 3	0.002*
	2 vs. 3	0.033*
4	1 vs. 2	0.155
	1 vs. 3	0.000*
	2 vs. 3	0.021*
6	1 vs. 2	0.218
	1 vs. 3	0.001*
	2 vs. 3	0.062

2.3 Individual Effect of #IL per *SndRepr*

(g) Errorbars



(h) ANOVA per *SndRepr*

<i>SndRepr</i>	<i>F</i>	<i>p</i>
1	2.399	0.076
2	4.481	0.006*
3	1.639	0.202

(i) PostHoc tests per *SndRepr*

<i>SndRepr</i>	#IL	<i>p</i>
1	2 vs. 3	0.341
	2 vs. 4	0.551
	2 vs. 6	0.157
	3 vs. 4	0.992
	3 vs. 6	1.000
	4 vs. 6	0.833
2	2 vs. 3	0.049*
	2 vs. 4	0.199
	2 vs. 6	0.018*
	3 vs. 4	0.961
	3 vs. 6	1.000
	4 vs. 6	0.757
3	2 vs. 3	0.868
	2 vs. 4	0.893
	2 vs. 6	0.224
	3 vs. 4	0.979
	3 vs. 6	0.983
	4 vs. 6	0.712

3. Analysis Results:

3.1 Hypothesis H5 (Communication Complexity has an impact on Quality of Experience) is supported, as #IL has a significant impact on *Overall Quality*, noting that the group sizes must differ substantially to explain this impact.

Rationale – main effect: Values decrease with increasing #IL (Step a). This effect is significant (Steps b), but only between the most extreme cases #IL = 2 vs. #IL = 6 (Step c).

Rationale – effect per *SndRepr*: While values decrease with increasing #IL for all three *SndRepr* (Step g), this effect is only significant for non-spatial wideband sound reproduction (Step h) between a few pairs of *SndRepr* (Step i).

3.2 Hypothesis H6 (Technical System Capability has an impact on Quality of Experience) is supported, more specifically the spatial sound reproduction, as *SndRepr* has a significant impact on *Overall Quality*.

Rationale – main effect: Values increase with increasing *SndRepr* (Step a). This effect is significant (Step b) and explained by significant differences between spatial and non-spatial sound reproduction, but not between the two non-spatial conditions (Step c).

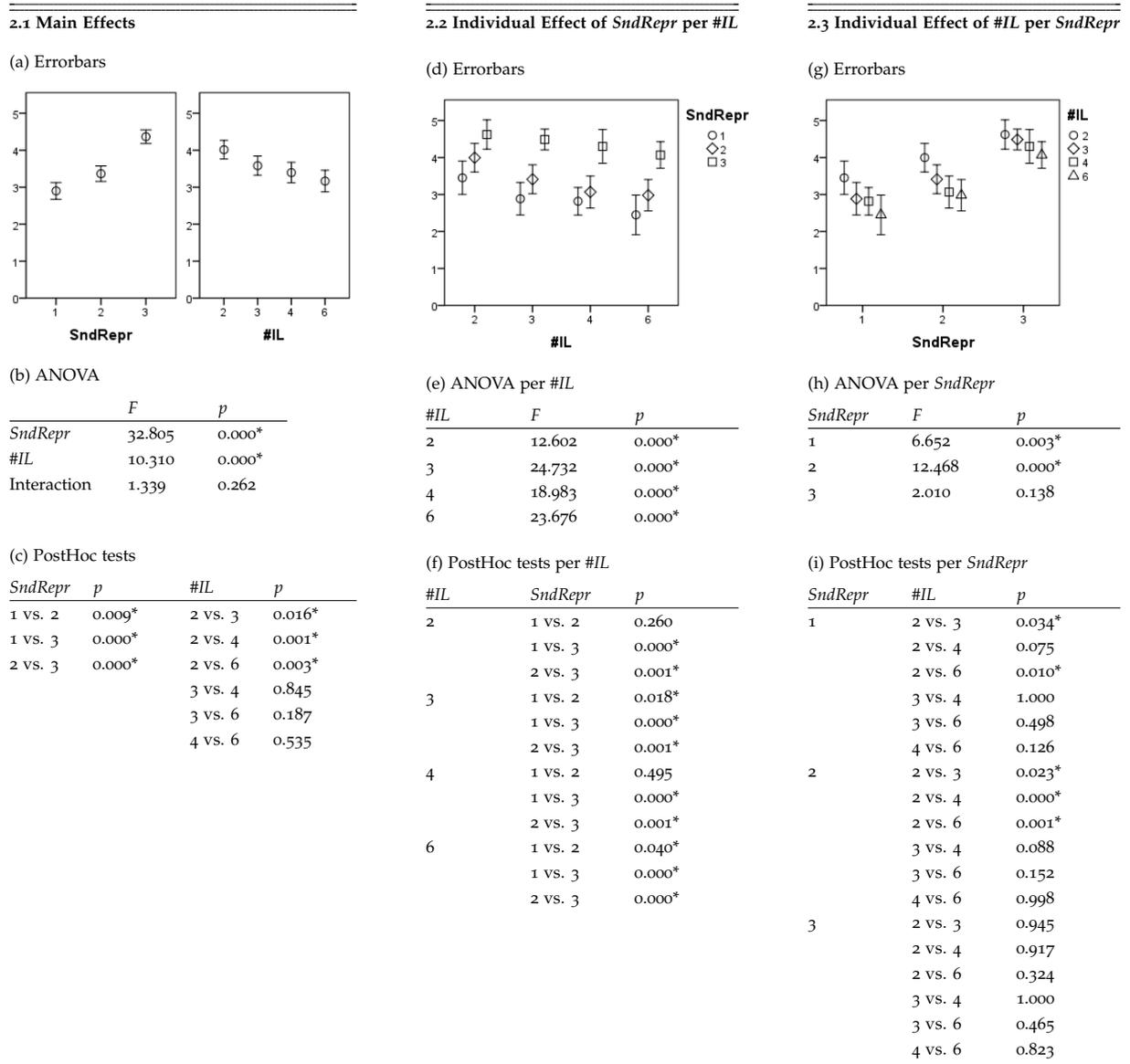
Rationale – effect per #IL: Values increase for all four #IL (Step d). This is significant (Step e) and is in mainly explained by significant differences between spatial and non-spatial sound reproduction (Step f).

Figure C.20: Results for the measure *Overall Quality*.

1. Analysis Steps

The present figure shows the data analysis for the measure *Satisfaction* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of *#IL* if all data over *SndRepr* is compiled and the effect of *SndRepr* if all data over *#IL* is compiled. Steps (d) to (f) concern the individual effect of *#IL* analyzed per *SndRepr*. Steps (g) to (i) concern the individual effect of *SndRepr* analyzed per *#IL*.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per *#IL* and *SndRepr*. Steps (b), (e), & (h): results of repeated-measures ANOVAs with *#IL* ∈ [2, 3, 4, 6] and *SndRepr* ∈ [1 = NB/non-spatial, 2 = FB/non-spatial, 3 = FB/spatial] as input, expressed by the *F*-measure and significance level *p*, and using the Greenhouse-Geisser correction when the sphericity criterion for repeated-measures ANOVAs is violated. Steps (c), (f), & (i): results of PostHoc tests (estimated marginal means) with Sidak correction for pairwise comparisons between all pairs of *#IL* or *SndRepr*, expressed by the significance level *p*; Significant differences ( $p \leq 0.05$ ) are indicated with an asterisk (\*).



3. Analysis Results:

3.1 Hypothesis H5 (Communication Complexity has an impact on Quality of Experience) is supported, as *#IL* has a significant impact on *Satisfaction* between two-party and multiparty groups.

Rationale – main effect: Values decrease with increasing *#IL* (Step a). This effect is significant (Step b) and explained by significant differences between two-party (*#IL* = 2) and multiparty (*#IL* > 2) groups (Step c).

Rationale – effect per *SndRepr*: While values decrease with increasing *#IL* for all three *SndRepr* (Step g), this effect is only significant for non-spatial sound reproduction (Step h). It is mainly explained by significant differences between two-party and multiparty groups (Step i).

3.2 Hypothesis H6 (Technical System Capability has an impact on Quality of Experience) is supported, as *SndRepr* has a significant impact on *Satisfaction*.

Rationale – main effect: Values increase with increasing *SndRepr* (Step a). This effect is significant (Step b) and is explained by significant differences between all pairs of *SndRepr* (Step c).

Rationale – effect per *#IL*: Values increase for all four *#IL* (Step d). This is significant (Step e) and is explained by significant differences between almost all pairs of *SndRepr* (Step f).

Figure C.21: Results for the measure *Satisfaction*.

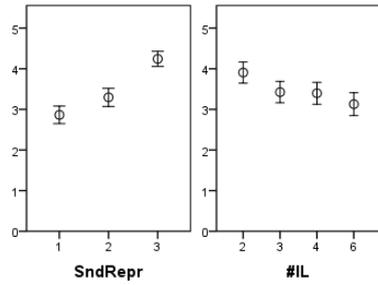
1. Analysis Steps

The present figure shows the data analysis for the measure *Pleasantness* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of *#IL* if all data over *SndRepr* is compiled and the effect of *SndRepr* if all data over *#IL* is compiled. Steps (d) to (f) concern the individual effect of *#IL* analyzed per *SndRepr*. Steps (g) to (i) concern the individual effect of *SndRepr* analyzed per *#IL*.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per *#IL* and *SndRepr*. Steps (b), (e), & (h): results of repeated-measures ANOVAs with *#IL* ∈ [2, 3, 4, 6] and *SndRepr* ∈ [1 = NB/non-spatial, 2 = FB/non-spatial, 3 = FB/spatial] as input, expressed by the *F*-measure and significance level *p*, and using the Greenhouse-Geisser correction when the sphericity criterion for repeated-measures ANOVAs is violated. Steps (c), (f), & (i): results of PostHoc tests (estimated marginal means) with Sidak correction for pairwise comparisons between all pairs of *#IL* or *SndRepr*, expressed by the significance level *p*; Significant differences (*p* ≤ 0.05) are indicated with an asterisk (\*).

2.1 Main Effects

(a) Errorbars



(b) ANOVA

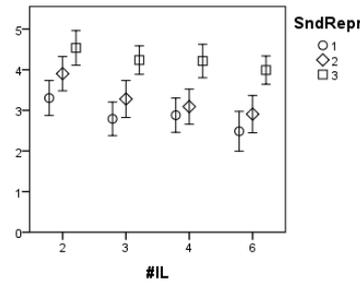
	<i>F</i>	<i>p</i>
<i>SndRepr</i>	26.049	0.000*
<i>#IL</i>	8.453	0.000*
Interaction	1.013	0.419

(c) PostHoc tests

<i>SndRepr</i>	<i>p</i>	<i>#IL</i>	<i>p</i>
1 vs. 2	0.007*	2 vs. 3	0.008*
1 vs. 3	0.000*	2 vs. 4	0.007*
2 vs. 3	0.001*	2 vs. 6	0.005*
		3 vs. 4	1.000
		3 vs. 6	0.672
		4 vs. 6	0.274

2.2 Individual Effect of *SndRepr* per *#IL*

(d) Errorbars



(e) ANOVA per *#IL*

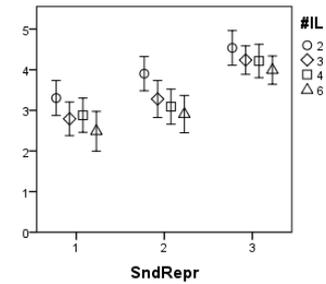
<i>#IL</i>	<i>F</i>	<i>p</i>
2	13.679	0.000*
3	16.131	0.000*
4	17.318	0.000*
6	21.368	0.000*

(f) PostHoc tests per *#IL*

<i>#IL</i>	<i>SndRepr</i>	<i>p</i>
2	1 vs. 2	0.129
	1 vs. 3	0.000*
	2 vs. 3	0.003*
3	1 vs. 2	0.016*
	1 vs. 3	0.000*
	2 vs. 3	0.017*
4	1 vs. 2	0.573
	1 vs. 3	0.000*
	2 vs. 3	0.002*
6	1 vs. 2	0.012*
	1 vs. 3	0.000*
	2 vs. 3	0.002*

2.3 Individual Effect of *#IL* per *SndRepr*

(g) Errorbars



(h) ANOVA per *SndRepr*

<i>SndRepr</i>	<i>F</i>	<i>p</i>
1	2.399	0.076
2	4.481	0.006*
3	1.639	0.202

(i) PostHoc tests per *SndRepr*

<i>SndRepr</i>	<i>#IL</i>	<i>p</i>
1	2 vs. 3	0.019*
	2 vs. 4	0.316
	2 vs. 6	0.008*
	3 vs. 4	0.999
2	3 vs. 6	0.744
	4 vs. 6	0.153
	2 vs. 3	0.026*
	2 vs. 4	0.005*
3	2 vs. 6	0.001*
	3 vs. 4	0.873
	3 vs. 6	0.312*
	4 vs. 6	0.926
	2 vs. 3	0.368
	2 vs. 4	0.886
	2 vs. 6	0.328
	3 vs. 4	0.959
3 vs. 6	0.964	
4 vs. 6	0.770	

3. Analysis Results:

3.1 Hypothesis H5 (Communication Complexity has an impact on Quality of Experience) is supported, as *#IL* has a significant impact on *Pleasantness* between two-party and multiparty groups.

Rationale – main effect: Values decrease with increasing *#IL* (Step a). This effect is significant (Step b) and explained by significant differences between two-party (*#IL* = 2) and multiparty (*#IL* > 2) groups (Step c).

Rationale – effect per *SndRepr*: While values decrease with increasing *#IL* for all three *SndRepr* (Step g), this effect is only significant for non-spatial wideband sound reproduction (Step h). It is mainly explained by significant differences between two-party and multiparty groups (Step i). There is a tendency that this holds also for non-spatial narrowband sound reproduction, but results are inconsistent (Steps h & i).

3.2 Hypothesis H6 (Technical System Capability has an impact on Quality of Experience) is supported, as *SndRepr* has a significant impact on *Pleasantness*.

Rationale – main effect: Values increase with increasing *SndRepr* (Step a). This effect is significant (Step b) and is explained by significant differences between all pairs of *SndRepr* (Step c).

Rationale – effect per *#IL*: Values increase for all four *#IL* (Step d). This is significant (Step e) and is explained by significant differences between almost all pairs of *SndRepr* (Step f).

Figure C.22: Results for the measure *Pleasantness*.

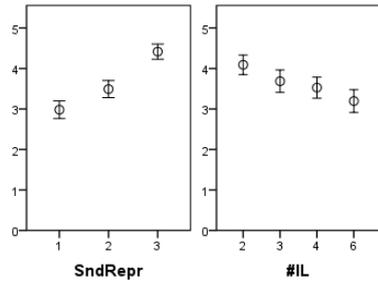
1. Analysis Steps

The present figure shows the data analysis for the measure *Acceptance* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of *#IL* if all data over *SndRepr* is compiled and the effect of *SndRepr* if all data over *#IL* is compiled. Steps (d) to (f) concern the individual effect of *#IL* analyzed per *SndRepr*. Steps (g) to (i) concern the individual effect of *SndRepr* analyzed per *#IL*.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per *#IL* and *SndRepr*. Steps (b), (e), & (h): results of repeated-measures ANOVAs with *#IL* ∈ [2, 3, 4, 6] and *SndRepr* ∈ [1 = NB/non-spatial, 2 = FB/non-spatial, 3 = FB/spatial] as input, expressed by the *F*-measure and significance level *p*, and using the Greenhouse-Geisser correction when the sphericity criterion for repeated-measures ANOVAs is violated. Steps (c), (f), & (i): results of PostHoc tests (estimated marginal means) with Sidak correction for pairwise comparisons between all pairs of *#IL* or *SndRepr*, expressed by the significance level *p*; Significant differences (*p* ≤ 0.05) are indicated with an asterisk (\*).

2.1 Main Effects

(a) Errorbars



(b) ANOVA

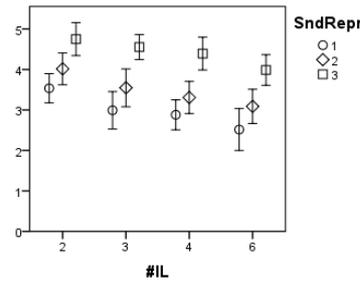
	<i>F</i>	<i>p</i>
<i>SndRepr</i>	27.689	0.000*
<i>#IL</i>	9.588	0.000*
Interaction	0.605	0.726

(c) PostHoc tests

<i>SndRepr</i>	<i>p</i>	<i>#IL</i>	<i>p</i>
1 vs. 2	0.004*	2 vs. 3	0.101
1 vs. 3	0.000*	2 vs. 4	0.001*
2 vs. 3	0.000*	2 vs. 6	0.002*
		3 vs. 4	0.976
		3 vs. 6	0.154
		4 vs. 6	0.213

2.2 Individual Effect of *SndRepr* per *#IL*

(d) Errorbars



(e) ANOVA per *#IL*

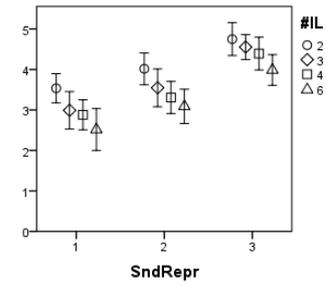
<i>#IL</i>	<i>F</i>	<i>p</i>
2	15.541	0.000*
3	18.022	0.000*
4	19.431	0.000*
6	18.457	0.000*

(f) PostHoc tests per *#IL*

<i>#IL</i>	<i>SndRepr</i>	<i>p</i>
2	1 vs. 2	0.212
	1 vs. 3	0.000*
	2 vs. 3	0.002*
3	1 vs. 2	0.009*
	1 vs. 3	0.000*
	2 vs. 3	0.012*
4	1 vs. 2	0.109
	1 vs. 3	0.000*
	2 vs. 3	0.002*
6	1 vs. 2	0.030*
	1 vs. 3	0.000*
	2 vs. 3	0.003*

2.3 Individual Effect of *#IL* per *SndRepr*

(g) Errorbars



(h) ANOVA per *SndRepr*

<i>SndRepr</i>	<i>F</i>	<i>p</i>
1	7.388	0.001*
2	7.322	0.000*
3	4.136	0.010*

(i) PostHoc tests per *SndRepr*

<i>SndRepr</i>	<i>#IL</i>	<i>p</i>
1	2 vs. 3	0.025*
	2 vs. 4	0.015*
	2 vs. 6	0.002*
	3 vs. 4	0.997
	3 vs. 6	0.490
	4 vs. 6	0.315
2	2 vs. 3	0.322
	2 vs. 4	0.012*
	2 vs. 6	0.001*
	3 vs. 4	0.858
	3 vs. 6	0.106
	4 vs. 6	0.835
3	2 vs. 3	0.814
	2 vs. 4	0.981
	2 vs. 6	0.061
	3 vs. 4	1.000
	3 vs. 6	0.214
	4 vs. 6	0.287

3. Analysis Results:

3.1 Hypothesis H5 (Communication Complexity has an impact on Quality of Experience) is supported, as *#IL* has a significant impact on *Acceptance* between two-party and multiparty groups.

Rationale – main effect: Values decrease with increasing *#IL* (Step a). This effect is significant (Step b) and explained by significant differences between two-party (*#IL* = 2) and multiparty (*#IL* > 2) groups (Step c).

Rationale – effect per *SndRepr*: While values decrease with increasing *#IL* for all three *SndRepr* (Step g), this effect is only significant for non-spatial sound reproduction (Step h). It is mainly explained by significant differences between two-party and multiparty groups (Step i).

3.2 Hypothesis H6 (Technical System Capability has an impact on Quality of Experience) is supported, as *SndRepr* has a significant impact on *Acceptance*.

Rationale – main effect: Values increase with increasing *SndRepr* (Step a). This effect is significant (Step b) and is explained by significant differences between all pairs of *SndRepr* (Step c).

Rationale – effect per *#IL*: Values increase for all four *#IL* (Step d). This is significant (Step e) and is explained by significant differences between almost all pairs of *SndRepr* (Step f).

Figure C.23: Results for the measure *Acceptance*.

## *D*

### *Detailed Results: Study on Impact of Involvement*

#### *What this chapter is about*

The present appendix presents per employed measure the detailed results of the experiment INV on the impact of involvement on quality perception. First, each figure provides an explanation of the analysis steps conducted. Then, each figure shows the results for two analysis levels: a) the main effects (i.e. effect of *Task* across all *SndTrans*, effect of *SndTrans* across all *Tasks*), and (b) the individual effects (effect of *Task* per *SndTrans*, effect of *SndTrans* per *Tasks*). Finally, each figure discusses how the current results relate to the different hypotheses formulated in Section 7.4. The individual results are then compiled in the main text in Section 7.4.

## D.1 Experiment INV

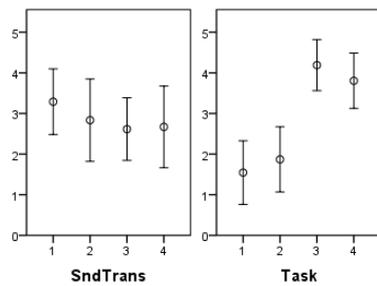
### 1. Analysis Steps

The present figure shows the data analysis for the measure *Perceived Involvement* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of *Task* if all data over *SndTrans* is compiled and the effect of *SndTrans* if all data over *Task* is compiled. Steps (d) to (f) concern the individual effect of *Task* analyzed per *SndTrans*. Steps (g) to (i) concern the individual effect of *SndTrans* analyzed per *Task*.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per *Task* and *SndTrans*. Steps (b), (e), & (h): results of one-way ANOVAs with  $Task \in [1..4]$  and  $SndTrans \in [1 = 0\%, 2 = 5\%, 3 = 10\%, 4 = 15\% \text{ packet loss rate}]$  as input, expressed by the *F*-measure and significance level *p*. Steps (c), (f), & (i): results of PostHoc tests with Sidak correction for pairwise comparisons between all pairs of *Task* or *SndTrans*, expressed by the significance level *p*. Significant differences ( $p \leq 0.05$ ) are indicated with an asterisk (\*).

### 2.1 Main Effects

(a) Errorbars



(b) ANOVA

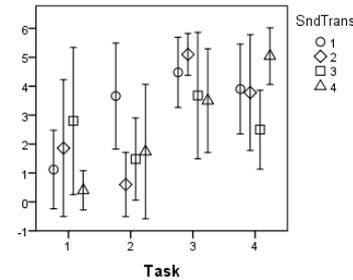
	<i>F</i>	<i>p</i>
<i>Task</i>	19.354	0.000*
<i>SndTrans</i>	1.014	0.392
Interaction	3.577	0.001*

(c) PostHoc tests

<i>SndTrans</i>	<i>p</i>	<i>Task</i>	<i>p</i>
1 vs. 2	0.876	1 vs. 2	0.973
1 vs. 3	0.540	1 vs. 3	0.000*
1 vs. 4	0.634	1 vs. 4	0.000*
2 vs. 3	0.997	2 vs. 3	0.000*
2 vs. 4	0.999	2 vs. 4	0.000*
3 vs. 4	1.000	3 vs. 4	0.940

### 2.2 Individual Effect of *SndTrans* per *Task*

(d) Errorbars



(e) ANOVA per *Task*

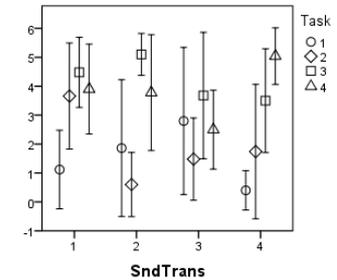
<i>Task</i>	<i>F</i>	<i>p</i>
1	2.263	0.120
2	4.263	0.022*
3	1.696	0.208
4	3.598	0.037*

(f) PostHoc tests per *Task*

<i>Task</i>	<i>SndTrans</i>	<i>p</i>
1	1 vs. 2	0.974
	1 vs. 3	0.472
	1 vs. 4	0.977
	2 vs. 3	0.921
	2 vs. 4	0.623
	3 vs. 4	0.138
2	1 vs. 2	0.019*
	1 vs. 3	0.142
	1 vs. 4	0.242
	2 vs. 3	0.913
	2 vs. 4	0.766
	3 vs. 4	1.000
3	1 vs. 2	0.973
	1 vs. 3	0.914
	1 vs. 4	0.809
	2 vs. 3	0.457
	2 vs. 4	0.329
	3 vs. 4	1.000
4	1 vs. 2	1.000
	1 vs. 3	0.430
	1 vs. 4	0.650
	2 vs. 3	0.529
	2 vs. 4	0.546
	3 vs. 4	0.028*

### 2.3 Individual Effect of *Task* per *SndTrans*

(g) Errorbars



(h) ANOVA per *SndTrans*

<i>SndTrans</i>	<i>F</i>	<i>p</i>
1	7.506	0.002*
2	10.827	0.000*
3	1.677	0.212
4	12.599	0.000*

(i) PostHoc tests per *SndTrans*

<i>SndTrans</i>	<i>Task</i>	<i>p</i>
1	1 vs. 2	0.026*
	1 vs. 3	0.003*
	1 vs. 4	0.014*
	2 vs. 3	0.884
	2 vs. 4	1.000
	3 vs. 4	0.975
2	1 vs. 2	0.653
	1 vs. 3	0.010*
	1 vs. 4	0.217
	2 vs. 3	0.000*
	2 vs. 4	0.011*
	3 vs. 4	0.606
3	1 vs. 2	0.741
	1 vs. 3	0.947
	1 vs. 4	1.000
	2 vs. 3	0.223
	2 vs. 4	0.900
	3 vs. 4	0.824
4	1 vs. 2	0.524
	1 vs. 3	0.009*
	1 vs. 4	0.000*
	2 vs. 3	0.239
	2 vs. 4	0.005*
	3 vs. 4	0.372

### 3. Analysis Results:

**3.1 Hypothesis H1** (Involvement has an impact on Measured Involvement) is supported, as *Task* has a significant impact on *Perceived Involvement*, noting that the difference between conversation and listening-only tasks explains this effect.

Rationale – main effect: Values for the conversation tasks T3 & T4 are higher than the values for the listening-only tasks T1 & T2 (Step a). This effect is significant (Step b) and the PostHoc tests confirm the differences between conversation and listening-only tasks (Step c).

Rationale – effect per *SndTrans*: Values increase with increasing task number in three of the four *SndTrans* (Step g). These effects are significant (Step h), whereas the task pairs showing those differences are not the same across the three *SndTrans* (Step i), i.e. the clear distinction seen for the main effect between conversation and listening-only tasks is only visible for *SndTrans* = 2.

**3.2 Hypothesis H2** (Technical System Capability has an impact on Measured Involvement) is not supported, as *SndTrans* has no consistent significant impact on *Perceived Involvement*.

Rationale – main effect: Values do not change with changing *SndTrans* (Step a) and no significant differences are found (Steps b & c).

Rationale – effect per *Task*: No consistent trend is visible (Step g), while significant effects can be found (Step e), which, however, are very few (i.e. 2 of 18 cases) and not consistent across *Tasks* (Step f).

Figure D.1: Results for the measure *Perceived Involvement*.

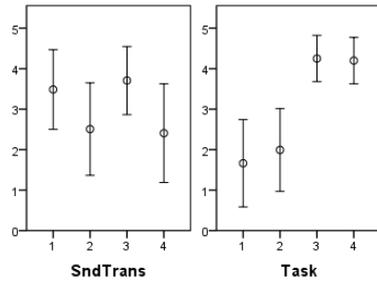
1. Analysis Steps

The present figure shows the data analysis for the measure *Perceived Involvement in Retrospect* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of *Task* if all data over *SndTrans* is compiled and the effect of *SndTrans* if all data over *Task* is compiled. Steps (d) to (f) concern the individual effect of *Task* analyzed per *SndTrans*. Steps (g) to (i) concern the individual effect of *SndTrans* analyzed per *Task*.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per *Task* and *SndTrans*. Steps (b), (e), & (h): results of one-way ANOVAs with *Task* ∈ [1..4] and *SndTrans* ∈ [1 = 0%, 2 = 5%, 3 = 10%, 4 = 15% packet loss rate] as input, expressed by the *F*-measure and significance level *p*. Steps (c), (f), & (i): results of PostHoc tests with Sidak correction for pairwise comparisons between all pairs of *Task* or *SndTrans*, expressed by the significance level *p*. Significant differences ( $p \leq 0.05$ ) are indicated with an asterisk (\*).

2.1 Main Effects

(a) Errorbars



(b) ANOVA

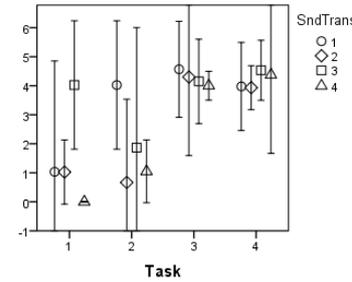
	<i>F</i>	<i>p</i>
<i>Task</i>	22.770	0.000*
<i>SndTrans</i>	4.454	0.009*
Interaction	3.586	0.002*

(c) PostHoc tests

<i>SndTrans</i>	<i>p</i>	<i>Task</i>	<i>p</i>
1 vs. 2	0.156	1 vs. 2	0.972
1 vs. 3	0.996	1 vs. 3	0.000*
1 vs. 4	0.092	1 vs. 4	0.000*
2 vs. 3	0.046*	2 vs. 3	0.000*
2 vs. 4	1.000	2 vs. 4	0.000*
3 vs. 4	0.025*	3 vs. 4	1.000

2.2 Individual Effect of *SndTrans* per *Task*

(d) Errorbars



(e) ANOVA per *Task*

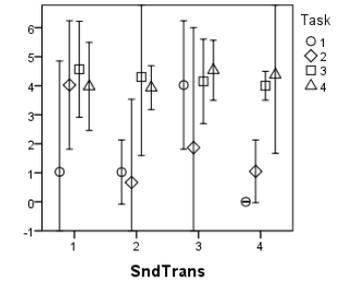
<i>Task</i>	<i>F</i>	<i>p</i>
1	9.300	0.003*
2	5.494	0.017*
3	0.148	0.929
4	0.243	0.864

(f) PostHoc tests per *Task*

<i>Task</i>	<i>SndTrans</i>	<i>p</i>
1	1 vs. 2	1.000
	1 vs. 3	0.030*
	1 vs. 4	0.854
	2 vs. 3	0.018*
	2 vs. 4	0.820
	3 vs. 4	0.004*
2	1 vs. 2	0.032*
	1 vs. 3	0.246
	1 vs. 4	0.041*
	2 vs. 3	0.841
	2 vs. 4	0.999
	3 vs. 4	0.957
3	1 vs. 2	1.000
	1 vs. 3	0.997
	1 vs. 4	0.991
	2 vs. 3	1.000
	2 vs. 4	1.000
	3 vs. 4	1.000
4	1 vs. 2	1.000
	1 vs. 3	0.988
	1 vs. 4	0.997
	2 vs. 3	0.987
	2 vs. 4	0.996
	3 vs. 4	1.000

2.3 Individual Effect of *Task* per *SndTrans*

(g) Errorbars



(h) ANOVA per *SndTrans*

<i>SndTrans</i>	<i>F</i>	<i>p</i>
1	5.576	0.016*
2	9.734	0.003*
3	3.143	0.074
4	15.703	0.000*

(i) PostHoc tests per *SndTrans*

<i>SndTrans</i>	<i>Task</i>	<i>p</i>
1	1 vs. 2	0.048*
	1 vs. 3	0.027*
	1 vs. 4	0.052
	2 vs. 3	0.993
	2 vs. 4	1.000
	3 vs. 4	0.989
2	1 vs. 2	0.999
	1 vs. 3	0.014*
	1 vs. 4	0.044*
	2 vs. 3	0.011*
	2 vs. 4	0.033*
	3 vs. 4	0.999
3	1 vs. 2	0.214
	1 vs. 3	1.000
	1 vs. 4	0.995
	2 vs. 3	0.172
	2 vs. 4	0.119
	3 vs. 4	0.999
4	1 vs. 2	0.743
	1 vs. 3	0.004*
	1 vs. 4	0.001*
	2 vs. 3	0.020*
	2 vs. 4	0.005*
	3 vs. 4	0.998

3. Analysis Results:

3.1 Hypothesis H1 (Involvement has an impact on Measured Involvement) is supported, as *Task* has a significant impact on *Perceived Involvement in Retrospect*, noting that the difference between conversation and listening-only tasks explains this effect.

Rationale – main effect: Values for the conversation tasks T3 & T4 are higher than the values for the listening-only tasks T1 & T2 (Step a). This effect is significant (Step b) and the PostHoc tests confirm the differences between conversation and listening-only tasks (Step c).

Rationale – effect per *SndTrans*: Values increase with increasing task number in three of the four *SndTrans* (Step g). These effects are significant (Step h), whereas the task pairs showing those differences are not the same across the three *SndTrans* (Step i), i.e. the clear distinction seen for the main effect between conversation and listening-only tasks is only visible for *SndTrans* = 2 & 4.

3.2 Hypothesis H2 (Technical System Capability has an impact on Measured Involvement) is not supported, as *SndTrans* has no consistent significant impact on *Perceived Involvement in Retrospect*.

Rationale – main effect: Values do change with changing *SndTrans* (Step a) but are not consistent with increasing impairment, i.e. *SndTrans* = 2 (5% Packet loss) is worse than *SndTrans* = 3 (10% packet loss). Significant effects are found (Step b) and are explained by differences between *SndTrans* = 1 & 3 vs. *SndTrans* = 2 & 4 (Step c).

Rationale – effect per *Task*: No consistent trend is visible (Step g), while significant effects can be found for the listening-only tasks (Step e), which, however, are very few and not consistent across *Tasks* (Step f).

Figure D.2: Results for the measure *Perceived Involvement in Retrospect*.

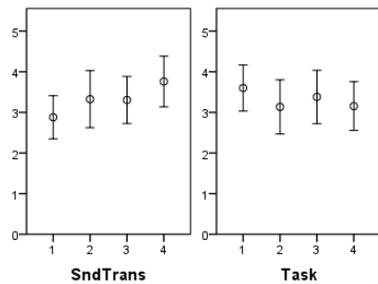
**1. Analysis Steps**

The present figure shows the data analysis for the measure *Influence of Content* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of *Task* if all data over *SndTrans* is compiled and the effect of *SndTrans* if all data over *Task* is compiled. Steps (d) to (f) concern the individual effect of *Task* analyzed per *SndTrans*. Steps (g) to (i) concern the individual effect of *SndTrans* analyzed per *Task*.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per *Task* and *SndTrans*. Steps (b), (e), & (h): results of one-way ANOVAs with *Task* ∈ [1...4] and *SndTrans* ∈ [1 = 0%, 2 = 5%, 3 = 10%, 4 = 15% packet loss rate] as input, expressed by the *F*-measure and significance level *p*. Steps (c), (f), & (i): results of PostHoc tests with Sidak correction for pairwise comparisons between all pairs of *Task* or *SndTrans*, expressed by the significance level *p*. Significant differences ( $p \leq 0.05$ ) are indicated with an asterisk (\*).

**2.1 Main Effects**

(a) Errorbars



(b) ANOVA

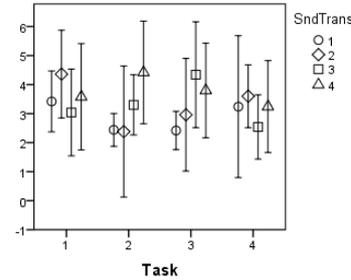
	<i>F</i>	<i>p</i>
<i>Task</i>	0.594	0.621
<i>SndTrans</i>	1.605	0.197
Interaction	1.689	0.110

(c) PostHoc tests

<i>SndTrans</i>	<i>p</i>	<i>Task</i>	<i>p</i>
1 vs. 2	0.850	1 vs. 2	0.823
1 vs. 3	0.876	1 vs. 3	0.995
1 vs. 4	0.177	1 vs. 4	0.850
2 vs. 3	1.000	2 vs. 3	0.991
2 vs. 4	0.863	2 vs. 4	1.000
3 vs. 4	0.837	3 vs. 4	0.994

**2.2 Individual Effect of *SndTrans* per *Task***

(d) Errorbars



(e) ANOVA per *Task*

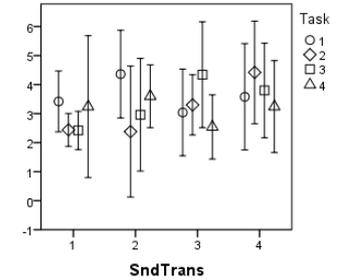
<i>Task</i>	<i>F</i>	<i>p</i>
1	1.057	0.395
2	2.917	0.066
3	2.215	0.126
4	0.557	0.651

(f) PostHoc tests per *Task*

<i>Task</i>	<i>SndTrans</i>	<i>p</i>
1	1 vs. 2	0.801
	1 vs. 3	0.997
	1 vs. 4	1.000
	2 vs. 3	0.479
	2 vs. 4	0.903
	3 vs. 4	0.982
2	1 vs. 2	1.000
	1 vs. 3	0.875
	1 vs. 4	0.132
	2 vs. 3	0.838
	2 vs. 4	0.114
	3 vs. 4	0.686
3	1 vs. 2	0.987
	1 vs. 3	0.173
	1 vs. 4	0.499
	2 vs. 3	0.499
	2 vs. 4	0.898
	3 vs. 4	0.987
4	1 vs. 2	0.999
	1 vs. 3	0.961
	1 vs. 4	1.000
	2 vs. 3	0.784
	2 vs. 4	0.999
	3 vs. 4	0.961

**2.3 Individual Effect of *Task* per *SndTrans***

(g) Errorbars



(h) ANOVA per *SndTrans*

<i>SndTrans</i>	<i>F</i>	<i>p</i>
1	1.083	0.384
2	1.813	0.185
3	2.253	0.122
4	0.652	0.593

(i) PostHoc tests per *SndTrans*

<i>SndTrans</i>	<i>Task</i>	<i>p</i>
1	1 vs. 2	0.714
	1 vs. 3	0.696
	1 vs. 4	1.000
	2 vs. 3	1.000
	2 vs. 4	0.859
	3 vs. 4	0.845
2	1 vs. 2	0.225
	1 vs. 3	0.587
	1 vs. 4	0.957
	2 vs. 3	0.989
	2 vs. 4	0.720
	3 vs. 4	0.981
3	1 vs. 2	1.000
	1 vs. 3	0.424
	1 vs. 4	0.983
	2 vs. 3	0.661
	2 vs. 4	0.886
	3 vs. 4	0.129
4	1 vs. 2	0.923
	1 vs. 3	1.000
	1 vs. 4	0.999
	2 vs. 3	0.982
	2 vs. 4	0.725
	3 vs. 4	0.989

**3. Analysis Results:**

**3.1 Hypothesis H1** (Involvement has an impact on Measured Involvement) is not supported, as *Task* has a no significant impact on *Influence of Content*.

Rationale – main effect: Values slightly increase with increasing *Task* number (Step a), but no significant effects can be found (Steps b & c).

Rationale – effect per *SndTrans*: Values do not change with changing *Tasks* (Step g), and no significant differences are found (Steps h & i).

**3.2 Hypothesis H2** (Technical System Capability has an impact on Measured Involvement) is not supported, as *SndTrans* has no significant impact on *Influence of Content*.

Rationale – main effect: Values slightly decrease with increasing *Tasks*, i.e. increasing impairment (Step a), but no significant differences are found (Steps b & c).

Rationale – effect per *Task*: Values do not change with changing *Tasks* (Step d), and no significant differences are found (Steps e & f).

Figure D.3: Results for the measure *Influence of Content*.

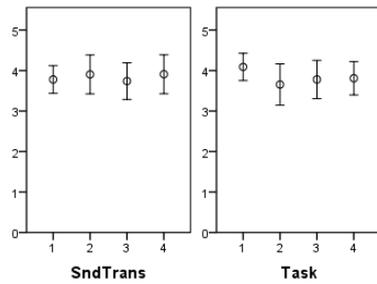
**1. Analysis Steps**

The present figure shows the data analysis for the measure *Rating Difficulty* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of *Task* if all data over *SndTrans* is compiled and the effect of *SndTrans* if all data over *Task* is compiled. Steps (d) to (f) concern the individual effect of *Task* analyzed per *SndTrans*. Steps (g) to (i) concern the individual effect of *SndTrans* analyzed per *Task*.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per *Task* and *SndTrans*. Steps (b), (e), & (h): results of one-way ANOVAs with *Task* ∈ [1...4] and *SndTrans* ∈ [1 = 0%, 2 = 5%, 3 = 10%, 4 = 15% packet loss rate] as input, expressed by the *F*-measure and significance level *p*. Steps (c), (f), & (i): results of PostHoc tests with Sidak correction for pairwise comparisons between all pairs of *Task* or *SndTrans*, expressed by the significance level *p*. Significant differences ( $p \leq 0.05$ ) are indicated with an asterisk (\*).

**2.1 Main Effects**

(a) Errorbars



(b) ANOVA

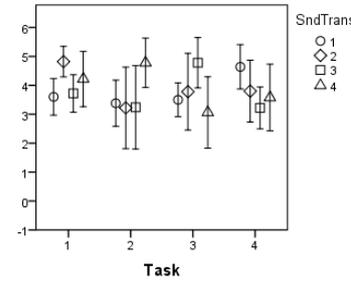
	<i>F</i>	<i>p</i>
<i>Task</i>	1.078	0.365
<i>SndTrans</i>	0.241	0.868
Interaction	4.646	0.000*

(c) PostHoc tests

<i>SndTrans</i>	<i>p</i>	<i>Task</i>	<i>p</i>
1 vs. 2	0.997	1 vs. 2	0.419
1 vs. 3	1.000	1 vs. 3	0.774
1 vs. 4	0.996	1 vs. 4	0.845
2 vs. 3	0.986	2 vs. 3	0.997
2 vs. 4	1.000	2 vs. 4	0.990
3 vs. 4	0.984	3 vs. 4	1.000

**2.2 Individual Effect of *SndTrans* per *Task***

(d) Errorbars



(e) ANOVA per *Task*

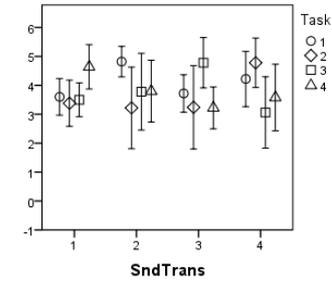
<i>Task</i>	<i>F</i>	<i>p</i>
1	4.731	0.015*
2	3.221	0.051
3	3.746	0.033*
4	3.123	0.055

(f) PostHoc tests per *Task*

<i>Task</i>	<i>SndTrans</i>	<i>p</i>	
1	1 vs. 2	0.023*	
	1 vs. 3	1.000	
	1 vs. 4	0.488	
	2 vs. 3	0.046*	
	2 vs. 4	0.524	
	3 vs. 4	0.708	
	2	1 vs. 2	1.000
		1 vs. 3	1.000
		1 vs. 4	0.174
		2 vs. 3	1.000
		2 vs. 4	0.105
		3 vs. 4	0.112
3		1 vs. 2	0.996
		1 vs. 3	0.161
		1 vs. 4	0.962
		2 vs. 3	0.390
		2 vs. 4	0.729
		3 vs. 4	0.031*
	4	1 vs. 2	0.471
		1 vs. 3	0.056
		1 vs. 4	0.232
		2 vs. 3	0.817
		2 vs. 4	0.998
		3 vs. 4	0.977

**2.3 Individual Effect of *Task* per *SndTrans***

(g) Errorbars



(h) ANOVA per *SndTrans*

<i>SndTrans</i>	<i>F</i>	<i>p</i>
1	5.283	0.010*
2	2.648	0.084
3	4.353	0.020*
4	3.854	0.030*

(i) PostHoc tests per *SndTrans*

<i>SndTrans</i>	<i>Task</i>	<i>p</i>	
1	1 vs. 2	0.991	
	1 vs. 3	1.000	
	1 vs. 4	0.059	
	2 vs. 3	1.000	
	2 vs. 4	0.017*	
	3 vs. 4	0.034*	
	2	1 vs. 2	0.080
		1 vs. 3	0.438
		1 vs. 4	0.459
		2 vs. 3	0.923
		2 vs. 4	0.911
		3 vs. 4	1.000
3		1 vs. 2	0.922
		1 vs. 3	0.256
		1 vs. 4	0.908
		2 vs. 3	0.040*
		2 vs. 4	1.000
		3 vs. 4	0.037*
	4	1 vs. 2	0.897
		1 vs. 3	0.252
		1 vs. 4	0.826
		2 vs. 3	0.034*
		2 vs. 4	0.222
		3 vs. 4	0.924

**3. Analysis Results:**

**3.1 Hypothesis H1** (Involvement has an impact on Measured Involvement and/or Cognitive Load) is not supported, as *Task* has no significant impact on *Rating Difficulty*.

Rationale – main effect: Values do not change with changing *Tasks* (Step a), and no significant differences are found (Steps b & c).

Rationale – effect per *SndTrans*: No consistent trend is visible (Step g), while significant effects can be found (Step h), which, however, are few (i.e. 5 of 18 cases) and not consistent across *Tasks* (Step f).

**3.2 Hypothesis H2** (Technical System Capability has an impact on Measured Involvement and/or Cognitive Load) is not supported, as *SndTrans* has no significant impact on *Rating Difficulty*.

Rationale – main effect: Values do not change with changing *Tasks* (Step a), and no significant differences are found (Steps b & c).

Rationale – effect per *Task*: No consistent trend is visible (Step g), while significant effects can be found for the listening-only tasks (Step e), which, however, are very few (i.e. 3 of 18 cases) and not consistent across *Tasks* (Step f).

Figure D.4: Results for the measure *Rating Difficulty*.

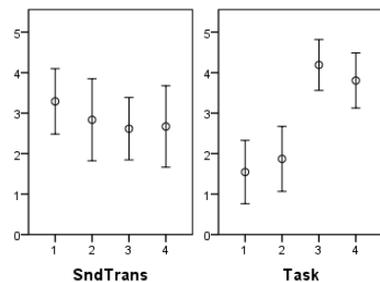
1. Analysis Steps

The present figure shows the data analysis for the measure *Quality* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of *Task* if all data over *SndTrans* is compiled and the effect of *SndTrans* if all data over *Task* is compiled. Steps (d) to (f) concern the individual effect of *Task* analyzed per *SndTrans*. Steps (g) to (i) concern the individual effect of *SndTrans* analyzed per *Task*.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per *Task* and *SndTrans*. Steps (b), (e), & (h): results of one-way ANOVAs with *Task* ∈ [1..4] and *SndTrans* ∈ [1 = 0%, 2 = 5%, 3 = 10%, 4 = 15% packet loss rate] as input, expressed by the *F*-measure and significance level *p*. Steps (c), (f), & (i): results of PostHoc tests with Sidak correction for pairwise comparisons between all pairs of *Task* or *SndTrans*, expressed by the significance level *p*. Significant differences ( $p \leq 0.05$ ) are indicated with an asterisk (\*).

2.1 Main Effects

(a) Errorbars



(b) ANOVA

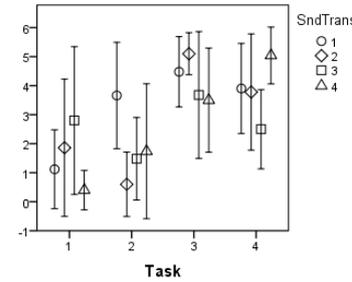
	<i>F</i>	<i>p</i>
<i>Task</i>	12.047	0.000*
<i>SndTrans</i>	3.211	0.029*
Interaction	1.528	0.157

(c) PostHoc tests

<i>SndTrans</i>	<i>p</i>	<i>Task</i>	<i>p</i>
1 vs. 2	0.998	1 vs. 2	0.859
1 vs. 3	0.458	1 vs. 3	0.000*
1 vs. 4	0.039*	1 vs. 4	0.000*
2 vs. 3	0.776	2 vs. 3	0.004*
2 vs. 4	0.121	2 vs. 4	0.003*
3 vs. 4	0.844	3 vs. 4	1.000

2.2 Individual Effect of *SndTrans* per *Task*

(d) Errorbars



(e) ANOVA per *Task*

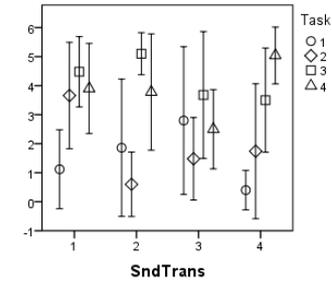
<i>Task</i>	<i>F</i>	<i>p</i>
1	1.492	0.255
2	3.364	0.045*
3	2.500	0.097
4	0.634	0.604

(f) PostHoc tests per *Task*

<i>Task</i>	<i>SndTrans</i>	<i>p</i>
1	1 vs. 2	0.405
	1 vs. 3	0.932
	1 vs. 4	0.455
	2 vs. 3	0.941
	2 vs. 4	1.000
	3 vs. 4	0.962
2	1 vs. 2	0.866
	1 vs. 3	0.214
	1 vs. 4	0.055
	2 vs. 3	0.851
	2 vs. 4	0.408
	3 vs. 4	0.982
3	1 vs. 2	0.982
	1 vs. 3	0.935
	1 vs. 4	0.397
	2 vs. 3	0.538
	2 vs. 4	0.117
	3 vs. 4	0.935
4	1 vs. 2	0.728
	1 vs. 3	0.987
	1 vs. 4	0.946
	2 vs. 3	0.984
	2 vs. 4	0.998
	3 vs. 4	1.000

2.3 Individual Effect of *Task* per *SndTrans*

(g) Errorbars



(h) ANOVA per *SndTrans*

<i>SndTrans</i>	<i>F</i>	<i>p</i>
1	0.874	0.475
2	9.877	0.001*
3	4.314	0.021*
4	3.589	0.037*

(i) PostHoc tests per *SndTrans*

<i>SndTrans</i>	<i>Task</i>	<i>p</i>
1	1 vs. 2	0.974
	1 vs. 3	0.590
	1 vs. 4	0.999
	2 vs. 3	0.969
2	2 vs. 3	0.969
	2 vs. 4	1.000
	3 vs. 4	0.851
	1 vs. 2	0.356
	1 vs. 3	0.001*
	1 vs. 4	0.004*
3	2 vs. 3	0.057
	2 vs. 4	0.209
	3 vs. 4	0.986
	1 vs. 2	1.000
4	1 vs. 3	0.271
	1 vs. 4	0.119
	2 vs. 3	0.145
	2 vs. 4	0.059
4	3 vs. 4	0.998
	1 vs. 2	1.000
	1 vs. 3	0.689
	1 vs. 4	0.082
4	2 vs. 3	0.640
	2 vs. 4	0.071
4	3 vs. 4	0.737

3. Analysis Results:

3.1 Hypothesis H5b (Involvement has an impact on Speech Communication Quality and/or Quality of Experience) is supported, as *Task* has a significant impact on *Quality*, noting that the difference between conversation and listening-only tasks explains this effect.

Rationale – main effect: Values for the conversation tasks T3 & T4 are higher than the values for the listening-only tasks T1 & T2 (Step a). This effect is significant (Step b) and the PostHoc tests confirm the differences between conversation and listening-only tasks (Step c).

Rationale – effect per *SndTrans*: Values increase with increasing task number in three of the four *SndTrans* (Step g). These effects are significant for three of the four *SndTrans* (Step h), whereas differences are strong enough to be detected by the PostHoc tests in only two cases.

3.2 Hypothesis H6b (Technical System Capability has an impact on Speech Communication Quality and/or Quality of Experience) is supported, as *SndTrans* has a significant impact on *Quality*, noting that this support can only be found in the main effects between the most extreme cases.

Rationale – main effect: Values decrease with increasing *SndTrans*, i.e. increasing packet loss rate (Step a). This effect is significant (Step b) and is explained by a significant difference between the two most extreme *SndTrans*.

Rationale – effect per *Task*: No consistent trend is visible (Step g), while significant effects can be found for one task (Step e), which, however, is not strong enough to be detected by the PostHoc tests (Step f).

Figure D.5: Results for the measure *Quality*.

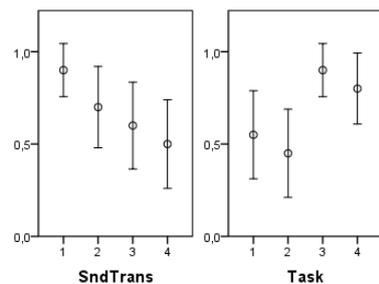
1. Analysis Steps

The present figure shows the data analysis for the measure *Own Conversation Effort* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of *Task* if all data over *SndTrans* is compiled and the effect of *SndTrans* if all data over *Task* is compiled. Steps (d) to (f) concern the individual effect of *Task* analyzed per *SndTrans*. Steps (g) to (i) concern the individual effect of *SndTrans* analyzed per *Task*.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per *Task* and *SndTrans*. Steps (b), (e), & (h): results of one-way ANOVAs with  $Task \in [1..4]$  and  $SndTrans \in [1 = 0\%, 2 = 5\%, 3 = 10\%, 4 = 15\% \text{ packet loss rate}]$  as input, expressed by the *F*-measure and significance level *p*. Steps (c), (f), & (i): results of PostHoc tests with Sidak correction for pairwise comparisons between all pairs of *Task* or *SndTrans*, expressed by the significance level *p*. Significant differences ( $p \leq 0.05$ ) are indicated with an asterisk (\*).

2.1 Main Effects

(a) Errorbars



(b) ANOVA

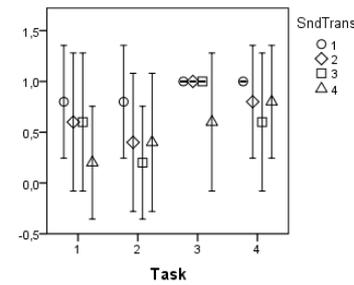
	<i>F</i>	<i>p</i>
<i>Task</i>	4.711	0.005*
<i>SndTrans</i>	3.111	0.032*
Interaction	0.681	0.723

(c) PostHoc tests

<i>SndTrans</i>	<i>p</i>	<i>Task</i>	<i>p</i>
1 vs. 2	0.620	1 vs. 2	0.977
1 vs. 3	0.178	1 vs. 4	0.075
1 vs. 4	0.029*	1 vs. 3	0.364
2 vs. 3	0.977	2 vs. 3	0.010*
2 vs. 4	0.620	2 vs. 4	0.075
3 vs. 4	0.977	3 vs. 4	0.977

2.2 Individual Effect of *SndTrans* per *Task*

(d) Errorbars



(e) ANOVA per *Task*

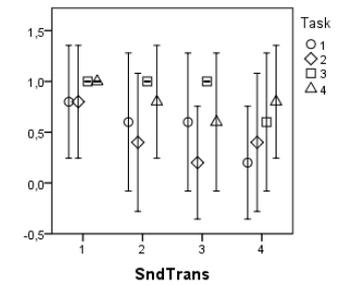
<i>Task</i>	<i>F</i>	<i>p</i>
1	1.267	0.319
2	1.267	0.319
3	2.667	0.083
4	0.762	0.532

(f) PostHoc tests per *Task*

<i>Task</i>	<i>SndTrans</i>	<i>p</i>
1	1 vs. 2	0.990
	1 vs. 3	0.990
	1 vs. 4	0.378
	2 vs. 3	1.000
	2 vs. 4	0.782
	3 vs. 4	0.782
2	1 vs. 2	0.782
	1 vs. 3	0.378
	1 vs. 4	0.782
	2 vs. 3	0.990
	2 vs. 4	1.000
	3 vs. 4	0.990
3	1 vs. 2	1.000
	1 vs. 3	1.000
	1 vs. 4	0.190
	2 vs. 3	1.000
	2 vs. 4	0.190
	3 vs. 4	0.190
4	1 vs. 2	0.975
	1 vs. 3	0.623
	1 vs. 4	0.975
	2 vs. 3	0.975
	2 vs. 4	1.000
	3 vs. 4	0.975

2.3 Individual Effect of *Task* per *SndTrans*

(g) Errorbars



(h) ANOVA per *SndTrans*

<i>SndTrans</i>	<i>F</i>	<i>p</i>
1	0.667	0.585
2	1.667	0.214
3	2.667	0.083
4	1.333	0.299

(i) PostHoc tests per *SndTrans*

<i>SndTrans</i>	<i>Task</i>	<i>p</i>
1	1 vs. 2	1.000
	1 vs. 3	0.911
	1 vs. 4	0.911
	2 vs. 3	0.911
2	2 vs. 4	0.911
	3 vs. 4	1.000
	1 vs. 2	0.982
	1 vs. 3	0.688
3	1 vs. 4	0.982
	2 vs. 3	0.264
	2 vs. 4	0.688
	3 vs. 4	0.982
4	1 vs. 2	0.688
	1 vs. 3	0.688
	1 vs. 4	1.000
	2 vs. 3	0.070
4	2 vs. 4	0.688
	3 vs. 4	0.688
	1 vs. 2	0.990
	1 vs. 3	0.782
4	1 vs. 4	0.378
	2 vs. 3	0.990
	2 vs. 4	0.782
	3 vs. 4	0.990

3. Analysis Results:

3.1 Hypothesis H5b (Involvement has an impact on Speech Communication Quality with focus on Group-Communication Component) is supported, as *Task* has a significant impact on *Own Conversation Effort*, noting that the difference between conversation and listening-only tasks explains this effect.

Rationale – main effect: Values for the conversation tasks T3 & T4 are higher than the values for the listening-only tasks T1 & T2 (Step a). This effect is significant (Step b) and the PostHoc tests confirm the differences between conversation and listening-only tasks (Step c).

Rationale – effect per *SndTrans*: No consistent trend is visible (Step g), and no significant effects are found (Steps h & i).

3.2 Hypothesis H6b (Technical System Capability has an impact on Speech Communication Quality with focus on Group-Communication Component) is supported, as *SndTrans* has a significant impact on *Own Conversation Effort*, noting that this support can only be found in the main effects between the most extreme cases.

Rationale – main effect: Values decrease with increasing *SndTrans*, i.e. increasing packet loss rate (Step a). This effect is significant (Step b) and is explained by a significant difference between the two most extreme *SndTrans*.

Rationale – effect per *Task*: No consistent trend is visible (Step d), and no significant effects are found (Steps e & f).

Figure D.6: Results for the measure *Own Conversation Effort*.

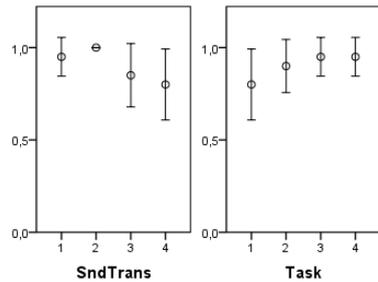
**1. Analysis Steps**

The present figure shows the data analysis for the measure *Partners Conversation Effort* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of *Task* if all data over *SndTrans* is compiled and the effect of *SndTrans* if all data over *Task* is compiled. Steps (d) to (f) concern the individual effect of *Task* analyzed per *SndTrans*. Steps (g) to (i) concern the individual effect of *SndTrans* analyzed per *Task*.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per *Task* and *SndTrans*. Steps (b), (e), & (h): results of one-way ANOVAs with *Task* ∈ [1...4] and *SndTrans* ∈ [1 = 0%, 2 = 5%, 3 = 10%, 4 = 15% packet loss rate] as input, expressed by the *F*-measure and significance level *p*. Steps (c), (f), & (i): results of PostHoc tests with Sidak correction for pairwise comparisons between all pairs of *Task* or *SndTrans*, expressed by the significance level *p*. Significant differences ( $p \leq 0.05$ ) are indicated with an asterisk (\*).

**2.1 Main Effects**

(a) Errorbars



(b) ANOVA

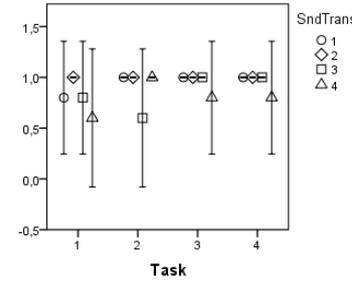
	<i>F</i>	<i>p</i>
<i>Task</i>	1.143	0.339
<i>SndTrans</i>	1.905	0.138
Interaction	1.016	0.437

(c) PostHoc tests

<i>SndTrans</i>	<i>p</i>	<i>Task</i>	<i>p</i>
1 vs. 2	0.996	1 vs. 2	0.871
1 vs. 3	0.871	1 vs. 3	0.515
1 vs. 4	0.515	1 vs. 4	0.515
2 vs. 3	0.515	2 vs. 3	0.996
2 vs. 4	0.199	2 vs. 4	0.996
3 vs. 4	0.996	3 vs. 4	1.000

**2.2 Individual Effect of *SndTrans* per *Task***

(d) Errorbars



(e) ANOVA per *Task*

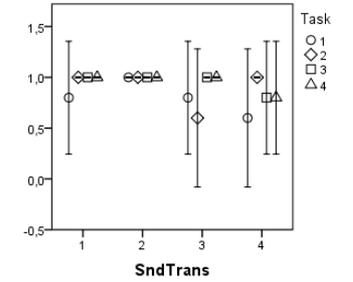
<i>Task</i>	<i>F</i>	<i>p</i>
1	0.762	0.532
2	2.667	0.083
3	1.000	0.418
4	1.000	0.418

(f) PostHoc tests per *Task*

<i>Task</i>	<i>SndTrans</i>	<i>p</i>
1	1 vs. 2	0.975
	1 vs. 3	1.000
	1 vs. 4	0.975
	2 vs. 3	0.975
	2 vs. 4	0.623
	3 vs. 4	0.975
2	1 vs. 2	1.000
	1 vs. 3	0.190
	1 vs. 4	1.000
	2 vs. 3	0.190
	2 vs. 4	1.000
	3 vs. 4	0.190
3	1 vs. 2	1.000
	1 vs. 3	1.000
	1 vs. 4	0.688
	2 vs. 3	1.000
	2 vs. 4	0.688
	3 vs. 4	0.688
4	1 vs. 2	1.000
	1 vs. 3	1.000
	1 vs. 4	0.688
	2 vs. 3	1.000
	2 vs. 4	0.688
	3 vs. 4	0.688

**2.3 Individual Effect of *Task* per *SndTrans***

(g) Errorbars



(h) ANOVA per *SndTrans*

<i>SndTrans</i>	<i>F</i>	<i>p</i>
1	1.000	0.418
2	-	-
3	1.467	0.261
4	0.762	0.532

(i) PostHoc tests per *SndTrans*

<i>SndTrans</i>	<i>Task</i>	<i>p</i>
1	1 vs. 2	0.688
	1 vs. 3	0.688
	1 vs. 4	0.688
	2 vs. 3	1.000
2	2 vs. 4	1.000
	3 vs. 4	1.000
	1 vs. 2	-
	1 vs. 3	-
3	2 vs. 3	-
	2 vs. 4	-
	3 vs. 4	-
	1 vs. 2	0.946
4	1 vs. 3	0.946
	1 vs. 4	0.946
	2 vs. 3	0.442
	2 vs. 4	0.442
4	3 vs. 4	1.000
	1 vs. 2	0.623
	1 vs. 3	0.975
	1 vs. 4	0.975
4	2 vs. 3	0.975
	2 vs. 4	0.975
	3 vs. 4	1.000

**3. Analysis Results:**

**3.1 Hypothesis H5b** (Involvement has an impact on Speech Communication Quality with focus on Group-Communication Component) is not supported, as *Task* has no significant impact on *Partners Conversation Effort*.

Rationale – main effect: Values slightly increase with increasing *Task* number (Step a), but no significant effects can be found (Steps b & c).

Rationale – effect per *SndTrans*: Values are in many cases equal across *Tasks* (Step g), and no significant effects are found (Steps h & i).

**3.2 Hypothesis H6b** (Technical System Capability has an impact on Speech Communication Quality with focus on Group-Communication Component) is not supported, as *SndTrans* has no significant impact on *Partners Conversation Effort*.

Rationale – main effect: Values slightly decrease with increasing *SndTrans*, i.e. increasing packet loss rate (Step a), but no significant effects can be found (Steps b & c).

Rationale – effect per *Task*: Values are in many cases equal across *SndTrans* (Step d), and no significant effects are found (Steps e & f).

Figure D.7: Results for the measure *Partners Conversation Effort*.

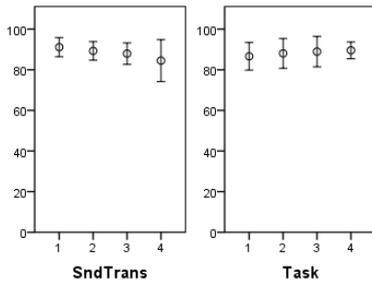
**1. Analysis Steps**

The present figure shows the data analysis for the measure *Partners Conversation Effort Confidence* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of *Task* if all data over *SndTrans* is compiled and the effect of *SndTrans* if all data over *Task* is compiled. Steps (d) to (f) concern the individual effect of *Task* analyzed per *SndTrans*. Steps (g) to (i) concern the individual effect of *SndTrans* analyzed per *Task*.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per *Task* and *SndTrans*. Steps (b), (e), & (h): results of one-way ANOVAs with *Task* ∈ [1..4] and *SndTrans* ∈ [1 = 0%, 2 = 5%, 3 = 10%, 4 = 15% packet loss rate] as input, expressed by the *F*-measure and significance level *p*. Steps (c), (f), & (i): results of PostHoc tests with Sidak correction for pairwise comparisons between all pairs of *Task* or *SndTrans*, expressed by the significance level *p*. Significant differences ( $p \leq 0.05$ ) are indicated with an asterisk (\*).

**2.1 Main Effects**

(a) Errorbars



(b) ANOVA

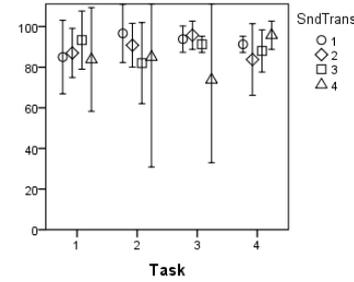
	<i>F</i>	<i>p</i>
<i>Task</i>	0.104	0.957
<i>SndTrans</i>	0.875	0.460
Interaction	1.290	0.265

(c) PostHoc tests

<i>SndTrans</i>	<i>p</i>	<i>Task</i>	<i>p</i>
1 vs. 2	0.999	1 vs. 2	1.000
1 vs. 3	0.976	1 vs. 3	0.995
1 vs. 4	0.586	1 vs. 4	0.982
2 vs. 3	1.000	2 vs. 3	1.000
2 vs. 4	0.853	2 vs. 4	1.000
3 vs. 4	0.968	3 vs. 4	1.000

**2.2 Individual Effect of *SndTrans* per *Task***

(d) Errorbars



(e) ANOVA per *Task*

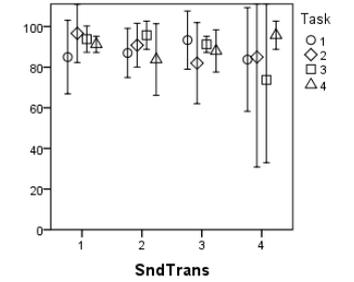
<i>Task</i>	<i>F</i>	<i>p</i>
1	0.295	0.828
2	0.803	0.516
3	2.512	0.104
4	1.866	0.185

(f) PostHoc tests per *Task*

<i>Task</i>	<i>SndTrans</i>	<i>p</i>
1	1 vs. 2	1.000
	1 vs. 3	0.972
	1 vs. 4	1.000
	2 vs. 3	0.993
2	2 vs. 4	1.000
	3 vs. 4	0.947
	1 vs. 2	0.994
	1 vs. 3	0.689
3	1 vs. 4	0.907
	2 vs. 3	0.917
	2 vs. 4	0.995
	3 vs. 4	1.000
4	1 vs. 2	1.000
	1 vs. 3	1.000
	1 vs. 4	0.203
	2 vs. 3	0.997
	2 vs. 4	0.172
	3 vs. 4	0.380
	1 vs. 2	0.693
	1 vs. 3	0.989
	1 vs. 4	0.958
	2 vs. 3	0.958
	2 vs. 4	0.220
	3 vs. 4	0.611

**2.3 Individual Effect of *Task* per *SndTrans***

(g) Errorbars



(h) ANOVA per *SndTrans*

<i>SndTrans</i>	<i>F</i>	<i>p</i>
1	1.309	0.313
2	1.378	0.290
3	0.955	0.443
4	0.833	0.501

(i) PostHoc tests per *SndTrans*

<i>SndTrans</i>	<i>Task</i>	<i>p</i>
1	1 vs. 2	0.462
	1 vs. 3	0.608
	1 vs. 4	0.899
	2 vs. 3	0.999
2	2 vs. 4	0.970
	3 vs. 4	0.999
	1 vs. 2	0.986
	1 vs. 3	0.660
3	1 vs. 4	0.996
	2 vs. 3	0.962
	2 vs. 4	0.832
	3 vs. 4	0.382
4	1 vs. 2	0.644
	1 vs. 3	1.000
	1 vs. 4	0.983
	2 vs. 3	0.750
	2 vs. 4	0.941
	3 vs. 4	0.998
	1 vs. 2	1.000
	1 vs. 3	0.976
	1 vs. 4	0.945
	2 vs. 3	0.978
	2 vs. 4	0.982
	3 vs. 4	0.597

**3. Analysis Results:**

**3.1 Hypothesis H5b** (Involvement has an impact on Speech Communication Quality with focus on Group-Communication Component) is not supported, as *Task* has no significant impact on *Partners Conversation Effort Confidence*.

Rationale – main effect: Values increase very slightly with increasing *Task* number (Step a), but no significant effects can be found (Steps b & c).

Rationale – effect per *SndTrans*: No consistent trend is visible (Step g), and no significant effects are found (Steps h & i).

**3.2 Hypothesis H6b** (Technical System Capability has an impact on Speech Communication Quality with focus on Group-Communication Component) is not supported, as *SndTrans* has no significant impact on *Partners Conversation Effort Confidence*.

Rationale – main effect: Values decrease very slightly with increasing *SndTrans*, i.e. increasing packet loss rate (Step a), but no significant effects can be found (Steps b & c).

Rationale – effect per *Task*: No consistent trend is visible (Step d), and no significant effects are found (Steps e & f).

Figure D.8: Results for the measure *Partners Conversation Effort Confidence*.

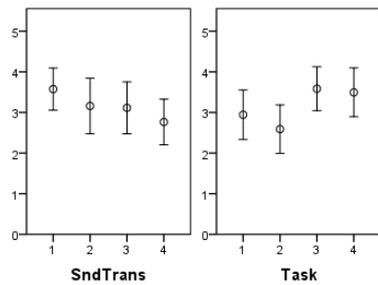
1. Analysis Steps

The present figure shows the data analysis for the measure *Perceived Effort* in nine steps (a) to (i). Steps (a) to (c) concern the main effects; that means the effect of *Task* if all data over *SndTrans* is compiled and the effect of *SndTrans* if all data over *Task* is compiled. Steps (d) to (f) concern the individual effect of *Task* analyzed per *SndTrans*. Steps (g) to (i) concern the individual effect of *SndTrans* analyzed per *Task*.

Steps (a), (d), & (g): errorbar plots, showing the mean values and 95% confidence intervals per *Task* and *SndTrans*. Steps (b), (e), & (h): results of one-way ANOVAs with  $Task \in [1..4]$  and  $SndTrans \in [1 = 0\%, 2 = 5\%, 3 = 10\%, 4 = 15\% \text{ packet loss rate}]$  as input, expressed by the *F*-measure and significance level *p*. Steps (c), (f), & (i): results of PostHoc tests with Sidak correction for pairwise comparisons between all pairs of *Task* or *SndTrans*, expressed by the significance level *p*. Significant differences ( $p \leq 0.05$ ) are indicated with an asterisk (\*).

2.1 Main Effects

(a) Errorbars



(b) ANOVA

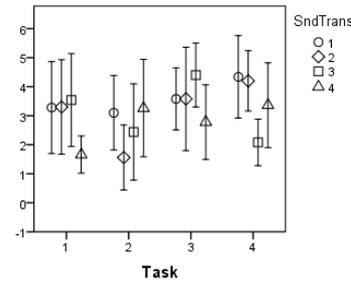
	<i>F</i>	<i>p</i>
<i>Task</i>	3.679	0.016*
<i>SndTrans</i>	1.830	0.151
Interaction	3.353	0.002*

(c) PostHoc tests

<i>SndTrans</i>	<i>p</i>	<i>Task</i>	<i>p</i>
1 vs. 2	0.801	1 vs. 2	0.892
1 vs. 3	0.716	1 vs. 4	0.351
1 vs. 4	0.129	1 vs. 3	0.528
2 vs. 3	1.000	2 vs. 3	0.033*
2 vs. 4	0.834	2 vs. 4	0.066
3 vs. 4	0.898	3 vs. 4	1.000

2.2 Individual Effect of *SndTrans* per *Task*

(d) Errorbars



(e) ANOVA per *Task*

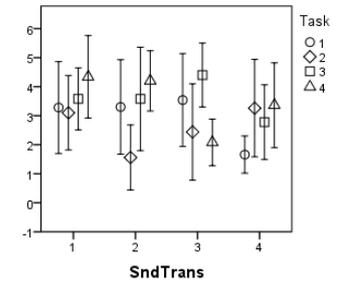
<i>Task</i>	<i>F</i>	<i>p</i>
1	2.834	0.071
2	2.173	0.131
3	1.879	0.174
4	5.650	0.008*

(f) PostHoc tests per *Task*

<i>Task</i>	<i>SndTrans</i>	<i>p</i>
1	1 vs. 2	1.000
	1 vs. 3	1.000
	1 vs. 4	0.219
	2 vs. 3	1.000
	2 vs. 4	0.209
	3 vs. 4	0.113
2	1 vs. 2	0.285
	1 vs. 3	0.947
	1 vs. 4	1.000
	2 vs. 3	0.826
	2 vs. 4	0.196
	3 vs. 4	0.867
3	1 vs. 2	1.000
	1 vs. 3	0.818
	1 vs. 4	0.833
	2 vs. 3	0.818
	2 vs. 4	0.833
	3 vs. 4	0.169
4	1 vs. 2	1.000
	1 vs. 3	0.013*
	1 vs. 4	0.573
	2 vs. 3	0.020*
	2 vs. 4	0.723
	3 vs. 4	0.286

2.3 Individual Effect of *Task* per *SndTrans*

(g) Errorbars



(h) ANOVA per *SndTrans*

<i>SndTrans</i>	<i>F</i>	<i>p</i>
1	1.261	0.321
2	4.832	0.014*
3	4.817	0.014*
4	2.664	0.083

(i) PostHoc tests per *SndTrans*

<i>SndTrans</i>	<i>Task</i>	<i>p</i>
1	1 vs. 2	1.000
	1 vs. 3	0.999
	1 vs. 4	0.605
	2 vs. 3	0.984
2	2 vs. 4	0.435
	3 vs. 4	0.868
	1 vs. 2	0.164
	1 vs. 3	0.999
3	1 vs. 4	0.798
	2 vs. 3	0.078
	2 vs. 4	0.013*
	3 vs. 4	0.956
4	1 vs. 2	0.555
	1 vs. 3	0.784
	1 vs. 4	0.255
	2 vs. 3	0.064
	2 vs. 4	0.996
	3 vs. 4	0.022*
	1 vs. 2	0.170
	1 vs. 3	0.524
	1 vs. 4	0.129
	2 vs. 3	0.982
	2 vs. 4	1.000
	3 vs. 4	0.955

3. Analysis Results:

3.1 Hypothesis H3 (Involvement has an impact on Cognitive Load) is supported, as *Task* has a significant impact on *Quality*, noting that the difference between conversation and listening-only tasks explains this effect.

Rationale – main effect: Values for the conversation tasks T3 & T4 are higher than the values for the listening-only tasks T1 & T2 (Step a). This effect is significant (Step b), but the PostHoc tests confirm the differences only between tasks T2 and T3 (Step c).

Rationale – effect per *SndTrans*: No consistent trend is visible *SndTrans* (Step g). Significant are found for two of the four *SndTrans* (Step h), whereas differences are strong enough to be detected by the PostHoc tests in only two cases (Step i).

3.2 Hypothesis H4 (Technical System Capability has an impact on Cognitive Load) is not supported, as *SndTrans* has no significant impact on *Partners Conversation Effort Confidence*.

Rationale – main effect: Values slightly decrease with increasing *SndTrans*, i.e. increasing packet loss rate (Step a), but no significant effects can be found (Steps b & c).

Rationale – effect per *Task*: No consistent trend is visible (Step d), and no significant effects are found except in two cases for *SndTrans* = 4 (Steps e & f).

Figure D.9: Results for the measure *Perceived Effort*.

## *E*

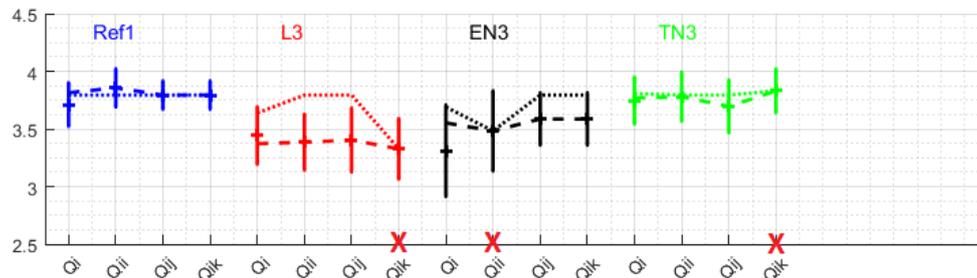
# *Detailed Results: Study on Impact of Audio-Only Telemeeting - Conversation Tests*

### *What this chapter is about*

The present appendix presents the detailed results of the three audio-only conversation tests ACT<sub>1</sub>, ACT<sub>2</sub> & ACT<sub>3</sub> on the impact of technical conditions on quality perception. First, each figure shows a visualization of the obtained data and the expectations according to Section 8.3. Then, each figure discusses per technical condition, in how far the expectations are fulfilled and how the current results relate to the different hypotheses formulated in Section 8.3. The individual results are then compiled in the main text in Section 8.3.

E.1 Experiment ACT<sub>1</sub>

## 1. Visualization of Results and Expectations

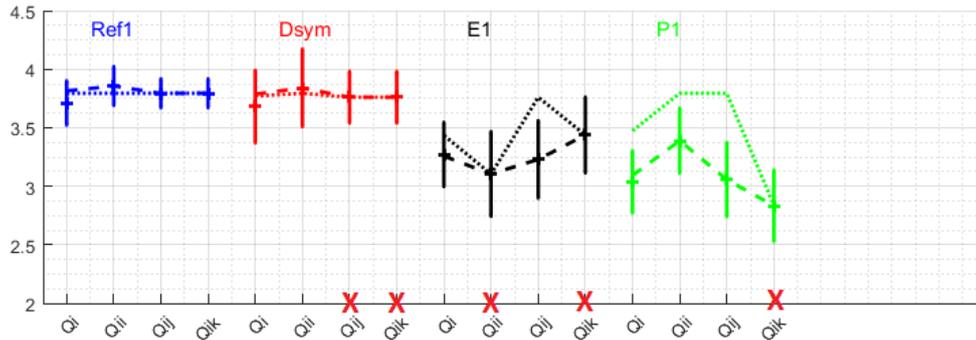


## 2. Detailed Analysis

Ref <sub>1</sub>	All expected values lie within confidence intervals of Q <sub>i</sub> , Q <sub>ii</sub> , & Q <sub>ij</sub> =Q <sub>ik</sub> , and all observed values are not significantly different from maximum value of whole test (3.8557, Q <sub>ii</sub> of Ref <sub>1</sub> ). ⇒ Ref <sub>1</sub> is confirmed as reference condition for this test.
L <sub>3</sub>	Expectations: No impairment on own connection, asymmetric condition concerning other connections. ⇒ Q <sub>ik</sub> determines lowest value, Q <sub>ii</sub> & Q <sub>ij</sub> should be equal to reference value Q <sub>ik</sub> of Ref <sub>1</sub> .  Testing Hypothesis H <sub>1</sub> (dotted line): Q <sub>ii</sub> & Q <sub>ij</sub> are significantly different (Q <sub>ii</sub> : p=0.001, Q <sub>ij</sub> : p=0.007) from the expected values. ⇒ Support for H <sub>1</sub> .  Testing Hypothesis H <sub>2</sub> (dashed line, since this is independent from H <sub>1</sub> ): Q <sub>i</sub> is not significantly different from the expected value (p=0.572). ⇒ Support for H <sub>2</sub> .
EN <sub>3</sub>	Expectations: Impairment on own connection, no impairments on other connections. ⇒ Q <sub>ii</sub> determines lowest value, Q <sub>ij</sub> = Q <sub>ik</sub> should be equal to reference value Q <sub>ik</sub> of Ref <sub>1</sub> .  Testing Hypothesis H <sub>1</sub> (dotted line): Q <sub>ij</sub> = Q <sub>ik</sub> are not significantly different from the expected value (p=0.071). ⇒ No support for H <sub>1</sub> .  Testing Hypothesis H <sub>2</sub> (dashed line, since this is independent from H <sub>1</sub> ): Q <sub>i</sub> is not significantly different from the expected value (p=0.221). ⇒ Support for H <sub>2</sub> .
TN <sub>3</sub>	Expectations: No impairment on own connection, asymmetric condition concerning other connections. ⇒ Q <sub>ik</sub> determines lowest value, Q <sub>ii</sub> & Q <sub>ij</sub> should be equal to reference value Q <sub>ik</sub> of Ref <sub>1</sub> .  Testing Hypothesis H <sub>1</sub> (dotted line): Q <sub>ii</sub> & Q <sub>ij</sub> are not significantly different (Q <sub>ii</sub> : p=0.894, Q <sub>ij</sub> : p=0.389) from the expected values. ⇒ No support for H <sub>1</sub> .  Testing Hypothesis H <sub>2</sub> (dashed line, since this is independent from H <sub>1</sub> ): Q <sub>i</sub> is not significantly different from the expected value (p=0.824). ⇒ Support for H <sub>2</sub> .  Special observation: All values are essentially equal to reference condition. ⇒ Impairment not strong enough to elicit effect in quality ratings.

Figure E.1: Detailed results for Experiment ACT<sub>1</sub>, Part 1.

1. Visualization of Results and Expectations



Note: The reference condition Ref1 is repeated in this plot for better visual comparison.

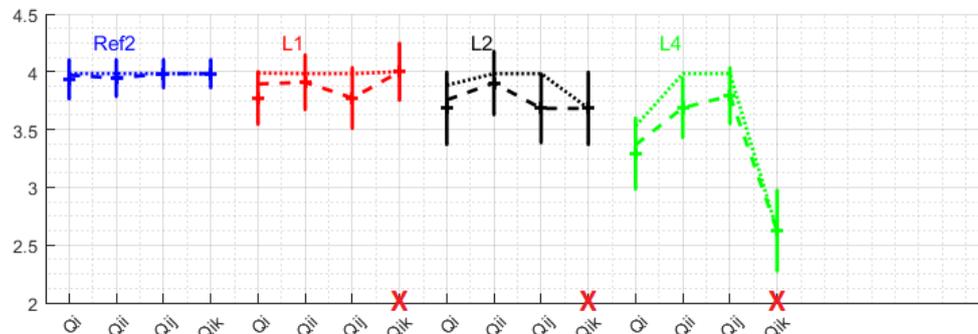
2. Detailed Analysis

- Dsym**      Expectations: No impairment on own connection, symmetric condition concerning other connections.  
 ⇒  $Q_{ij} = Q_{ik}$  determine lowest value,  $Q_{ii}$  should be equal to reference value ( $Q_{ik}$  of Ref1).  
 Testing Hypothesis H1 (dotted line):  $Q_{ii}$  is not significantly different from the expected value ( $p=0.775$ ).  
 ⇒ No support for H1.  
 Testing Hypothesis H2 (dashed line, since this is independent from H1):  $Q_i$  is not significantly different from the expected value ( $p=0.483$ ).  
 ⇒ Support for H2.  
 Special observation: All values are essentially equal to reference condition. ⇒ Impairment not strong enough to elicit effect in quality ratings.
- E1**      Expectations: Impairment on own connection, asymmetric condition concerning other connections.  
 ⇒  $Q_{ii}$  &  $Q_{ik}$  are lower than reference value  $Q_{ik}$  of Ref1,  $Q_{ij}$  should be equal to reference value ( $Q_{ik}$  of Ref1).  
 Testing Hypothesis H1 (dotted line):  $Q_{ij}$  is significantly different from the expected value ( $p=0.002$ ).  
 ⇒ Support for H1.  
 Testing Hypothesis H2 (dashed line, since this is independent from H1):  $Q_i$  is not significantly different from the expected value ( $p=0.919$ ).  
 ⇒ Support for H2.
- P1**      Expectations: No impairment on own connection, asymmetric condition concerning other connections.  
 ⇒  $Q_{ik}$  determines lowest value,  $Q_{ii}$  &  $Q_{ij}$  should be equal to reference value ( $Q_{ik}$  of Ref1).  
 Testing Hypothesis H1 (dotted line):  $Q_{ii}$  &  $Q_{ij}$  are significantly different ( $Q_{ii}$ :  $p=0.005$ ,  $Q_{ij}$ :  $p=0.000$ ) from expected values for dotted line.  
 ⇒ Support for H1.  
 Testing Hypothesis H2 (dashed line, since this is independent from H1):  $Q_i$  is not significantly different from the expected value ( $p=0.676$ ).  
 ⇒ Support for H2.

Figure E.2: Detailed results for Experiment ACT1, Part 2.

## E.2 Experiment ACT2

### 1. Visualization of Results and Expectations



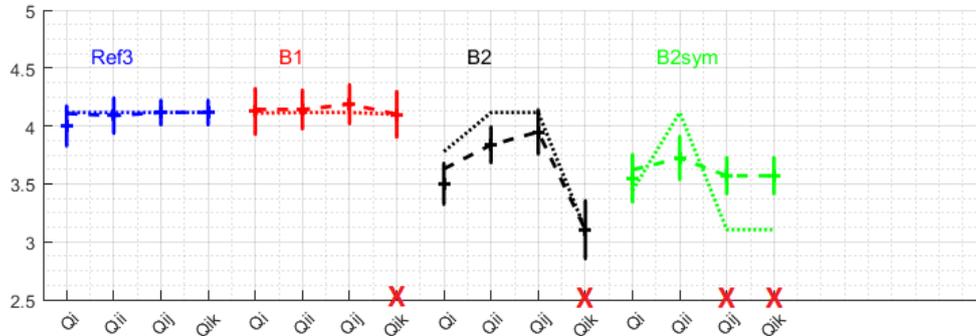
### 2. Detailed Analysis

Ref1	All expected values lie within confidence intervals of $Q_i$ , $Q_{ii}$ , & $Q_{ij}=Q_{ik}$ , and all observed values are not significantly different from maximum value of whole test (3.8557, $Q_{ii}$ of Ref1). ⇒ Ref1 is confirmed as reference condition for this test.
L1	Expectations: No impairment on own connection, asymmetric condition concerning other connections. ⇒ $Q_{ik}$ determines lowest value, $Q_{ii}$ & $Q_{ij}$ should be equal to reference value ( $Q_{ik}$ of Ref1).  Testing Hypothesis H1 (dotted line): $Q_{ii}$ & $Q_{ij}$ are not significantly different ( $Q_{ii}$ : $p=0.526$ , $Q_{ij}$ : $p=0.111$ ) from the expected values. ⇒ No support for H1.  Testing Hypothesis H2 (dashed line, since this is independent from H1): $Q_i$ is not significantly different from the expected value ( $p=0.285$ ). ⇒ Support for H2.
L2	Expectations: No impairment on own connection, asymmetric condition concerning other connections. ⇒ $Q_{ik}$ determines lowest value, $Q_{ii}$ & $Q_{ij}$ should be equal to reference value ( $Q_{ik}$ of Ref1).  Testing Hypothesis H1 (dotted line): $Q_{ii}$ is not significantly different ( $Q_{ii}$ : $p=0.550$ ) from the expected value, $Q_{ij}$ is significantly different ( $p=0.046$ ) from the expected value. ⇒ Support for H1.  Testing Hypothesis H2 (dashed line, since this is independent from H1): $Q_i$ is not significantly different from the expected value ( $p=0.637$ ). ⇒ Support for H2.
L4	Expectations: No impairment on own connection, asymmetric condition concerning other connections. ⇒ $Q_{ik}$ determines lowest value, $Q_{ii}$ & $Q_{ij}$ should be equal to reference value ( $Q_{ik}$ of Ref1).  Testing Hypothesis H1 (dotted line): $Q_{ii}$ is significantly different ( $Q_{ii}$ : $p=0.024$ ) from the expected value, $Q_{ij}$ is significantly different ( $p=0.113$ ) from the expected value. ⇒ No support for H1.  Testing Hypothesis H2 (dashed line, since this is independent from H1): $Q_i$ is not significantly different from the expected value ( $p=0.617$ ). ⇒ Support for H2.

Figure E.3: Detailed results for Experiment ACT2.

### E.3 Experiment ACT<sub>3</sub>

#### 1. Visualization of Results and Expectations

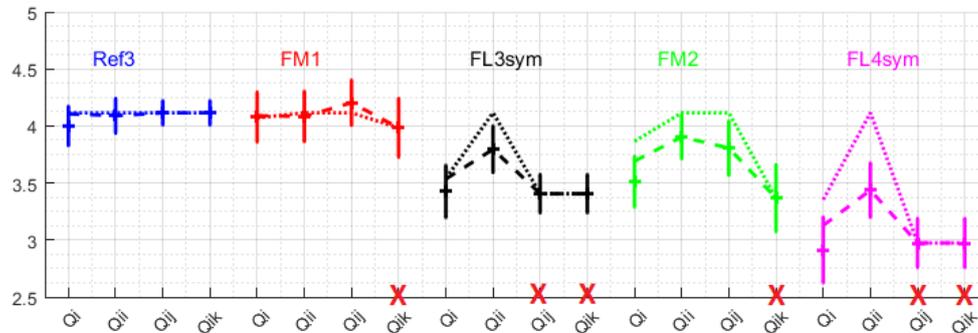


#### 2. Detailed Analysis

Ref <sub>3</sub>	<p>All expected values lie within confidence intervals of Qi, Qii, &amp; Qij=Qik, the observed values for Qii &amp; Qij=Qik are not significantly different from maximum value of whole test (4.2036, Qij of FM<sub>1</sub>), and the observed value for Qi is close but significantly different from that maximum value, but it is not significantly different from the maximum value for Qi (4.1250, Qi of B<sub>1</sub>).</p> <p>⇒ Ref<sub>3</sub> is confirmed as reference condition for this test.</p>
B <sub>1</sub>	<p>Expectations: No impairment on own connection, asymmetric condition concerning other connections.                      ⇒ Qik determines lowest value, Qii &amp; Qij should be equal to reference value (Qik of Ref<sub>1</sub>).</p> <p>Testing Hypothesis H<sub>1</sub> (dotted line): Qii &amp; Qij are not significantly different (Qii: p=0.753, Qij: p=0.388) from the expected values.                      ⇒ No support for H<sub>1</sub>.</p> <p>Testing Hypothesis H<sub>2</sub> (dashed line, since this is independent from H<sub>1</sub>): Qi is not significantly different from the expected value (p=0.854).                      ⇒ Support for H<sub>2</sub>.</p> <p>Special observation: All values are essentially equal to reference condition. ⇒ Impairment not strong enough to elicit effect in quality ratings.</p>
B <sub>2</sub>	<p>Expectations: No impairment on own connection, asymmetric condition concerning other connections.                      ⇒ Qik determines lowest value, Qii &amp; Qij should be equal to reference value (Qik of Ref<sub>1</sub>).</p> <p>Testing Hypothesis H<sub>1</sub> (dotted line): Qii is significantly different (Qii: p=0.001) from the expected value, Qij is not significantly different (p=0.081) from the expected value.                      ⇒ Support for H<sub>1</sub>.</p> <p>Testing Hypothesis H<sub>2</sub> (dashed line, since this is independent from H<sub>1</sub>): Qi is not significantly different from the expected value (p=0.141).                      ⇒ Support for H<sub>2</sub>.</p>
B <sub>2sym</sub>	<p>Expectations: No impairment on own connection, symmetric condition concerning other connections.                      ⇒ Qij = Qik should be equal to corresponding value of asymmetric condition (Qik of B<sub>2</sub>), Qii should be equal to reference value (Qik of Ref<sub>1</sub>).</p> <p>Testing Hypothesis H<sub>1</sub> (dotted line): Qii &amp; Qij are significantly different (Qii: p=0.000, Qij: p=0.000) from the expected values.                      ⇒ Support for H<sub>1</sub>.</p> <p>Testing Hypothesis H<sub>2</sub> (dashed line, since this is independent from H<sub>1</sub>): Qi is not significantly different from the expected value (p=0.480).                      ⇒ Support for H<sub>2</sub>.</p>

Figure E.4: Detailed results for Experiment ACT<sub>3</sub>, Part 1.

## 1. Visualization of Results and Expectations



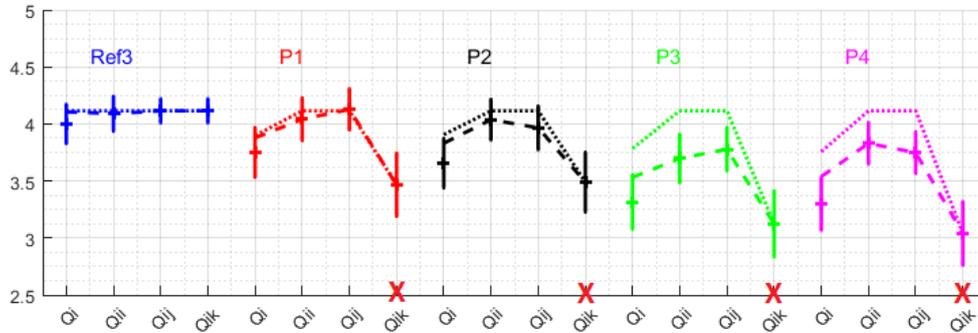
Note: The reference condition Ref3 is repeated in this plot for better visual comparison.

## 2. Detailed Analysis

FM1	<p>Expectations: No impairment on own connection, asymmetric condition concerning other connections.  <math>\Rightarrow</math> Qik determines lowest value, Qii &amp; Qij should be equal to reference value (Qik of Ref1).</p> <p>Testing Hypothesis H1 (dotted line): Qii &amp; Qij are not significantly different (Qii: <math>p=0.770</math>, Qij: <math>p=0.370</math>) from the expected values.  <math>\Rightarrow</math> No support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is not significantly different from the expected value (<math>p=0.909</math>).  <math>\Rightarrow</math> Support for H2.</p> <p>Special observation: All values are essentially equal to reference condition. <math>\Rightarrow</math> Impairment not strong enough to elicit effect in quality ratings.</p>
FL3sym	<p>Expectations: No impairment on own connection, symmetric condition concerning other connections.  <math>\Rightarrow</math> Qij = Qik determine the lowest value, Qii should be equal to reference value (Qik of Ref1).</p> <p>Testing Hypothesis H1 (dotted line): Qii is significantly different from the expected value (<math>p=0.002</math>).  <math>\Rightarrow</math> Support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is not significantly different from the expected value (<math>p=0.333</math>).  <math>\Rightarrow</math> Support for H2.</p>
FM2	<p>Expectations: No impairment on own connection, asymmetric condition concerning other connections.  <math>\Rightarrow</math> Qik determines lowest value, Qii &amp; Qij should be equal to reference value (Qik of Ref1).</p> <p>Testing Hypothesis H1 (dotted line): Qii &amp; Qij are significantly different (Qii: <math>p=0.029</math>, Qij: <math>p=0.011</math>) from the expected values.  <math>\Rightarrow</math> Support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is not significantly different from the expected value (<math>p=0.103</math>).  <math>\Rightarrow</math> Support for H2.</p>
FL3sym	<p>Expectations: No impairment on own connection, symmetric condition concerning other connections.  <math>\Rightarrow</math> Qij = Qik determine the lowest value, Qii should be equal to reference value (Qik of Ref1).</p> <p>Testing Hypothesis H1 (dotted line): Qii is significantly different from the expected value (<math>p=0.000</math>).  <math>\Rightarrow</math> Support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is not significantly different from the expected value (<math>p=0.136</math>).  <math>\Rightarrow</math> Support for H2.</p>

Figure E.5: Detailed results for Experiment ACT3, Part 2.

1. Visualization of Results and Expectations



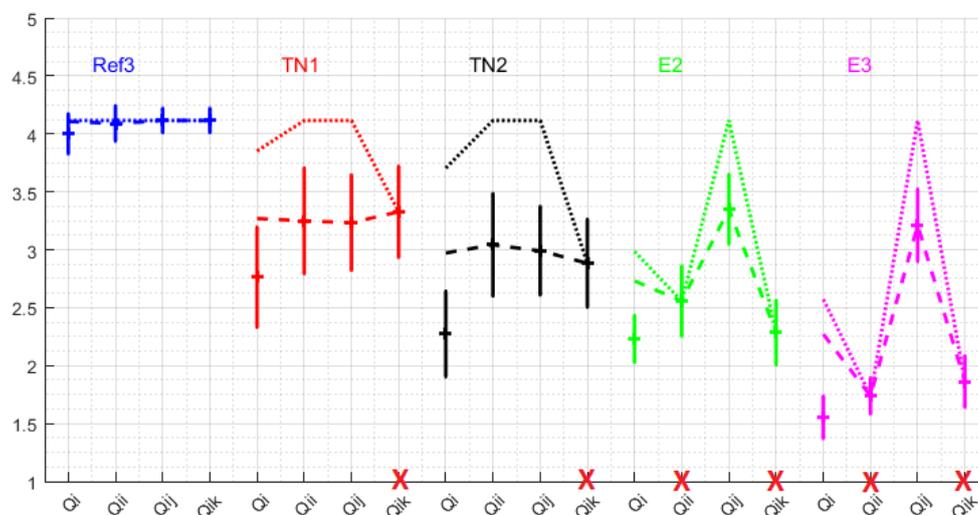
Note: The reference condition Ref<sub>3</sub> is repeated in this plot for better visual comparison.

2. Detailed Analysis

- P<sub>1</sub>**      Expectations: No impairment on own connection, asymmetric condition concerning other connections.  
 ⇒ Q<sub>ik</sub> determines lowest value, Q<sub>ii</sub> & Q<sub>ij</sub> should be equal to reference value (Q<sub>ik</sub> of Ref<sub>1</sub>).
- Testing Hypothesis H<sub>1</sub> (dotted line): Q<sub>ii</sub> & Q<sub>ij</sub> are not significantly different (Q<sub>ii</sub>: p=0.424, Q<sub>ij</sub>: p=0.887) from the expected values.  
 ⇒ No support for H<sub>1</sub>.
- Testing Hypothesis H<sub>2</sub> (dashed line, since this is independent from H<sub>1</sub>): Q<sub>i</sub> is not significantly different from the expected value (p=0.240).  
 ⇒ Support for H<sub>2</sub>.
- P<sub>2</sub>**      Expectations: No impairment on own connection, asymmetric condition concerning other connections.  
 ⇒ Q<sub>ik</sub> determines lowest value, Q<sub>ii</sub> & Q<sub>ij</sub> should be equal to reference value (Q<sub>ik</sub> of Ref<sub>1</sub>).
- Testing Hypothesis H<sub>1</sub> (dotted line): Q<sub>ii</sub> & Q<sub>ij</sub> are not significantly different (Q<sub>ii</sub>: p=0.381, Q<sub>ij</sub>: p=0.119) from the expected values.  
 ⇒ No support for H<sub>1</sub>.
- Testing Hypothesis H<sub>2</sub> (dashed line, since this is independent from H<sub>1</sub>): Q<sub>i</sub> is not significantly different from the expected value (p=0.107).  
 ⇒ Support for H<sub>2</sub>.
- P<sub>3</sub>**      Expectations: No impairment on own connection, asymmetric condition concerning other connections.  
 ⇒ Q<sub>ik</sub> determines lowest value, Q<sub>ii</sub> & Q<sub>ij</sub> should be equal to reference value (Q<sub>ik</sub> of Ref<sub>1</sub>).
- Testing Hypothesis H<sub>1</sub> (dotted line): Q<sub>ii</sub> & Q<sub>ij</sub> are significantly different (Q<sub>ii</sub>: p=0.000, Q<sub>ij</sub>: p=0.001) from the expected values.  
 ⇒ Support for H<sub>1</sub>.
- Testing Hypothesis H<sub>2</sub> (dashed line, since this is independent from H<sub>1</sub>): Q<sub>i</sub> is not significantly different from the expected value (p=0.072).  
 ⇒ Support for H<sub>2</sub>.
- P<sub>4</sub>**      Expectations: No impairment on own connection, asymmetric condition concerning other connections.  
 ⇒ Q<sub>ik</sub> determines lowest value, Q<sub>ii</sub> & Q<sub>ij</sub> should be equal to reference value (Q<sub>ik</sub> of Ref<sub>1</sub>).
- Testing Hypothesis H<sub>1</sub> (dotted line): Q<sub>ii</sub> & Q<sub>ij</sub> are significantly different (Q<sub>ii</sub>: p=0.003, Q<sub>ij</sub>: p=0.000) from the expected values.  
 ⇒ Support for H<sub>1</sub>.
- Testing Hypothesis H<sub>2</sub> (dashed line, since this is independent from H<sub>1</sub>): Q<sub>i</sub> is significantly different from the expected value (p=0.040).  
 ⇒ No support for H<sub>2</sub>.

Figure E.6: Detailed results for Experiment ACT<sub>3</sub>, Part 3.

## 1. Visualization of Results and Expectations



Note: The reference condition Ref<sub>3</sub> is repeated in this plot for better visual comparison.

## 2. Detailed Analysis

- TN<sub>1</sub>**  
 Expectations: No impairment on own connection, asymmetric condition concerning other connections.  
 ⇒ Q<sub>ik</sub> determines lowest value, Q<sub>ii</sub> & Q<sub>ij</sub> should be equal to reference value (Q<sub>ik</sub> of Ref<sub>1</sub>).
- Testing Hypothesis H<sub>1</sub> (dotted line): Q<sub>ii</sub> & Q<sub>ij</sub> are significantly different (Q<sub>ii</sub>: p=0.001, Q<sub>ij</sub>: p=0.000) from the expected values.  
 ⇒ Support for H<sub>1</sub>.
- Testing Hypothesis H<sub>2</sub> (dashed line, since this is independent from H<sub>1</sub>): Q<sub>i</sub> is significantly different from the expected value (p=0.023).  
 ⇒ No support for H<sub>2</sub>.
- TN<sub>2</sub>**  
 Expectations: No impairment on own connection, asymmetric condition concerning other connections.  
 ⇒ Q<sub>ik</sub> determines lowest value, Q<sub>ii</sub> & Q<sub>ij</sub> should be equal to reference value (Q<sub>ik</sub> of Ref<sub>1</sub>).
- Testing Hypothesis H<sub>1</sub> (dotted line): Q<sub>ii</sub> & Q<sub>ij</sub> are significantly different (Q<sub>ii</sub>: p=0.000, Q<sub>ij</sub>: p=0.000) from the expected values.  
 ⇒ Support for H<sub>1</sub>.
- Testing Hypothesis H<sub>2</sub> (dashed line, since this is independent from H<sub>1</sub>): Q<sub>i</sub> is significantly different from the expected value (p=0.001).  
 ⇒ No support for H<sub>2</sub>.
- E<sub>1</sub>**  
 Expectations: Impairment on own connection, asymmetric condition concerning other connections.  
 ⇒ Q<sub>ii</sub> & Q<sub>ik</sub> are lower than reference value Q<sub>ik</sub> of Ref<sub>1</sub>, Q<sub>ij</sub> should be equal to reference value (Q<sub>ik</sub> of Ref<sub>1</sub>).
- Testing Hypothesis H<sub>1</sub> (dotted line): Q<sub>ij</sub> is significantly different from the expected value (p=0.000).  
 ⇒ Support for H<sub>1</sub>.
- Testing Hypothesis H<sub>2</sub> (dashed line, since this is independent from H<sub>1</sub>): Q<sub>i</sub> is significantly different from the expected value (p=0.000).  
 ⇒ No support for H<sub>2</sub>.
- E<sub>2</sub>**  
 Expectations: Impairment on own connection, asymmetric condition concerning other connections.  
 ⇒ Q<sub>ii</sub> & Q<sub>ik</sub> are lower than reference value Q<sub>ik</sub> of Ref<sub>1</sub>, Q<sub>ij</sub> should be equal to reference value (Q<sub>ik</sub> of Ref<sub>1</sub>).
- Testing Hypothesis H<sub>1</sub> (dotted line): Q<sub>ij</sub> is significantly different from the expected value (p=0.000).  
 ⇒ Support for H<sub>1</sub>.
- Testing Hypothesis H<sub>2</sub> (dashed line, since this is independent from H<sub>1</sub>): Q<sub>i</sub> is significantly different from the expected value (p=0.000).  
 ⇒ No support for H<sub>2</sub>.

Figure E.7: Detailed results for Experiment ACT<sub>3</sub>, Part 4.

*F*

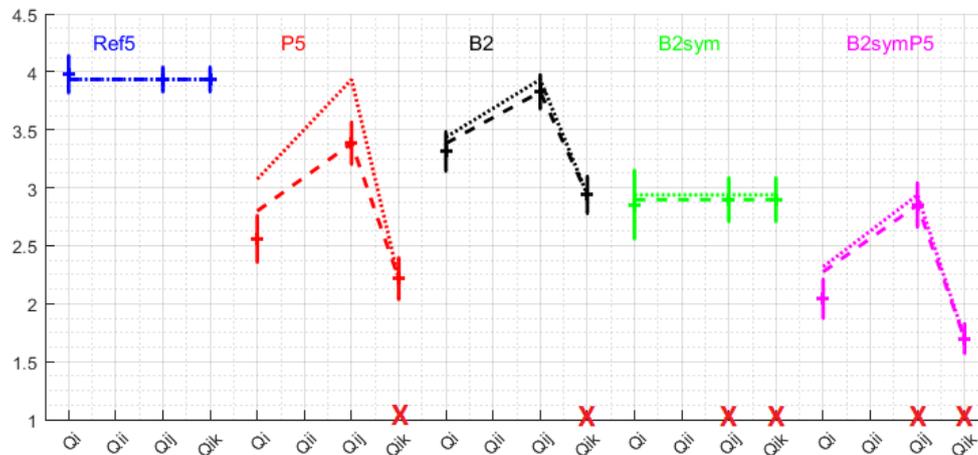
## *Detailed Results: Study on Impact of Audio-Only Telemeeting - Listening-Only Tests*

*What this chapter is about*

The present appendix presents the detailed results of the two listening-only tests LOT<sub>1</sub> & LOT<sub>2</sub> on the impact of technical conditions on quality perception. First, each figure shows a visualization of the obtained data and the expectations according to Section 8.4. Then, each figure discusses per technical condition, in how far the expectations are fulfilled and how the current results relate to the different hypotheses formulated in Section 8.4. The individual results are then compiled in the main text in Section 8.4.

F.1 Experiment LOT<sub>1</sub>

## 1. Visualization of Results and Expectations

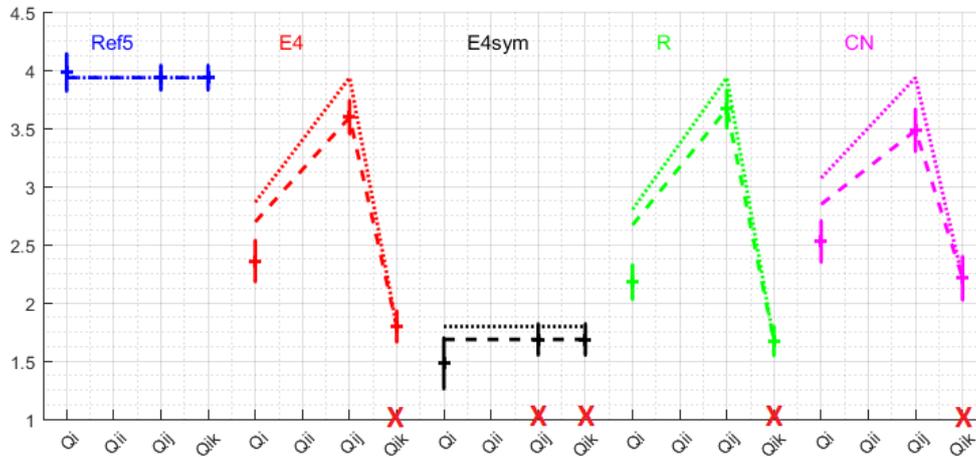


## 2. Detailed Analysis

Ref5	All expected values lie within confidence intervals of $Q_i$ & $Q_{ij}=Q_{ik}$ , and the observed values for $Q_i$ & $Q_{ij}=Q_{ik}$ are not significantly different from maximum value of whole test (3.9792, $Q_i$ of Ref5). ⇒ Ref5 is confirmed as reference condition for this test.
P5	Expectations: Asymmetric condition. ⇒ $Q_{ik}$ determines lowest value, $Q_{ij}$ should be equal to reference value ( $Q_{ik}$ of Ref5). Testing Hypothesis H1 (dotted line): $Q_{ij}$ is significantly different ( $p=0.000$ ) from the expected value. ⇒ Support for H1. Testing Hypothesis H2 (dashed line, since this is independent from H1): $Q_i$ is significantly different from the expected value ( $p=0.000$ ). ⇒ No support for H2.
B2	Expectations: Asymmetric condition. ⇒ $Q_{ik}$ determines lowest value, $Q_{ij}$ should be equal to reference value ( $Q_{ik}$ of Ref5). Testing Hypothesis H1 (dotted line): $Q_{ij}$ is not significantly different ( $p=0.145$ ) from the expected value. ⇒ No support for H1. Testing Hypothesis H2 (dashed line, since this is independent from H1): $Q_i$ is not significantly different from the expected value ( $p=0.408$ ). ⇒ Support for H2.
B2sym	Expectations: Symmetric condition. ⇒ $Q_{ij} = Q_{ik}$ should be equal to corresponding value of asymmetric condition ( $Q_{ik}$ of B2). Testing Hypothesis H1 (dotted line): $Q_{ij}$ is not significantly different ( $p=0.657$ ) from the expected value. ⇒ No support for H1. Testing Hypothesis H2 (dashed line, since this is independent from H1): $Q_i$ is not significantly different from the expected value ( $p=0.774$ ). ⇒ Support for H2.
B2symP5	Expectations: Asymmetric condition concerning P5, symmetric condition concerning B2. ⇒ $Q_{ij}$ should be equal to corresponding value of asymmetric condition concerning B2 ( $Q_{ik}$ of B2). Testing Hypothesis H1 (dotted line): $Q_{ij}$ is not significantly different ( $p=0.360$ ) from the expected value. ⇒ No support for H1. Testing Hypothesis H2 (dashed line, since this is independent from H1): $Q_i$ is significantly different from the expected value ( $p=0.006$ ). ⇒ No support for H2.

Figure F.1: Detailed results for Experiment LOT<sub>1</sub>, Part 1.

1. Visualization of Results and Expectations



Note: The reference condition Ref5 is repeated in this plot for better visual comparison.

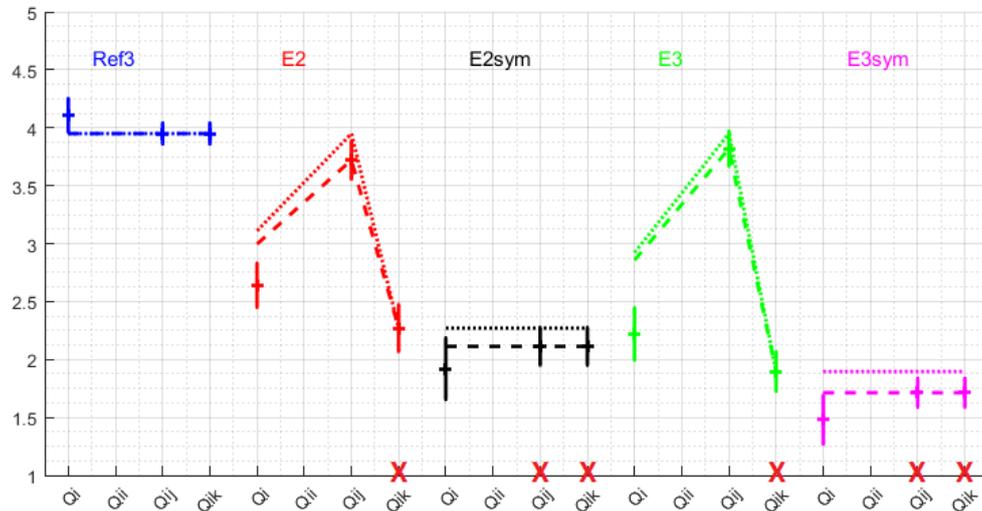
2. Detailed Analysis

E4	<p>Expectations: Asymmetric condition.                  ⇒ Qik determines lowest value, Qij should be equal to reference value (Qik of Ref5).</p> <p>Testing Hypothesis H1 (dotted line): Qij is significantly different (p=0.000) from the expected value.                  ⇒ Support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is significantly different from the expected value (p=0.000).                  ⇒ No support for H2.</p>
E4sym	<p>Expectations: Symmetric condition.                  ⇒ Qij = Qik should be equal to corresponding value of asymmetric condition (Qik of E4).</p> <p>Testing Hypothesis H1 (dotted line): Qij is not significantly different (p=0.098) from the expected value.                  ⇒ No support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is not significantly different from the expected value (p=0.062).                  ⇒ Support for H2.</p>
R	<p>Expectations: Asymmetric condition.                  ⇒ Qik determines lowest value, Qij should be equal to reference value (Qik of Ref5).</p> <p>Testing Hypothesis H1 (dotted line): Qij is significantly different (p=0.001) from the expected value.                  ⇒ Support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is significantly different from the expected value (p=0.000).                  ⇒ No support for H2.</p>
CN	<p>Expectations: Asymmetric condition.                  ⇒ Qik determines lowest value, Qij should be equal to reference value (Qik of Ref5).</p> <p>Testing Hypothesis H1 (dotted line): Qij is significantly different (p=0.000) from the expected value.                  ⇒ Support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is significantly different from the expected value (p=0.000).                  ⇒ No support for H2.</p>

Figure F.2: Detailed results for Experiment LOT1, Part 2.

## F.2 Experiment LOT<sub>2</sub>

### 1. Visualization of Results and Expectations

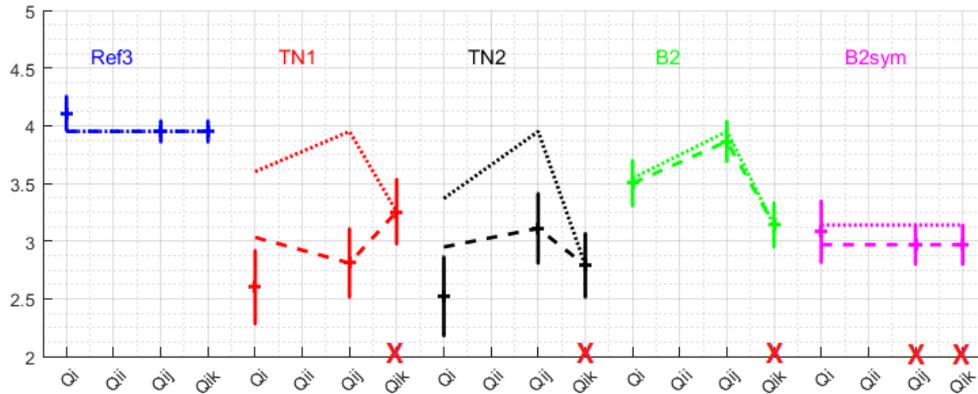


### 2. Detailed Analysis

Ref <sub>3</sub>	The expected value for $Q_{ij}=Q_{ik}$ lies within the confidence interval, the expected value for $Q_i$ lies outside the confidence interval ( $p=0.034$ ), whereas the observed value is better than the expected one, the observed values for $Q_i=4.1067$ and $Q_{ij}=Q_{ik}=3.9499$ are the maximum values of the whole test $\Rightarrow$ Ref <sub>3</sub> is confirmed as reference condition for this test. However, $Q_i$ is significantly better than the individual $Q_{ij}$ , suggesting some scale interpretation effects.
E <sub>2</sub>	Expectations: Asymmetric condition. $\Rightarrow Q_{ik}$ determines lowest value, $Q_{ij}$ should be equal to reference value ( $Q_{ik}$ of Ref <sub>5</sub> ). Testing Hypothesis H <sub>1</sub> (dotted line): $Q_{ij}$ is significantly different ( $p=0.007$ ) from the expected value. $\Rightarrow$ Support for H <sub>1</sub> . Testing Hypothesis H <sub>2</sub> (dashed line, since this is independent from H <sub>1</sub> ): $Q_i$ is significantly different from the expected value ( $p=0.000$ ). $\Rightarrow$ No support for H <sub>2</sub> .
E <sub>2sym</sub>	Expectations: Symmetric condition. $\Rightarrow Q_{ij} = Q_{ik}$ should be equal to corresponding value of asymmetric condition ( $Q_{ik}$ of E <sub>2</sub> ). Testing Hypothesis H <sub>1</sub> (dotted line): $Q_{ij}$ is not significantly different ( $p=0.055$ ) from the expected value. $\Rightarrow$ No support for H <sub>1</sub> . Testing Hypothesis H <sub>2</sub> (dashed line, since this is independent from H <sub>1</sub> ): $Q_i$ is not significantly different from the expected value ( $p=0.144$ ). $\Rightarrow$ Support for H <sub>2</sub> .
E <sub>3</sub>	Expectations: Asymmetric condition. $\Rightarrow Q_{ik}$ determines lowest value, $Q_{ij}$ should be equal to reference value ( $Q_{ik}$ of Ref <sub>5</sub> ). Testing Hypothesis H <sub>1</sub> (dotted line): $Q_{ij}$ is not significantly different ( $p=0.084$ ) from the expected value. $\Rightarrow$ No support for H <sub>1</sub> . Testing Hypothesis H <sub>2</sub> (dashed line, since this is independent from H <sub>1</sub> ): $Q_i$ is significantly different from the expected value ( $p=0.000$ ). $\Rightarrow$ No support for H <sub>2</sub> .
E <sub>3sym</sub>	Expectations: Symmetric condition. $\Rightarrow Q_{ij} = Q_{ik}$ should be equal to corresponding value of asymmetric condition ( $Q_{ik}$ of E <sub>3</sub> ). Testing Hypothesis H <sub>1</sub> (dotted line): $Q_{ij}$ is significantly different ( $p=0.004$ ) from the expected value. $\Rightarrow$ Support for H <sub>1</sub> . Testing Hypothesis H <sub>2</sub> (dashed line, since this is independent from H <sub>1</sub> ): $Q_i$ is significantly different from the expected value ( $p=0.032$ ). $\Rightarrow$ No support for H <sub>2</sub> .

Figure F.3: Detailed results for Experiment LOT<sub>2</sub>, Part 1.

1. Visualization of Results and Expectations



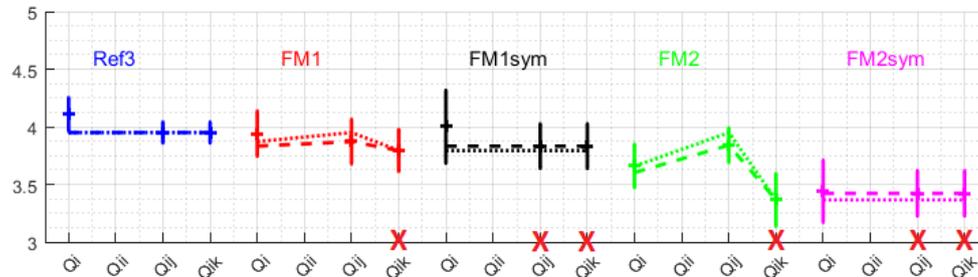
Note: The reference condition Ref5 is repeated in this plot for better visual comparison.

2. Detailed Analysis

TN1	<p>Expectations: Asymmetric condition.                  ⇒ Qik determines lowest value, Qij should be equal to reference value (Qik of Ref5).</p> <p>Testing Hypothesis H1 (dotted line): Qij is significantly different (p=0.000) from the expected value.                  ⇒ Support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is significantly different from the expected value (p=0.009).                  ⇒ No support for H2.</p>
TN2	<p>Expectations: Asymmetric condition.                  ⇒ Qik determines lowest value, Qij should be equal to reference value (Qik of Ref5).</p> <p>Testing Hypothesis H1 (dotted line): Qij is significantly different (p=0.000) from the expected value.                  ⇒ Support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is significantly different from the expected value (p=0.016).                  ⇒ No support for H2.</p>
B2	<p>Expectations: Asymmetric condition.                  ⇒ Qik determines lowest value, Qij should be equal to reference value (Qik of Ref5).</p> <p>Testing Hypothesis H1 (dotted line): Qij is not significantly different (p=0.311) from the expected value.                  ⇒ No support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is not significantly different from the expected value (p=0.994).                  ⇒ Support for H2.</p>
B2sym	<p>Expectations: Symmetric condition.                  ⇒ Qij = Qik should be equal to corresponding value of asymmetric condition (Qik of B2).</p> <p>Testing Hypothesis H1 (dotted line): Qij is significantly different (p=0.045) from the expected value.                  ⇒ Support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is not significantly different from the expected value (p=0.389).                  ⇒ Support for H2.</p>

Figure F.4: Detailed results for Experiment LOT2, Part 2.

## 1. Visualization of Results and Expectations



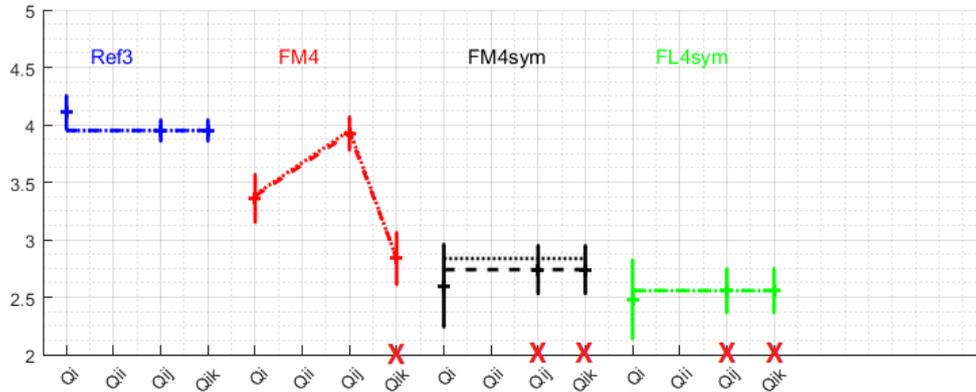
Note: The reference condition Ref5 is repeated in this plot for better visual comparison.

## 2. Detailed Analysis

FM1	<p>Expectations: Asymmetric condition.  <math>\Rightarrow</math> Qik determines lowest value, Qij should be equal to reference value (Qik of Ref5).</p> <p>Testing Hypothesis H1 (dotted line): Qij is not significantly different (<math>p=0.411</math>) from the expected value.  <math>\Rightarrow</math> No support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is not significantly different from the expected value (<math>p=0.269</math>).  <math>\Rightarrow</math> Support for H2.</p> <p>Special observation: All values are essentially equal to reference condition. <math>\Rightarrow</math> Impairment not strong enough to elicit effect in quality ratings.</p>
FM1sym	<p>Expectations: Symmetric condition.  <math>\Rightarrow</math> Qij = Qik should be equal to corresponding value of asymmetric condition (Qik of B2).</p> <p>Testing Hypothesis H1 (dotted line): Qij is not significantly different (<math>p=0.692</math>) from the expected value.  <math>\Rightarrow</math> No support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is not significantly different from the expected value (<math>p=0.282</math>).  <math>\Rightarrow</math> Support for H2.</p> <p>Special observation: All values are essentially equal to reference condition. <math>\Rightarrow</math> Impairment not strong enough to elicit effect in quality ratings.</p>
FM2	<p>Expectations: Asymmetric condition.  <math>\Rightarrow</math> Qik determines lowest value, Qij should be equal to reference value (Qik of Ref5).</p> <p>Testing Hypothesis H1 (dotted line): Qij is not significantly different (<math>p=0.141</math>) from the expected value.  <math>\Rightarrow</math> No support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is not significantly different from the expected value (<math>p=0.541</math>).  <math>\Rightarrow</math> Support for H2.</p>
FM2sym	<p>Expectations: Symmetric condition.  <math>\Rightarrow</math> Qij = Qik should be equal to corresponding value of asymmetric condition (Qik of B2).</p> <p>Testing Hypothesis H1 (dotted line): Qij is not significantly different (<math>p=0.563</math>) from the expected value.  <math>\Rightarrow</math> No support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is not significantly different from the expected value (<math>p=0.891</math>).  <math>\Rightarrow</math> Support for H2.</p>

Figure F.5: Detailed results for Experiment LOT2, Part 3.

1. Visualization of Results and Expectations



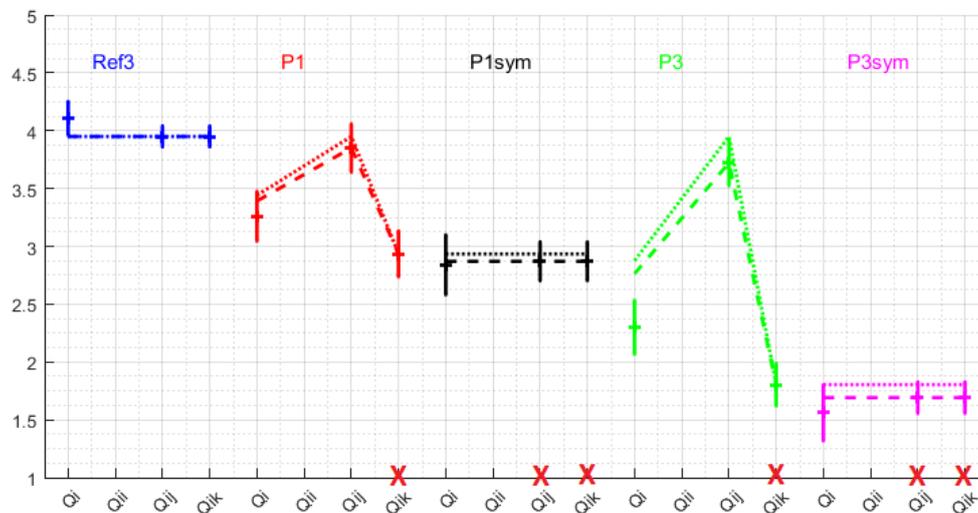
Note: The reference condition Ref5 is repeated in this plot for better visual comparison.

2. Detailed Analysis

FM4	<p>Expectations: Asymmetric condition.                      ⇒ Qik determines lowest value, Qij should be equal to reference value (Qik of Ref5).</p> <p>Testing Hypothesis H1 (dotted line): Qij is not significantly different (p=0.712) from the expected value.                      ⇒ No support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is not significantly different from the expected value (p=0.845).                      ⇒ Support for H2.</p> <p>Special observation: All values are essentially equal to reference condition. ⇒ Impairment not strong enough to elicit effect in quality ratings.</p>
FM4sym	<p>Expectations: Symmetric condition.                      ⇒ Qij = Qik should be equal to corresponding value of asymmetric condition (Qik of B2).</p> <p>Testing Hypothesis H1 (dotted line): Qij is not significantly different (p=0.351) from the expected value.                      ⇒ No support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is not significantly different from the expected value (p=0.428).                      ⇒ Support for H2.</p> <p>Special observation: All values are essentially equal to reference condition. ⇒ Impairment not strong enough to elicit effect in quality ratings.</p>
FL4sym	<p>Expectations: Symmetric condition.                      ⇒ Qij = Qik is lower than reference condition. There is no corresponding asymmetric condition to compared with.</p> <p>Testing Hypothesis H1 (dotted line): Can not be tested.                      ⇒ No support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is not significantly different from the expected value (p=0.647).                      ⇒ Support for H2.</p>

Figure F.6: Detailed results for Experiment LOT2, Part 4.

## 1. Visualization of Results and Expectations



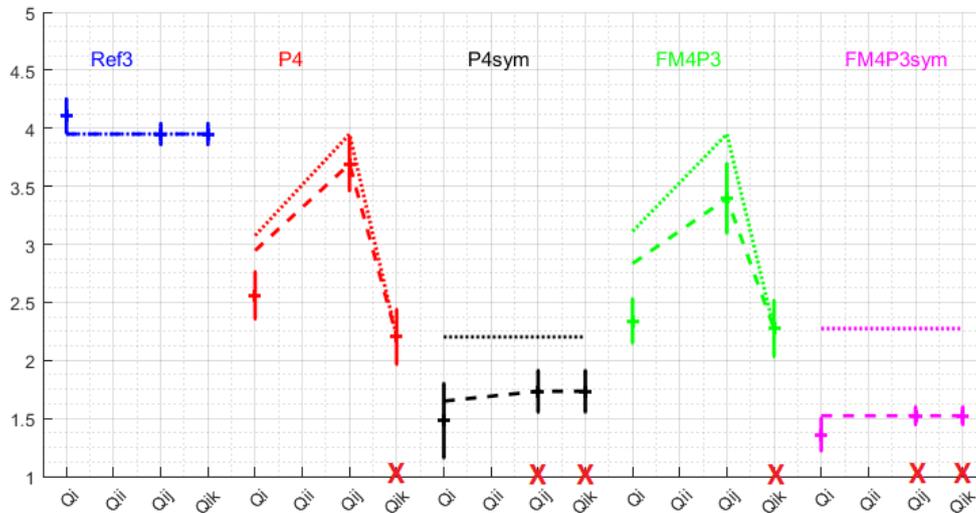
Note: The reference condition Ref5 is repeated in this plot for better visual comparison.

## 2. Detailed Analysis

P1	<p>Expectations: Asymmetric condition.  <math>\Rightarrow</math> Qik determines lowest value, Qij should be equal to reference value (Qik of Ref5).</p> <p>Testing Hypothesis H1 (dotted line): Qij is not significantly different (<math>p=0.335</math>) from the expected value.  <math>\Rightarrow</math> No support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is not significantly different from the expected value (<math>p=0.220</math>).  <math>\Rightarrow</math> Support for H2.</p> <p>Special observation: All values are essentially equal to reference condition. <math>\Rightarrow</math> Impairment not strong enough to elicit effect in quality ratings.</p>
P1sym	<p>Expectations: Symmetric condition.  <math>\Rightarrow</math> Qij = Qik should be equal to corresponding value of asymmetric condition (Qik of B2).</p> <p>Testing Hypothesis H1 (dotted line): Qij is not significantly different (<math>p=0.441</math>) from the expected value.  <math>\Rightarrow</math> No support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is not significantly different from the expected value (<math>p=0.813</math>).  <math>\Rightarrow</math> Support for H2.</p> <p>Special observation: All values are essentially equal to reference condition. <math>\Rightarrow</math> Impairment not strong enough to elicit effect in quality ratings.</p>
P3	<p>Expectations: Asymmetric condition.  <math>\Rightarrow</math> Qik determines lowest value, Qij should be equal to reference value (Qik of Ref5).</p> <p>Testing Hypothesis H1 (dotted line): Qij is significantly different (<math>p=0.024</math>) from the expected value.  <math>\Rightarrow</math> Support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is significantly different from the expected value (<math>p=0.000</math>).  <math>\Rightarrow</math> No support for H2.</p>
P3sym	<p>Expectations: Symmetric condition.  <math>\Rightarrow</math> Qij = Qik should be equal to corresponding value of asymmetric condition (Qik of B2).</p> <p>Testing Hypothesis H1 (dotted line): Qij is not significantly different (<math>p=0.099</math>) from the expected value.  <math>\Rightarrow</math> No support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is not significantly different from the expected value (<math>p=0.271</math>).  <math>\Rightarrow</math> Support for H2.</p>

Figure F.7: Detailed results for Experiment LOT2, Part 5.

1. Visualization of Results and Expectations



Note: The reference condition Ref5 is repeated in this plot for better visual comparison.

2. Detailed Analysis

P4	<p>Expectations: Asymmetric condition.                  ⇒ <math>Q_{ik}</math> determines lowest value, <math>Q_{ij}</math> should be equal to reference value (<math>Q_{ik}</math> of Ref5).</p> <p>Testing Hypothesis H1 (dotted line): <math>Q_{ij}</math> is significantly different (<math>p=0.026</math>) from the expected value.                  ⇒ Support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): <math>Q_i</math> is significantly different from the expected value (<math>p=0.000</math>).                  ⇒ No support for H2.</p> <p>Special observation: All values are essentially equal to reference condition. ⇒ Impairment not strong enough to elicit effect in quality ratings.</p>
P4sym	<p>Expectations: Symmetric condition.                  ⇒ <math>Q_{ij} = Q_{ik}</math> should be equal to corresponding value of asymmetric condition (<math>Q_{ik}</math> of B2).</p> <p>Testing Hypothesis H1 (dotted line): <math>Q_{ij}</math> is significantly different (<math>p=0.000</math>) from the expected value.                  ⇒ Support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): <math>Q_i</math> is not significantly different from the expected value (<math>p=0.284</math>).                  ⇒ Support for H2.</p> <p>Special observation: All values are essentially equal to reference condition. ⇒ Impairment not strong enough to elicit effect in quality ratings.</p>
FM4P3	<p>Expectations: Asymmetric condition.                  ⇒ <math>Q_{ik}</math> determines lowest value, <math>Q_{ij}</math> should be equal to reference value (<math>Q_{ik}</math> of Ref5).</p> <p>Testing Hypothesis H1 (dotted line): <math>Q_{ij}</math> is significantly different (<math>p=0.000</math>) from the expected value.                  ⇒ Support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): <math>Q_i</math> is significantly different from the expected value (<math>p=0.000</math>).                  ⇒ No support for H2.</p>
FM4P3sym	<p>Expectations: Symmetric condition.                  ⇒ <math>Q_{ij} = Q_{ik}</math> should be equal to corresponding value of asymmetric condition (<math>Q_{ik}</math> of B2).</p> <p>Testing Hypothesis H1 (dotted line): <math>Q_{ij}</math> is significantly different (<math>p=0.000</math>) from the expected value.                  ⇒ Support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): <math>Q_i</math> is significantly different from the expected value (<math>p=0.022</math>).                  ⇒ No support for H2.</p>

Figure F8: Detailed results for Experiment LOT2, Part 6.

## G

# *Detailed Results: Study on Impact of Audiovisual Telemeeting - Conversation Tests*

### *What this chapter is about*

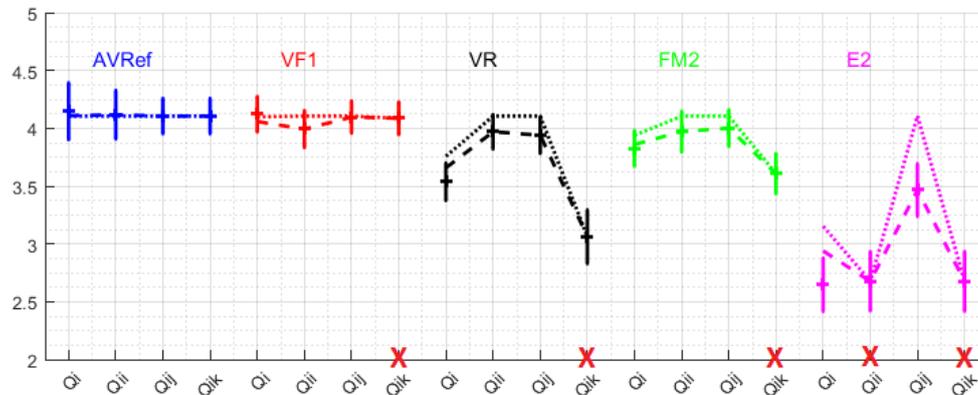
The present appendix presents the detailed results of the two audiovisual conversation tests AVCT<sub>1</sub> & AVCT<sub>2</sub> on the impact of the conversation task and the technical conditions on quality perception. First, each section shows in a table the results of the comparison of the two used conversation tasks *Survival Task* and *Celebrity Name Guessing Game*. Then, each section shows in a number of figure the impact of the technical conditions. Here, each figure shows a visualization of the obtained data and the expectations according to Section 8.5. Furthermore, each figure discusses per technical condition, in how far the expectations are fulfilled and how the current results relate to the different hypotheses formulated in Section 8.5. The individual results are then compiled in the main text in Section 8.5.

G.1 Experiment AVCT<sub>1</sub>

Condition	Aggregation Level $Q_x$	Impairment definition			p
<i>AVRef</i>	$Q_i$	0	0	0	0.921
	$Q_{ii}$	0	0	0	0.121
	$Q_{ij} = Q_{ik}$	0	0	0	0.593
<i>VF1</i>	$Q_i$	0	0	<i>VF1</i>	0.695
	$Q_{ii}$	0	0	<i>VF1</i>	0.534
	$Q_{ij}$	0	0	<i>VF1</i>	0.612
	$Q_{ik}$	<i>VF1</i>	0	0	0.416
<i>VR</i>	$Q_i$	0	0	<i>VR</i>	0.083
	$Q_{ii}$	0	0	<i>VR</i>	0.502
	$Q_{ij}$	0	0	<i>Vr</i>	0.066
	$Q_{ik}$	<i>VR</i>	0	0	0.928
<i>FM2</i>	$Q_i$	0	0	<i>FM2</i>	0.613
	$Q_{ii}$	0	0	<i>FM2</i>	0.850
	$Q_{ij}$	0	0	<i>FM2</i>	0.608
	$Q_{ik}$	<i>FM2</i>	0	0	0.534
<i>E2</i>	$Q_i$	<i>TE2</i>	0	<i>LE2</i>	0.526
	$Q_{ii}$	<i>TE2</i>	0	<i>LE2</i>	0.409
	$Q_{ij}$	0	<i>TE2</i>	<i>LE2</i>	0.581
	$Q_{ik}$	<i>LE2</i>	<i>TE2</i>	0	0.167
<i>FM2 – VF1</i>	$Q_i$	0	0	<i>FM2 – VF1</i>	0.384
	$Q_{ii}$	0	0	<i>FM2 – VF1</i>	0.124
	$Q_{ij}$	0	0	<i>FM2 – VF1</i>	0.742
	$Q_{ik}$	<i>FM2 – VF1</i>	0	0	0.118
<i>E2 – VF1</i>	$Q_i$	<i>TE2</i>	0	<i>LE2 – VF2</i>	0.403
	$Q_{ii}$	<i>TE2</i>	0	<i>LE2 – VF2</i>	0.332
	$Q_{ij}$	0	<i>TE2</i>	<i>LE2 – VF2</i>	0.340
	$Q_{ik}$	<i>LE2 – VF2</i>	<i>TE2</i>	0	0.098
<i>FM2 – VR</i>	$Q_i$	0	0	<i>FM2 – VR</i>	0.933
	$Q_{ii}$	0	0	<i>FM2 – VR</i>	0.947
	$Q_{ij}$	0	0	<i>FM2 – VR</i>	0.975
	$Q_{ik}$	<i>FM2 – VR</i>	0	0	0.996
<i>E2 – VR</i>	$Q_i$	<i>TE2</i>	0	<i>LE2 – VR</i>	0.114
	$Q_{ii}$	<i>TE2</i>	0	<i>LE2 – VR</i>	0.092
	$Q_{ij}$	0	<i>TE2</i>	<i>LE2 – VR</i>	0.729
	$Q_{ik}$	<i>LE2 – VR</i>	<i>TE2</i>	0	0.089

Table G.1: Independent t-Tests between the two conversation scenarios Survival Task and Name Guessing Game, computed per condition for the audio-visual conversation test AVCT<sub>1</sub>.

## 1. Visualization of Results and Expectations

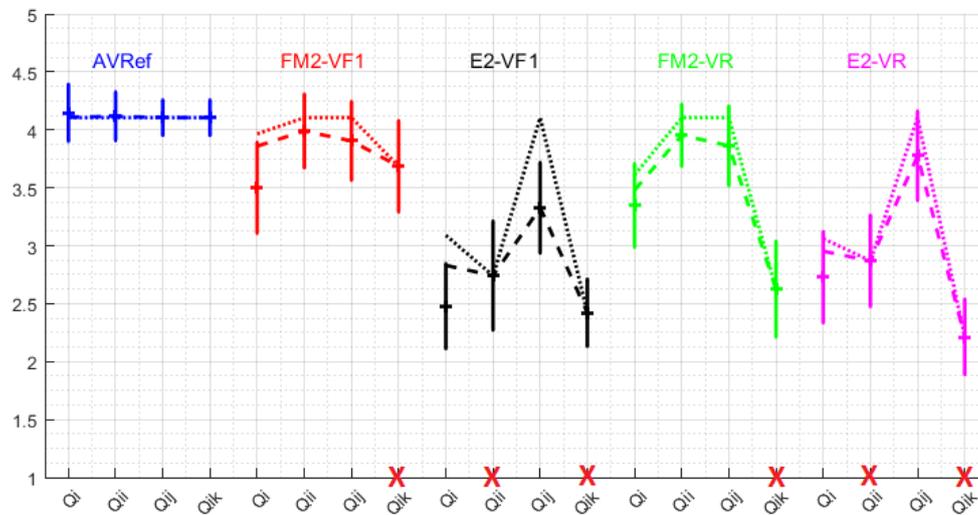


## 2. Detailed Analysis

AVRef	All expected values lie within confidence intervals of $Q_i$ , $Q_{ii}$ , & $Q_{ij}=Q_{ik}$ , and mean value of $Q_i=4.1471$ is maximum value of whole test and observed mean values of $Q_{ii}$ & $Q_{ij}=Q_{ik}$ are not significantly different that value. $\Rightarrow$ AVRef is confirmed as reference condition for this test.
VF1	Expectations: No impairment on own connection, asymmetric condition concerning other connections. $\Rightarrow$ $Q_{ik}$ determines lowest value, $Q_{ii}$ & $Q_{ij}$ should be equal to reference value $Q_{ik}$ of AVRef. Testing Hypothesis H1 (dotted line): $Q_{ii}$ & $Q_{ij}$ are not significantly different ( $Q_{ii}$ : $p=0.162$ , $Q_{ij}$ : $p=0.900$ ) from the expected values. $\Rightarrow$ No support for H1. Testing Hypothesis H2 (dashed line, since this is independent from H1): $Q_i$ is not significantly different from the expected value ( $p=0.413$ ). $\Rightarrow$ Support for H2. Special observation: All values are essentially equal to reference condition. $\Rightarrow$ Impairment not strong enough to elicit effect in quality ratings.
VR	Expectations: No impairment on own connection, asymmetric condition concerning other connections. $\Rightarrow$ $Q_{ik}$ determines lowest value, $Q_{ii}$ & $Q_{ij}$ should be equal to reference value $Q_{ik}$ of AVRef. Testing Hypothesis H1 (dotted line): $Q_{ii}$ is not significantly different ( $p=0.071$ ) from the expected value, $Q_{ij}$ is significantly different ( $p=0.038$ ) from the expected value. $\Rightarrow$ Support for H1. Testing Hypothesis H2 (dashed line, since this is independent from H1): $Q_i$ is not significantly different from the expected value ( $p=0.142$ ). $\Rightarrow$ Support for H2.
FM2	Expectations: No impairment on own connection, asymmetric condition concerning other connections. $\Rightarrow$ $Q_{ik}$ determines lowest value, $Q_{ii}$ & $Q_{ij}$ should be equal to reference value $Q_{ik}$ of AVRef. Testing Hypothesis H1 (dotted line): $Q_{ii}$ & $Q_{ij}$ are not significantly different ( $Q_{ii}$ : $p=0.131$ , $Q_{ij}$ : $p=0.190$ ) from the expected values. $\Rightarrow$ No support for H1. Testing Hypothesis H2 (dashed line, since this is independent from H1): $Q_i$ is not significantly different from the expected value ( $p=0.656$ ). $\Rightarrow$ Support for H2.
E2	Expectations: Impairment on own connection, asymmetric condition concerning other connections. $\Rightarrow$ $Q_{ii}$ & $Q_{ik}$ are lower than reference value $Q_{ik}$ of Ref1, $Q_{ij}$ should be equal to reference value ( $Q_{ik}$ of Ref1). Testing Hypothesis H1 (dotted line): $Q_{ij}$ is significantly different from the expected value ( $p=0.000$ ). $\Rightarrow$ Support for H1. Testing Hypothesis H2 (dashed line, since this is independent from H1): $Q_i$ is significantly different from the expected value ( $p=0.013$ ). $\Rightarrow$ No support for H2.

Figure G.1: Detailed results for Experiment AVCT1, Part 1.

1. Visualization of Results and Expectations



Note: The reference condition Ref1 is repeated in this plot for better visual comparison.

2. Detailed Analysis

FM2-VF1	<p>Expectations: No impairment on own connection, asymmetric condition concerning other connections.                  ⇒ Qik determines lowest value, Qii &amp; Qij should be equal to reference value Qik of AVRef.</p> <p>Testing Hypothesis H1 (dotted line): Qii &amp; Qij are not significantly different (Qii: p=0.450, Qij: p=0.227) from the expected values.                  ⇒ No support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is not significantly different from the expected value (p=0.069).                  ⇒ Support for H2.</p>
E2-VF1	<p>Expectations: Impairment on own connection, asymmetric condition concerning other connections.                  ⇒ Qii &amp; Qik are lower than reference value Qik of Ref1, Qij should be equal to reference value (Qik of Ref1).</p> <p>Testing Hypothesis H1 (dotted line): Qij is significantly different from the expected value (p=0.000).                  ⇒ Support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is not significantly different from the expected value (p=0.059).                  ⇒ Support for H2.</p>
FM2-VR	<p>Expectations: No impairment on own connection, asymmetric condition concerning other connections.                  ⇒ Qik determines lowest value, Qii &amp; Qij should be equal to reference value Qik of AVRef.</p> <p>Testing Hypothesis H1 (dotted line): Qii &amp; Qij are not significantly different (Qii: p=0.250, Qij: p=0.153) from the expected values.                  ⇒ No support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is not significantly different from the expected value (p=0.232).                  ⇒ Support for H2.</p>
E2-VR	<p>Expectations: Impairment on own connection, asymmetric condition concerning other connections.                  ⇒ Qii &amp; Qik are lower than reference value Qik of Ref1, Qij should be equal to reference value (Qik of Ref1).</p> <p>Testing Hypothesis H1 (dotted line): Qij is not significantly different from the expected value (p=0.089).                  ⇒ No support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is not significantly different from the expected value (p=0.245).                  ⇒ Support for H2.</p>

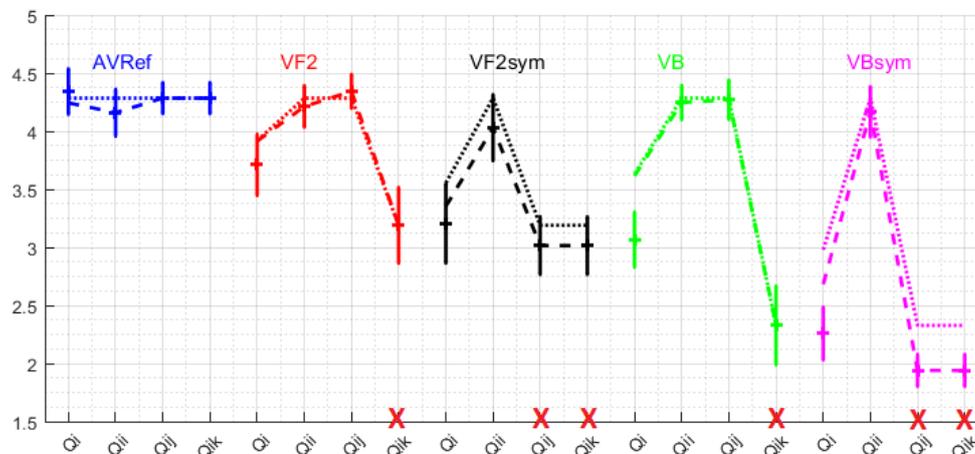
Figure G.2: Detailed results for Experiment AVCT1, Part 2.

G.2 Experiment AVCT<sub>2</sub>

Condition	Aggregation Level $Q_x$	Impairment definition			p
<i>AVRef</i>	$Q_i$	0	0	0	0.211
	$Q_{ii}$	0	0	0	0.189
	$Q_{ij} = Q_{ik}$	0	0	0	0.086
<i>VF2</i>	$Q_i$	0	0	<i>VF2</i>	0.552
	$Q_{ii}$	0	0	<i>VF2</i>	0.246
	$Q_{ij}$	0	0	<i>VF2</i>	0.220
	$Q_{ik}$	<i>VF2</i>	0	0	0.523
<i>VF2sym</i>	$Q_i$	0	<i>VF2</i>	<i>VF2</i>	0.511
	$Q_{ii}$	0	<i>VF2</i>	<i>VF2</i>	0.276
	$Q_{ij} = Q_{ik}$	<i>VF2</i>	0	<i>VF2</i>	0.062
<i>VB</i>	$Q_i$	0	0	<i>VB</i>	0.585
	$Q_{ii}$	0	0	<i>VB</i>	0.062
	$Q_{ij}$	0	0	<i>VB</i>	0.004*
	$Q_{ik}$	<i>VB</i>	0	0	0.507
<i>VF2sym</i>	$Q_i$	0	<i>VB</i>	<i>VB</i>	0.573
	$Q_{ii}$	0	<i>VB</i>	<i>VB</i>	0.827
	$Q_{ij} = Q_{ik}$	<i>VB</i>	0	<i>VB</i>	0.943
<i>E2</i>	$Q_i$	<i>TE2</i>	0	<i>LE2</i>	0.882
	$Q_{ii}$	<i>TE2</i>	0	<i>LE2</i>	0.922
	$Q_{ij}$	0	<i>TE2</i>	<i>LE2</i>	0.271
	$Q_{ik}$	<i>LE2</i>	<i>TE2</i>	0	0.525
<i>FM4</i>	$Q_i$	0	0	<i>FM4</i>	0.511
	$Q_{ii}$	0	0	<i>FM4</i>	0.048*
	$Q_{ij}$	0	0	<i>FM4</i>	0.480
	$Q_{ik}$	<i>FM4</i>	0	0	0.965
<i>FM4sym</i>	$Q_i$	0	<i>FM4</i>	<i>FM4</i>	0.742
	$Q_{ii}$	0	<i>FM4</i>	<i>FM4</i>	0.016*
	$Q_{ij} = Q_{ik}$	<i>FM4</i>	0	<i>FM4</i>	0.109
<i>FM4P3</i>	$Q_i$	0	0	<i>FM4P3</i>	0.122
	$Q_{ii}$	0	0	<i>FM4P3</i>	0.163
	$Q_{ij}$	0	0	<i>FM4P3</i>	0.276
	$Q_{ik}$	<i>FM4P3</i>	0	0	0.102
<i>FM4P3sym</i>	$Q_i$	0	<i>FM4P3</i>	<i>FM4P3</i>	0.793
	$Q_{ii}$	0	<i>FM4P3</i>	<i>FM4P3</i>	0.897
	$Q_{ij} = Q_{ik}$	<i>FM4P3</i>	0	<i>FM4P3</i>	0.994
<i>FM4 – AVP3</i>	$Q_i$	0	0	<i>FM4 – AVP3</i>	0.722
	$Q_{ii}$	0	0	<i>FM4 – AVP3</i>	0.294
	$Q_{ij}$	0	0	<i>FM4 – AVP3</i>	0.433
	$Q_{ik}$	<i>FM4 – AVP3</i>	0	0	0.380
<i>FM4 – VF2</i>	$Q_i$	0	0	<i>FM4 – VF2</i>	0.096
	$Q_{ii}$	0	0	<i>FM4 – VF2</i>	0.467
	$Q_{ij}$	0	0	<i>FM4 – VF2</i>	0.165
	$Q_{ik}$	<i>FM4 – VF2</i>	0	0	0.250
<i>FM4P3 – VF2</i>	$Q_i$	0	0	<i>FM4P3 – VF2</i>	0.122
	$Q_{ii}$	0	0	<i>FM4P3 – VF2</i>	0.066
	$Q_{ij}$	0	0	<i>FM4P3 – VF2</i>	0.001*
	$Q_{ik}$	<i>FM4P3 – VF2</i>	0	0	0.052
<i>FM4 – VB</i>	$Q_i$	0	0	<i>FM4 – VB</i>	0.157
	$Q_{ii}$	0	0	<i>FM4 – VB</i>	0.494
	$Q_{ij}$	0	0	<i>FM4 – VB</i>	0.960
	$Q_{ik}$	<i>FM4 – VB</i>	0	0	0.463
<i>FM4P3 – VB</i>	$Q_i$	0	0	<i>FM4P3 – VB</i>	0.930
	$Q_{ii}$	0	0	<i>FM4P3 – VB</i>	0.156
	$Q_{ij}$	0	0	<i>FM4P3 – VB</i>	0.058
	$Q_{ik}$	<i>FM4P3 – VB</i>	0	0	0.490

Table G.2: Independent t-Tests between the two conversation scenarios Survival Task and Name Guessing Game, computed per condition for the audio-visual conversation test AVCT<sub>2</sub>.

1. Visualization of Results and Expectations

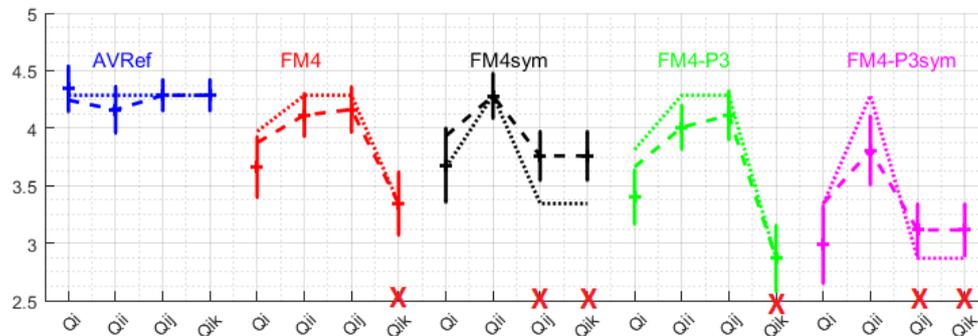


2. Detailed Analysis

AVRef	<p>All expected values lie within confidence intervals of <math>Q_i</math>, <math>Q_{ii}</math>, &amp; <math>Q_{ij}=Q_{ik}</math>, and mean value of <math>Q_i=4.1471</math> is maximum value of whole test and all observed mean values are not significantly different from maximum value of whole test (<math>Q_{ij}=4.3429</math> for VF2).  <math>\Rightarrow</math> AVRef is confirmed as reference condition for this test.</p>
VF2	<p>Expectations: No impairment on own connection, asymmetric condition concerning other connections.  <math>\Rightarrow</math> <math>Q_{ik}</math> determines lowest value, <math>Q_{ii}</math> &amp; <math>Q_{ij}</math> should be equal to reference value <math>Q_{ik}</math> of AVRef.</p> <p>Testing Hypothesis H1 (dotted line): <math>Q_{ii}</math> &amp; <math>Q_{ij}</math> are not significantly different (<math>Q_{ii}</math>: <math>p=0.418</math>, <math>Q_{ij}</math>: <math>p=0.426</math>) from the expected values.  <math>\Rightarrow</math> No support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): <math>Q_i</math> is not significantly different from the expected value (<math>p=0.122</math>).  <math>\Rightarrow</math> Support for H2.</p>
VF2sym	<p>Expectations: No impairment on own connection, symmetric condition concerning other connections.  <math>\Rightarrow</math> <math>Q_{ij} = Q_{ik}</math> should be equal to corresponding value of asymmetric condition (<math>Q_{ik}</math> of VF2), <math>Q_{ii}</math> should be equal to reference value (<math>Q_{ik}</math> of VF2).</p> <p>Testing Hypothesis H1 (dotted line): <math>Q_{ii}</math> &amp; <math>Q_{ij}</math> are not significantly different (<math>Q_{ii}</math>: <math>p=0.076</math>, <math>Q_{ij}</math>: <math>p=0.164</math>) from the expected values.  <math>\Rightarrow</math> No support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): <math>Q_i</math> is not significantly different from the expected value (<math>p=0.386</math>).  <math>\Rightarrow</math> Support for H2.</p>
VB	<p>Expectations: No impairment on own connection, asymmetric condition concerning other connections.  <math>\Rightarrow</math> <math>Q_{ik}</math> determines lowest value, <math>Q_{ii}</math> &amp; <math>Q_{ij}</math> should be equal to reference value <math>Q_{ik}</math> of AVRef.</p> <p>Testing Hypothesis H1 (dotted line): <math>Q_{ii}</math> &amp; <math>Q_{ij}</math> are not significantly different (<math>Q_{ii}</math>: <math>p=0.600</math>, <math>Q_{ij}</math>: <math>p=0.858</math>) from the expected values.  <math>\Rightarrow</math> No support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): <math>Q_i</math> is significantly different from the expected value (<math>p=0.000</math>).  <math>\Rightarrow</math> No support for H2.</p>
VBsym	<p>Expectations: No impairment on own connection, symmetric condition concerning other connections.  <math>\Rightarrow</math> <math>Q_{ij} = Q_{ik}</math> should be equal to corresponding value of asymmetric condition (<math>Q_{ik}</math> of VF2), <math>Q_{ii}</math> should be equal to reference value (<math>Q_{ik}</math> of VF2).</p> <p>Testing Hypothesis H1 (dotted line): <math>Q_{ii}</math> is not significantly different (<math>p=0.270</math>) from the expected value, <math>Q_{ij}</math> is significantly different (<math>p=0.000</math>) from the expected value.  <math>\Rightarrow</math> Support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): <math>Q_i</math> is significantly different from the expected value (<math>p=0.001</math>).  <math>\Rightarrow</math> No support for H2.</p>

Figure G.3: Detailed results for Experiment AVCT2, Part 1.

## 1. Visualization of Results and Expectations



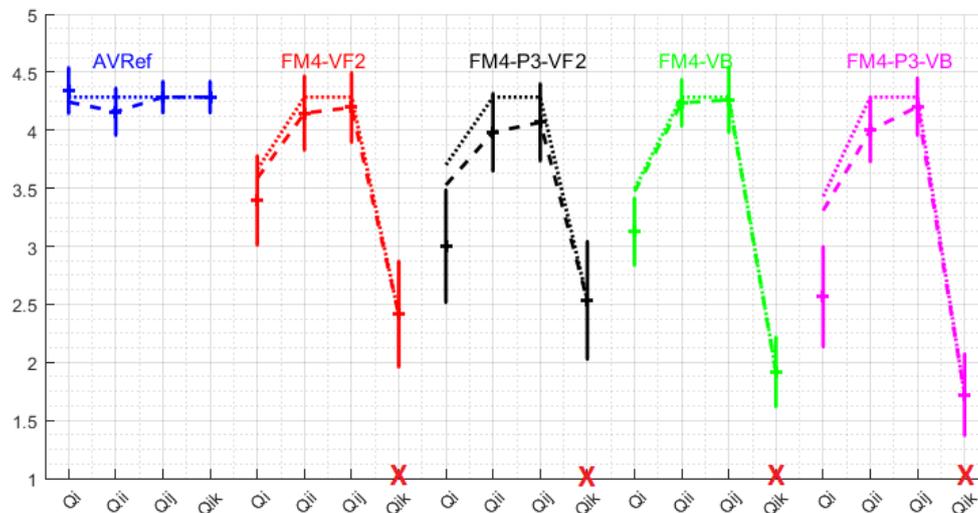
Note: The reference condition AVRef is repeated in this plot for better visual comparison.

## 2. Detailed Analysis

FM4	<p>Expectations: No impairment on own connection, asymmetric condition concerning other connections.  <math>\Rightarrow</math> <math>Q_{ik}</math> determines lowest value, <math>Q_{ii}</math> &amp; <math>Q_{ij}</math> should be equal to reference value <math>Q_{ik}</math> of AVRef.</p> <p>Testing Hypothesis H1 (dotted line): <math>Q_{ii}</math> &amp; <math>Q_{ij}</math> are not significantly different (<math>Q_{ii}</math>: <math>p=0.066</math>, <math>Q_{ij}</math>: <math>p=0.206</math>) from the expected values.  <math>\Rightarrow</math> No support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): <math>Q_i</math> is not significantly different from the expected value (<math>p=0.107</math>).  <math>\Rightarrow</math> Support for H2.</p>
FM4sym	<p>Expectations: No impairment on own connection, symmetric condition concerning other connections.  <math>\Rightarrow</math> <math>Q_{ij} = Q_{ik}</math> should be equal to corresponding value of asymmetric condition (<math>Q_{ik}</math> of VF2), <math>Q_{ii}</math> should be equal to reference value (<math>Q_{ik}</math> of VF2).</p> <p>Testing Hypothesis H1 (dotted line): <math>Q_{ii}</math> is not significantly different (<math>p=0.956</math>) from the expected value, <math>Q_{ij}</math> is significantly different (<math>p=0.000</math>) from the expected value.  <math>\Rightarrow</math> No support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): <math>Q_i</math> is not significantly different from the expected value (<math>p=0.114</math>).  <math>\Rightarrow</math> Support for H2.</p>
FM4-P3	<p>Expectations: No impairment on own connection, asymmetric condition concerning other connections.  <math>\Rightarrow</math> <math>Q_{ik}</math> determines lowest value, <math>Q_{ii}</math> &amp; <math>Q_{ij}</math> should be equal to reference value <math>Q_{ik}</math> of AVRef.</p> <p>Testing Hypothesis H1 (dotted line): <math>Q_{ii}</math> is significantly different (<math>p=0.005</math>) from the expected value, <math>Q_{ij}</math> is not significantly different (<math>p=0.102</math>) from the expected value.  <math>\Rightarrow</math> Support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): <math>Q_i</math> is significantly different from the expected value (<math>p=0.028</math>).  <math>\Rightarrow</math> No support for H2.</p>
FM4-P3sym	<p>Expectations: No impairment on own connection, symmetric condition concerning other connections.  <math>\Rightarrow</math> <math>Q_{ij} = Q_{ik}</math> should be equal to corresponding value of asymmetric condition (<math>Q_{ik}</math> of VF2), <math>Q_{ii}</math> should be equal to reference value (<math>Q_{ik}</math> of VF2).</p> <p>Testing Hypothesis H1 (dotted line): <math>Q_{ii}</math> &amp; <math>Q_{ij}</math> are significantly different (<math>Q_{ii}</math>: <math>p=0.002</math>, <math>Q_{ij}</math>: <math>p=0.031</math>) from the expected values.  <math>\Rightarrow</math> Support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): <math>Q_i</math> is significantly different from the expected value (<math>p=0.036</math>).  <math>\Rightarrow</math> No support for H2.</p>

Figure G.4: Detailed results for Experiment AVCT2, Part 2.

1. Visualization of Results and Expectations



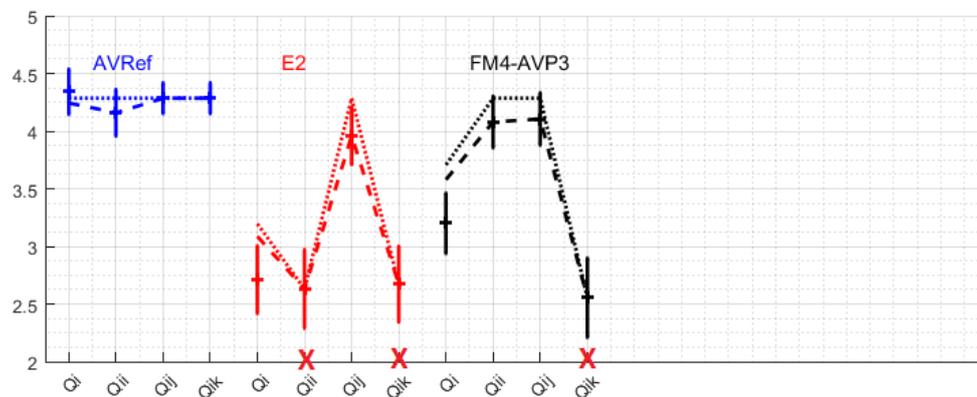
Note: The reference condition AVRef is repeated in this plot for better visual comparison.

2. Detailed Analysis

FM4-VF2	<p>Expectations: No impairment on own connection, asymmetric condition concerning other connections.                      ⇒ Qik determines lowest value, Qii &amp; Qij should be equal to reference value Qik of AVRef.</p> <p>Testing Hypothesis H1 (dotted line): Qii &amp; Qij are not significantly different (Qii: <math>p=0.372</math>, Qij: <math>p=0.534</math>) from the expected values.                      ⇒ No support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is not significantly different from the expected value (<math>p=0.307</math>).                      ⇒ Support for H2.</p>
FM4-P3-VF2	<p>Expectations: No impairment on own connection, asymmetric condition concerning other connections.                      ⇒ Qik determines lowest value, Qii &amp; Qij should be equal to reference value Qik of AVRef.</p> <p>Testing Hypothesis H1 (dotted line): Qii &amp; Qij are not significantly different (Qii: <math>p=0.070</math>, Qij: <math>p=0.180</math>) from the expected values.                      ⇒ No support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is significantly different from the expected value (<math>p=0.033</math>).                      ⇒ No support for H2.</p>
FM4-VB	<p>Expectations: No impairment on own connection, asymmetric condition concerning other connections.                      ⇒ Qik determines lowest value, Qii &amp; Qij should be equal to reference value Qik of AVRef.</p> <p>Testing Hypothesis H1 (dotted line): Qii &amp; Qij are not significantly different (Qii: <math>p=0.608</math>, Qij: <math>p=0.861</math>) from the expected values.                      ⇒ No support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is significantly different from the expected value (<math>p=0.020</math>).                      ⇒ No support for H2.</p>
FM4-P3-VB	<p>Expectations: No impairment on own connection, asymmetric condition concerning other connections.                      ⇒ Qik determines lowest value, Qii &amp; Qij should be equal to reference value Qik of AVRef.</p> <p>Testing Hypothesis H1 (dotted line): Qii is significantly different (<math>p=0.037</math>) from the expected value, Qij is not significantly different (<math>p=0.490</math>) from the expected value.                      ⇒ No support for H1.</p> <p>Testing Hypothesis H2 (dashed line, since this is independent from H1): Qi is significantly different from the expected value (<math>p=0.002</math>).                      ⇒ No support for H2.</p>

Figure G.5: Detailed results for Experiment AVCT2, Part 3.

## 1. Visualization of Results and Expectations



Note: The reference condition AVRef is repeated in this plot for better visual comparison.

## 2. Detailed Analysis

- E2**      Expectations: Impairment on own connection, asymmetric condition concerning other connections.  
 ⇒  $Q_{ii}$  &  $Q_{ik}$  are lower than reference value  $Q_{ik}$  of Ref<sub>1</sub>,  $Q_{ij}$  should be equal to reference value ( $Q_{ik}$  of Ref<sub>1</sub>).
- Testing Hypothesis H<sub>1</sub> (dotted line):  $Q_{ij}$  is significantly different from the expected value ( $p=0.011$ ).  
 ⇒ Support for H<sub>1</sub>.
- Testing Hypothesis H<sub>2</sub> (dashed line, since this is independent from H<sub>1</sub>):  $Q_i$  is significantly different from the expected value ( $p=0.013$ ).  
 ⇒ No support for H<sub>2</sub>.
- FM4-AVP<sub>3</sub>**      Expectations: No impairment on own connection, asymmetric condition concerning other connections.  
 ⇒  $Q_{ik}$  determines lowest value,  $Q_{ii}$  &  $Q_{ij}$  should be equal to reference value  $Q_{ik}$  of AVRef.
- Testing Hypothesis H<sub>1</sub> (dotted line):  $Q_{ii}$  &  $Q_{ij}$  are not significantly different ( $Q_{ii}$ :  $p=0.067$ ,  $Q_{ij}$ :  $p=0.114$ ) from the expected values.  
 ⇒ No support for H<sub>1</sub>.
- Testing Hypothesis H<sub>2</sub> (dashed line, since this is independent from H<sub>1</sub>):  $Q_i$  is significantly different from the expected value ( $p=0.006$ ).  
 ⇒ No support for H<sub>2</sub>.

Figure G.6: Detailed results for Experiment AVCT<sub>2</sub>, Part 4.

# *H*

## *Detailed Results: Model Performance*

### *What this chapter is about*

The present appendix presents the detailed results of the modeling experiments according to Chapter 9. The modeling experiments were conducted individually for each data set obtained from the seven experiments ACT<sub>1</sub>, ACT<sub>2</sub>, ACT<sub>3</sub>, LOT<sub>1</sub>, LOT<sub>2</sub>, AVCT<sub>1</sub> & AVCT<sub>2</sub> on the impact of technical conditions on perceived telemeeting quality.

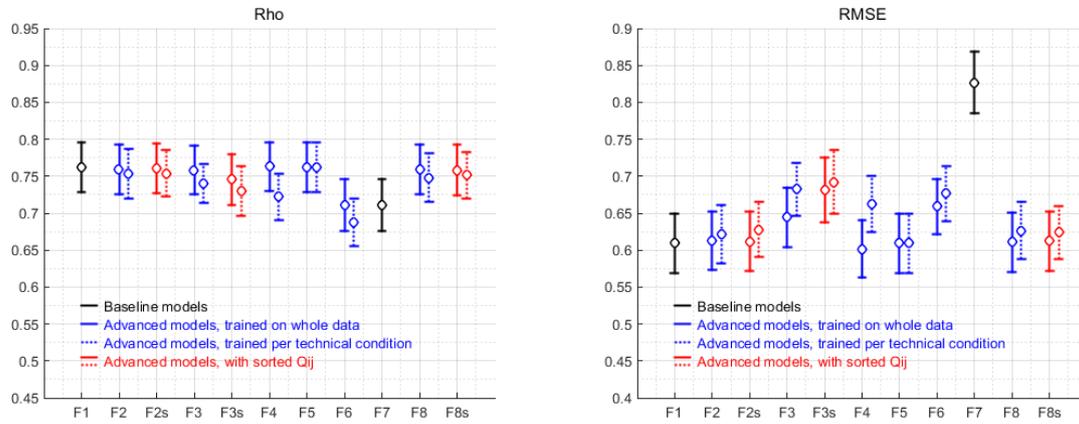
In each section – one section per experiment – the first figure shows the detailed model performance for the whole data set of that experiment. Then, the following figures in that section show the detailed model performance per technical condition.

Each individual figure visualizes first the results of the two employed performance measures Rho and RMSE for the different models. Then, each figure provides explanations of the conducted statistical analysis and the corresponding results. Finally, each figure presents the main conclusions that can be extracted from the current results. The individual results are then compiled in the main text in Chapter 9.

## H.1 Experiment ACT<sub>1</sub>

### H.1.1 Performance across technical conditions

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



**2.1 Testing for significant differences between functions F<sub>1</sub> - F<sub>8s</sub>:** Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
Rho	per condition	$p = 0.013$	$p = 0.041$	F6 vs. F1, F5
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. all other
RMSE	per condition	$p = 0.000$	$p = 0.000$	F7 vs. all other

**2.2 Testing for differences between two training modes:** T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

RMSE	F <sub>4</sub> : $p = 0.023$
------	------------------------------

### 3. Behavior of modeling functions

Measure	Baseline (F <sub>1</sub> & F <sub>7</sub> )	Advanced (F <sub>2</sub> -F <sub>6</sub> , F <sub>8</sub> , F <sub>8s</sub> )
Rho	F <sub>1</sub> slightly better than F <sub>7</sub> (not sig.)	No one outperforms F <sub>1</sub> .
RMSE	F <sub>1</sub> sig. better than F <sub>7</sub> .	No one outperforms F <sub>1</sub> .

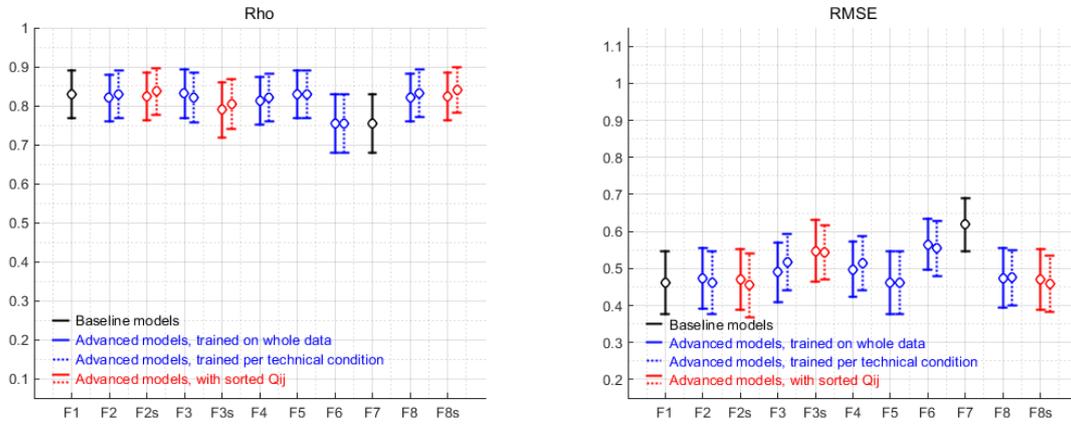
### 4. Observations and Conclusions

- No advanced model outperforms the baseline mean.
- Equal performance for sorting-based (F<sub>2s</sub>, F<sub>3s</sub>, F<sub>8s</sub>) and rule-based (F<sub>2</sub>, F<sub>3</sub>, F<sub>8</sub>) assignment of  $Q_{ij}$ .
- Training per condition does not improve performance, in case of RMSE for F<sub>4</sub> it even reduces performance.

Figure H.1: Modeling performance across all technical conditions of Experiment ACT<sub>1</sub>.

H.1.2 Performance per technical condition

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	per condition	$p = 0.032$	$p = 0.115$ (n.s.)	F2s vs. F7

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

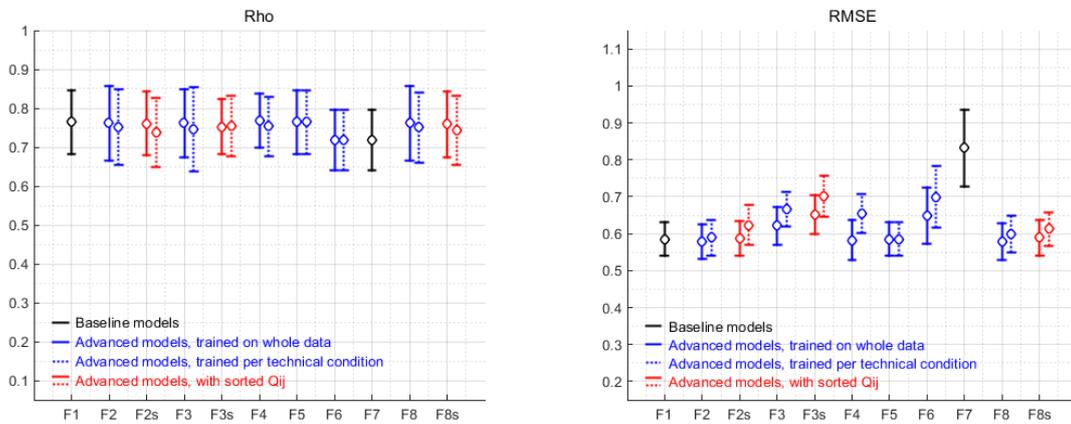
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho, RMSE	F1 slightly better than F7 (not sig.)	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition Ref1

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.001$	F7 vs. F3s
RMSE	per condition	$p = 0.000$	$p = 0.000$	F7 vs. all other F8s
			$p = 0.001$	F7 vs. F4
			$p = 0.004$	F7 vs. F3

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Slight (not sig.) differences found for: RMSE F4:  $p = 0.056$

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly better than F7 (not sig.)	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

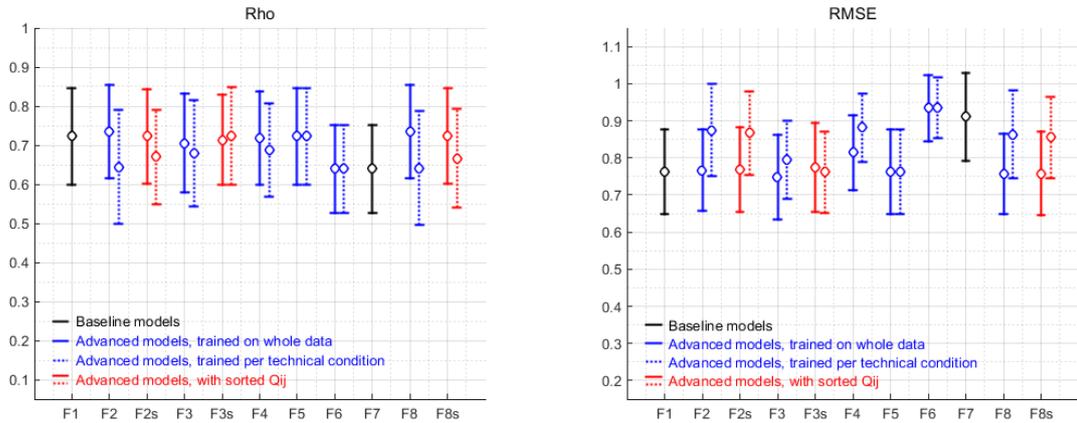
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Assignment of  $Q_{ij}$  to weights of F2, F3, F8 based on sorting of  $Q_{ij}$  shows equal performance than rule-based assignment.
- C) Training per condition does not improve performance.

Condition L3

Figure H.2: Modeling performance for conditions Ref1 (top panel) and L3 (bottom panel) of Experiment ACT1.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

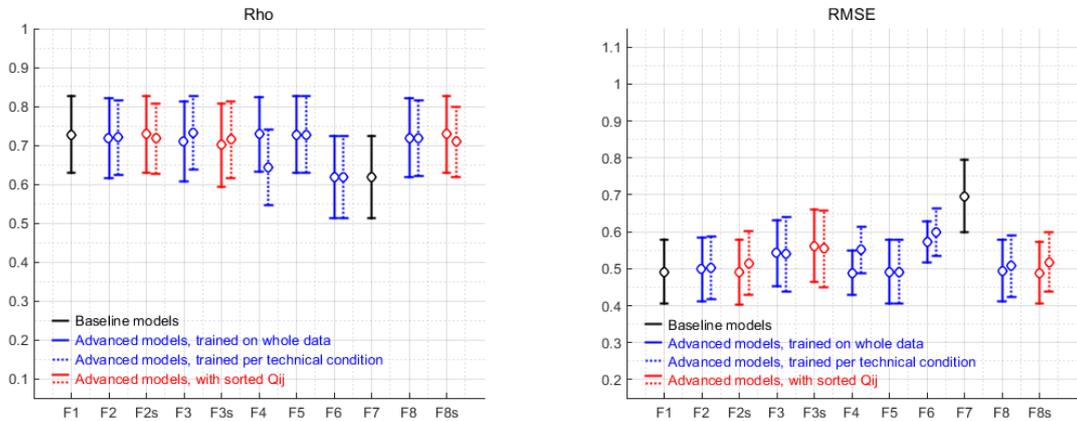
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho, RMSE	F1 slightly better than F7 (not sig.)	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition EN3

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.007$	$p = 0.017$	F7 vs. F4, F8s
			$p = 0.020$	F7 vs. F2s
			$p = 0.022$	F7 vs. F1, F5
			$p = 0.026$	F7 vs. F8
			$p = 0.033$	F7 vs. F2
RMSE	per condition	$p = 0.024$	$p = 0.031$	F7 vs. F1, F5

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly better than F7 (not sig.)	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

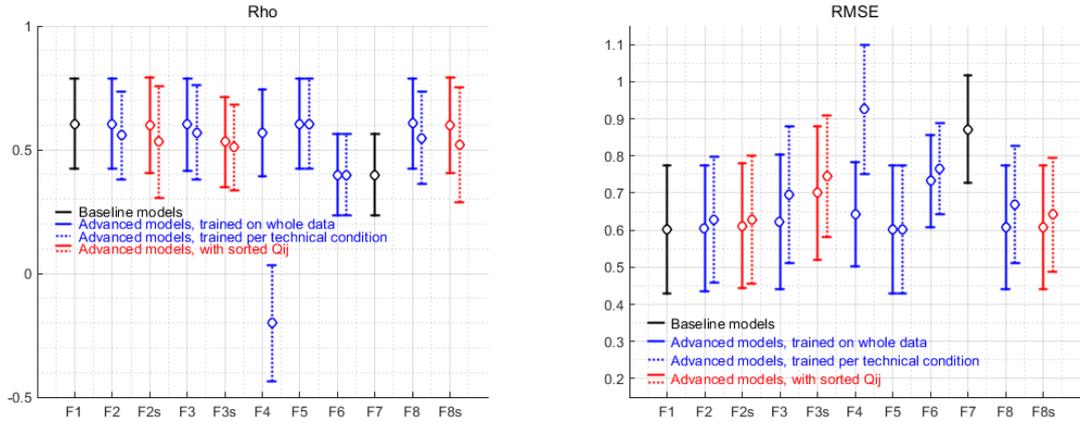
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition TN3

Figure H.3: Modeling performance for conditions EN3 (top panel) and TN3 (bottom panel) of Experiment ACT1.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
Rho	per condition	$p = 0.000$	$p = 0.000$	F4 vs. all other
RMSE	per condition	$p = 0.038$	$p = 0.186$ (n.s.)	F4 vs. F1, F5

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

Measure	T-test	Functions	Measure	T-test	Functions
Rho	$p = 0.000$	F4	RMSE	0.012	F4

3. Behavior of modeling functions

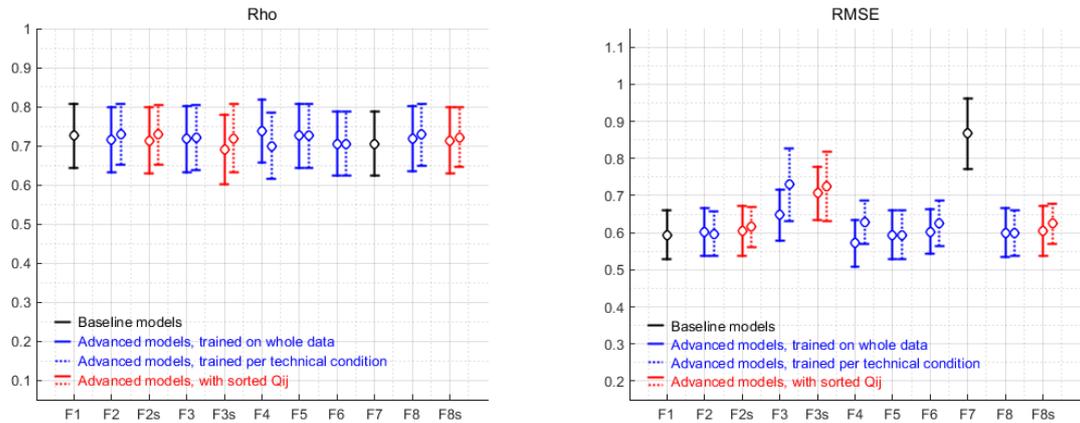
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho, RMSE	F1 slightly better than F7 (not sig.)	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance, in case of F4 it even reduces performance substantially.

Condition Dsym

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.036$	F7 vs. F3s
			$p = 0.000$	F7 vs. all other
RMSE	per condition	$p = 0.000$	$p = 0.000$	F7 vs. F1, F2, F2s, F4, F5, F6, F8, F8s

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

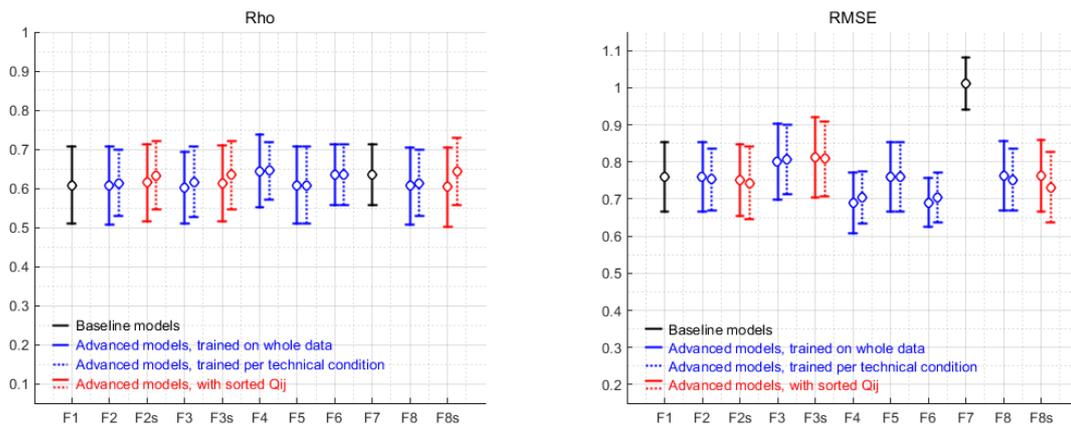
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition E1

Figure H.4: Modeling performance for conditions Dsym (top panel) and E1 (bottom panel) of Experiment ACT1.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. F4, F6
			$p = 0.002$	F7 vs. F2s
			$p = 0.004$	F7 vs. F1, F2, F5, F8
			$p = 0.005$	F7 vs. F8s
			$p = 0.046$	F7 vs. F3
RMSE	per condition	$p = 0.000$	$p = 0.000$	F7 vs. F4, F6
			$p = 0.001$	F7 vs. F2, F2s, F8, F8s
			$p = 0.002$	F7 vs. F1, F5
			$p = 0.034$	F7 vs. F3
			$p = 0.040$	F7 vs. F3s

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

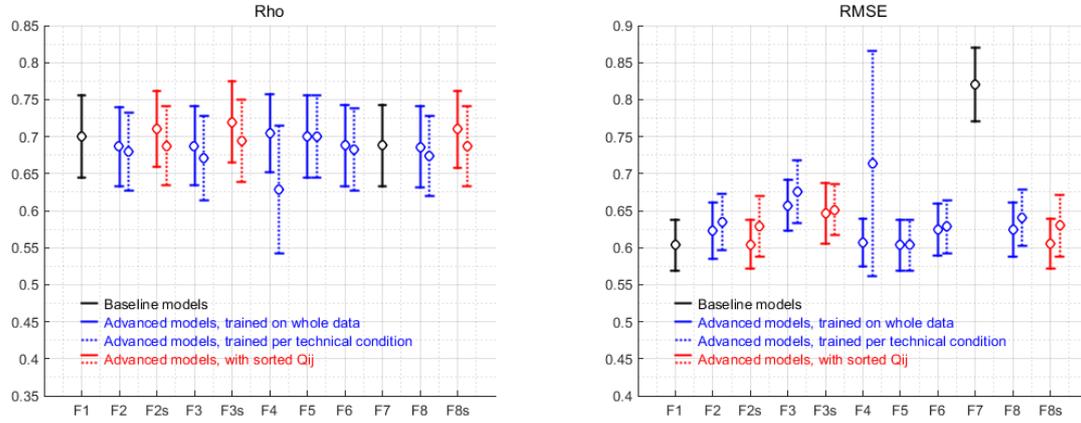
Condition P1

Figure H.5: Modeling performance for condition P1 of Experiment ACT1.

## H.2 Experiment ACT2

### H.2.1 Performance across technical conditions

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. all other
RMSE	per condition	$p = 0.000$	$p = 0.000$	F7 vs. F1, F2, F2s, F5, F6, F8s
			$p = 0.001$	F7 vs. F8
			$p = 0.002$	F7 vs. F3s
			$p = 0.020$	F7 vs. F3

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

### 3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

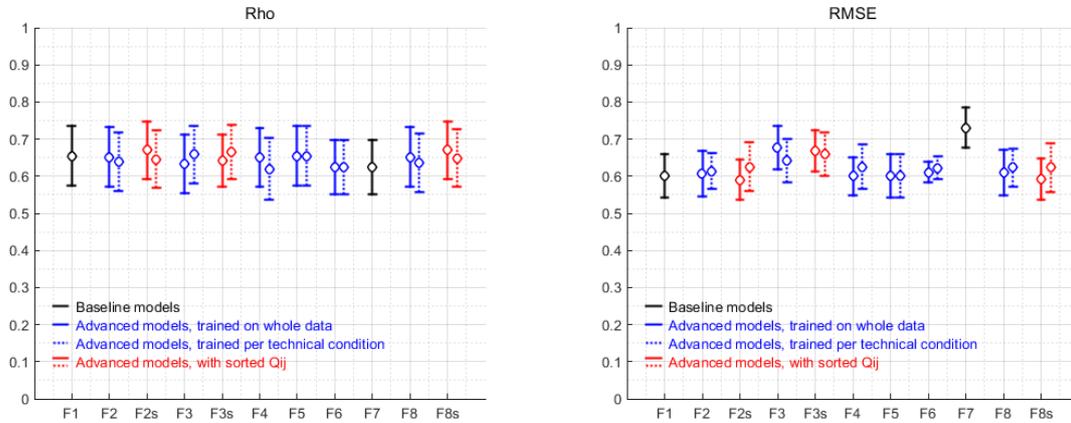
### 4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance, in case of F4 it leads to reduced and instable (large confidence intervals) performance.

Figure H.6: Modeling performance across all technical conditions of Experiment ACT2.

H.2.2 Performance per technical condition

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.002$	$p = 0.011$ $p = 0.013$ $p = 0.028$	F7 vs. F2s F7 vs. F8s F7 vs. F4
RMSE	per condition	$p = 0.054$ (n.s.)	$p = 0.031$ $p = 0.045$	F7 vs. F1, F5 F7 vs. F1, F5

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

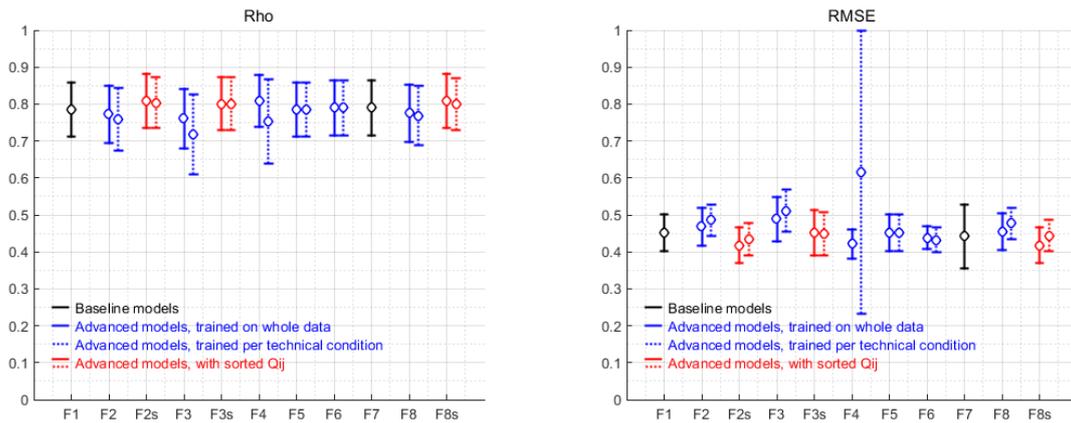
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition Ref2

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho, RMSE	F1 essentially equal to F7.	No one outperforms F1.

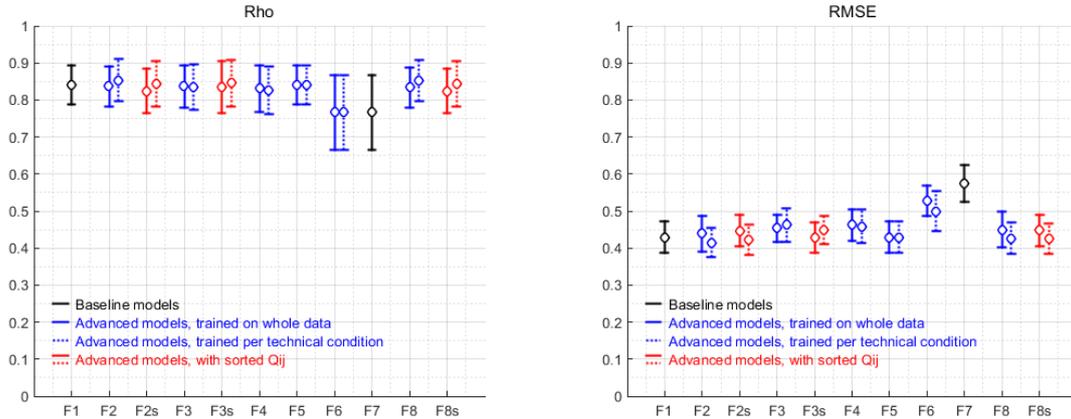
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance, in case of RMSE for F4 it even leads to reduced and instable (large confidence interval) performance.

Condition L1

Figure H.7: Modeling performance for conditions Ref2 (top panel) and L1 (bottom panel) of Experiment ACT2.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. F1, F2, F3s, F5
				F7 vs. F2s, F8s
				F7 vs. F8
				F7 vs. F3
RMSE	per condition	$p = 0.000$	$p = 0.000$	F7 vs. F1, F2, F2s, F5, F8, F8s
				F7 vs. F3s
				F7 vs. F4
				F7 vs. F3

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

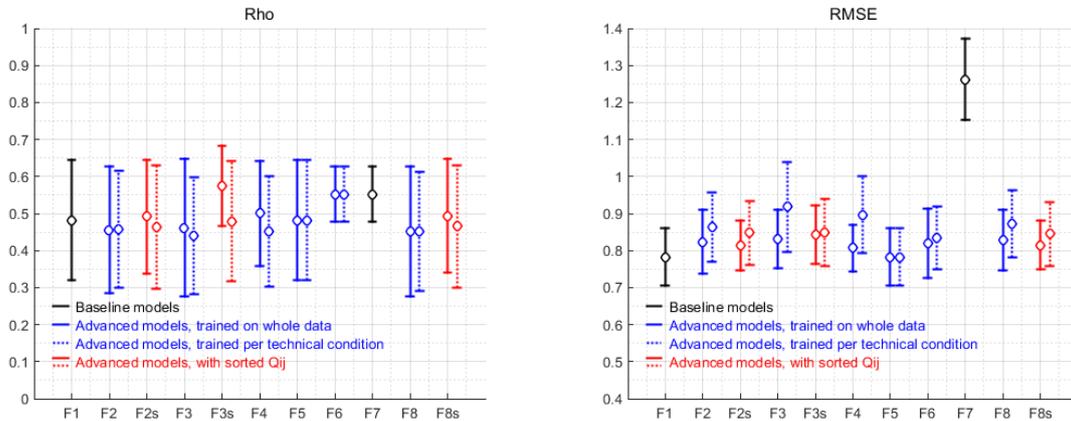
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly better than F7 (not sig.)	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition L2

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. all other
RMSE	per condition	$p = 0.000$	$p = 0.000$	F7 vs. all other

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly worse than F7 (not sig.)	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

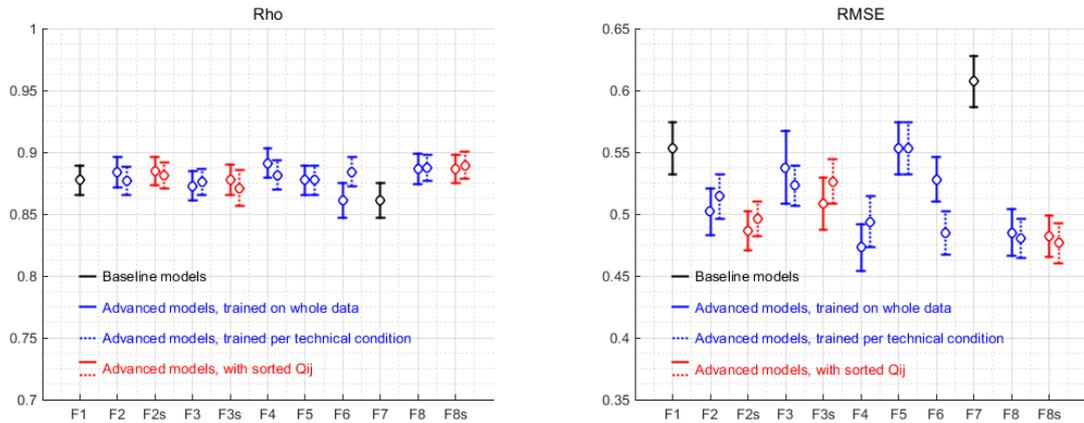
Condition L4

Figure H.8: Modeling performance for conditions L2 (top panel) and L4 (bottom panel) of Experiment ACT2.

### H.3 Experiment ACT<sub>3</sub>

#### H.3.1 Performance across technical conditions

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
Rho	whole data	$p = 0.003$	$p = 0.024$	F4 vs. F6, F7
Rho	per condition	$p = 0.050$ (n.s.)	$p = 0.033$	F7 vs. F8s
RMSE	whole data	$p = 0.000$	$p = 0.000$	F1, F5 vs. F2s, F4, F8, F8s
			$p = 0.005$	F1, F5 vs. F7
			$p = 0.012$	F1, F5 vs. F2
			$p = 0.000$	F7 vs. F2, F2s, F3, F3s, F4, F6, F8, F8s
			$p = 0.000$	F3 vs. F4
			$p = 0.004$	F3 vs. F8s
			$p = 0.009$	F3 vs. F8
			$p = 0.013$	F3 vs. F2s
			$p = 0.000$	F6 vs. F4
			$p = 0.000$	F6 vs. F8s
RMSE	per condition	$p = 0.000$	$p = 0.000$	F1, F5 vs. F2s, F4, F6, F8, F8s
			$p = 0.001$	F1, F5 vs. F7
			$p = 0.000$	F7 vs. F2, F2s, F3, F3s, F4, F6, F8, F8s
			$p = 0.013$	F8 vs. F3s
			$p = 0.035$	F8 vs. F3
			$p = 0.004$	F8s vs. F3s
			$p = 0.011$	F8s vs. F3

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

Measure	T-test	Functions	Measure	T-test	Functions
Rho	$p = 0.014$	F6	RMSE	$p = 0.001$	F6

#### 3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly better than F7 (not sig.)	No one outperforms F1.
RMSE	F1 sig. better than F7.	F2s, F4, F8, F8s & F2 (for training on whole data) & F6 (for training per condition) sig. outperform F1.

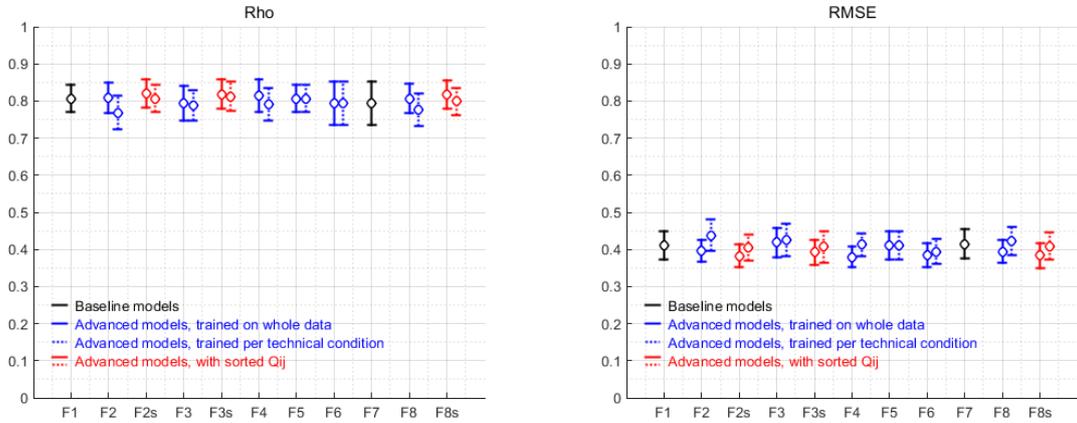
#### 4. Observations and Conclusions

- A) Six advanced models outperform the baseline mean. F4 is the best, though not sig. different from the other five.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition improves performance only for F6.

Figure H.9: Modeling performance across all technical conditions of Experiment ACT<sub>3</sub>.

H.3.2 Performance per technical condition

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

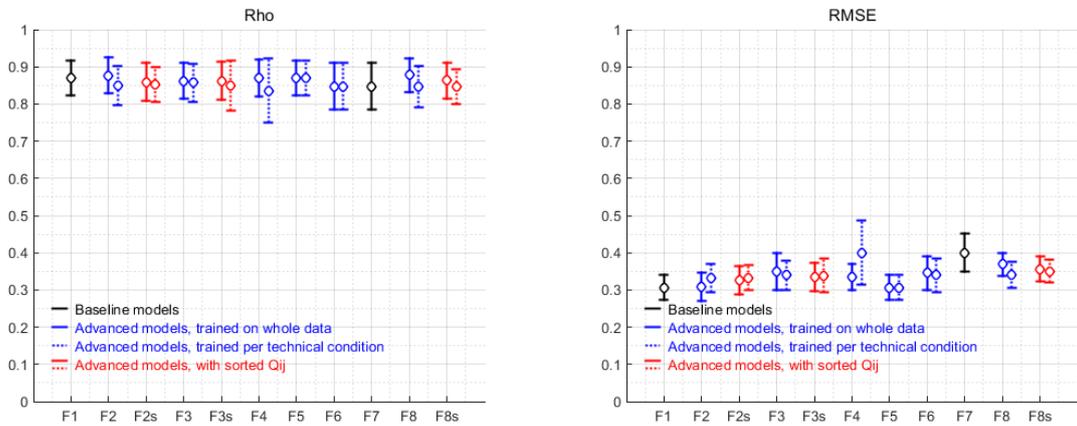
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho, RMSE	F1 essentially equal to F7.	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition Ref3

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.012$	$p = 0.028$	F7 vs. F1, F5
			$p = 0.041$	F7 vs. F2
RMSE	per condition	$p = 0.031$	$p = 0.127$ (n.s.)	F7 vs. F1, F5
			$p = 0.128$ (n.s.)	F4 vs. F1, F5

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

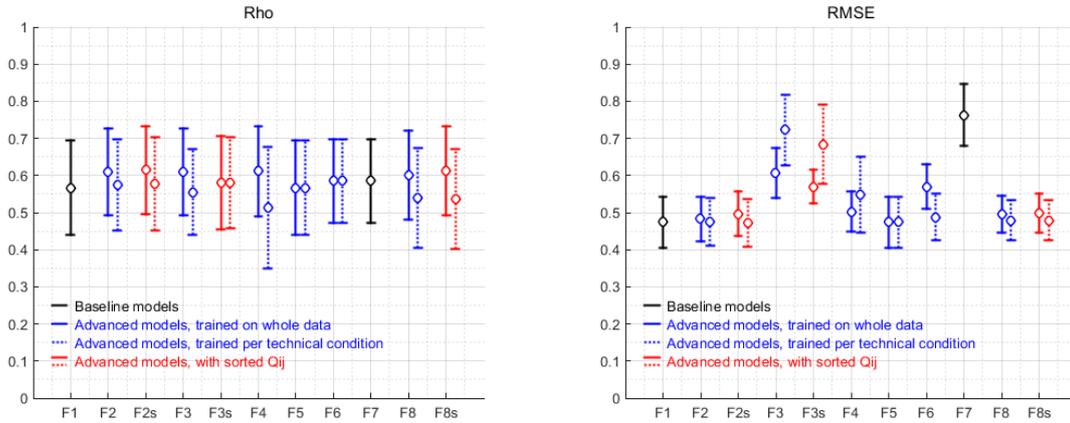
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition B1

Figure H.10: Modeling performance for conditions Ref3 (top panel) and B1 (bottom panel) of Experiment ACT3.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. F1, F2, F2s, F3s, F4, F5, F6, F8, F8s
			$p = 0.014$	F7 vs. F3
RMSE	per condition	$p = 0.000$	$p = 0.000$	F3 vs. F1, F2, F2s, F5, F8, F8s
			$p = 0.001$	F3 vs. F6
			$p = 0.049$	F3 vs. F4
			$p = 0.004$	F3s vs. F2s
			$p = 0.005$	F3s vs. F1, F2, F5
			$p = 0.006$	F3s vs. F8
			$p = 0.007$	F3s vs. F8s
			$p = 0.013$	F3s vs. F6
			$p = 0.000$	F7 vs. F1, F2, F2s, F5, F6, F8, F8s
			$p = 0.003$	F7 vs. F4

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

Measure	T-test	Functions	Measure	T-test	Functions
RMSE	$p = 0.043$	F3	RMSE	$p = 0.057$ (n.s.)	F6

3. Behavior of modeling functions

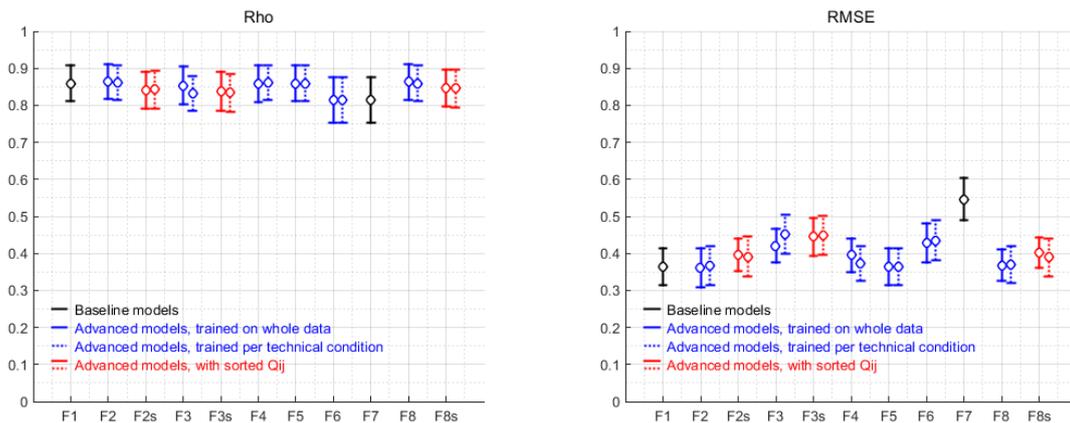
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance, in case of RMSE it even reduces performance of F3.

Condition B2

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. F1, F2, F2s, F4, F5, F8
			$p = 0.001$	F7 vs. F8s
			$p = 0.009$	F7 vs. F3
			$p = 0.022$	F7 vs. F6
RMSE	per condition	$p = 0.000$	$p = 0.000$	F7 vs. F1, F2, F4, F5, F8
			$p = 0.001$	F7 vs. F2s, F8s

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

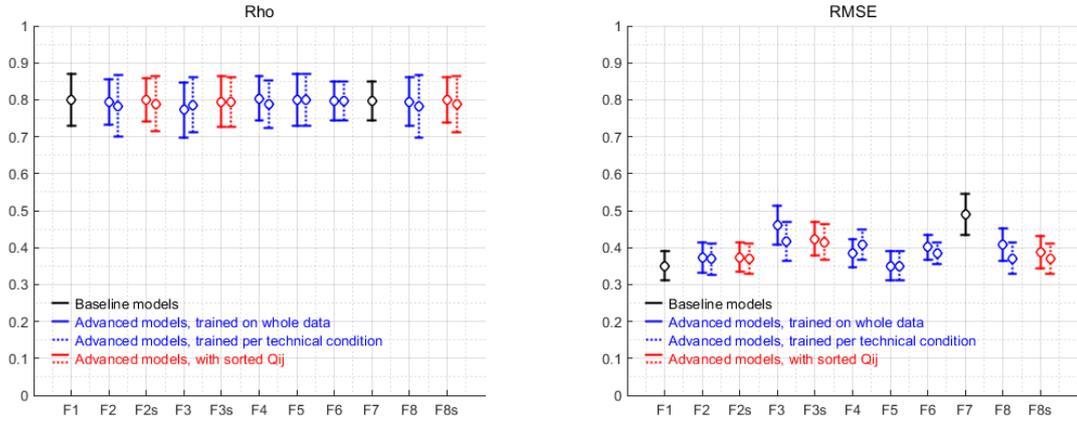
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition B2sym

Figure H.11: Modeling performance for conditions B2 (top panel) and B2sym (bottom panel) of Experiment ACT3.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.014$	F3 vs. F1, F5
			$p = 0.000$	F7 vs. F1, F5
			$p = 0.005$	F7 vs. F2
			$p = 0.006$	F7 vs. F2s
			$p = 0.020$	f7 vs. F4
RMSE	per condition	$p = 0.000$	$p = 0.035$	F7 vs. F8s
			$p = 0.000$	F7 vs. F1, F5
			$p = 0.003$	F7 vs. F2, F2s, F8
			$p = 0.004$	f7 vs. F8
			$p = 0.024$	F7 vs. F6

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

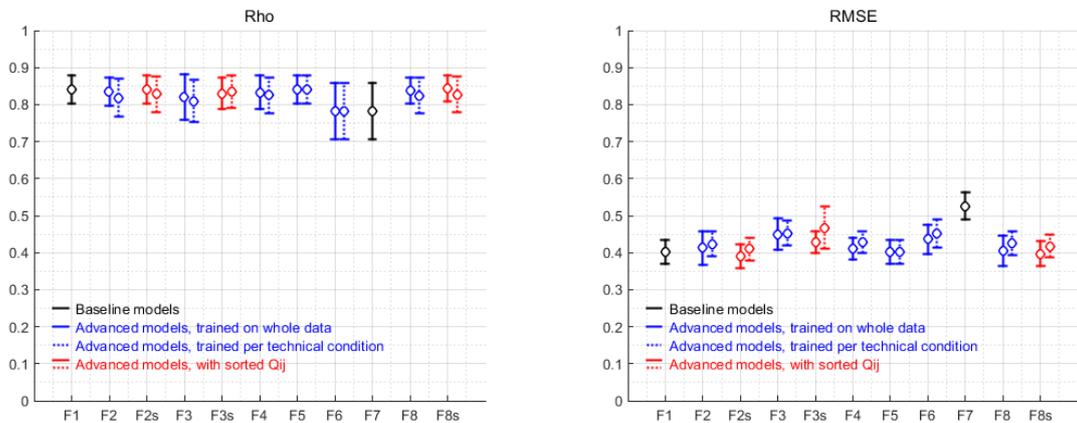
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition FM1

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. F1, F2, F2s, F4, F5, F8, F8s
			$p = 0.005$	F7 vs. F3s
			$p = 0.017$	F7 vs. F6
			$p = 0.000$	F7 vs. F1, F2s, F5
			$p = 0.001$	F7 vs. F8s
RMSE	per condition	$p = 0.000$	$p = 0.002$	F7 vs. F2
			$p = 0.003$	F7 vs. F8
			$p = 0.004$	F7 vs. F4

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly better than F7 (not sig.)	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

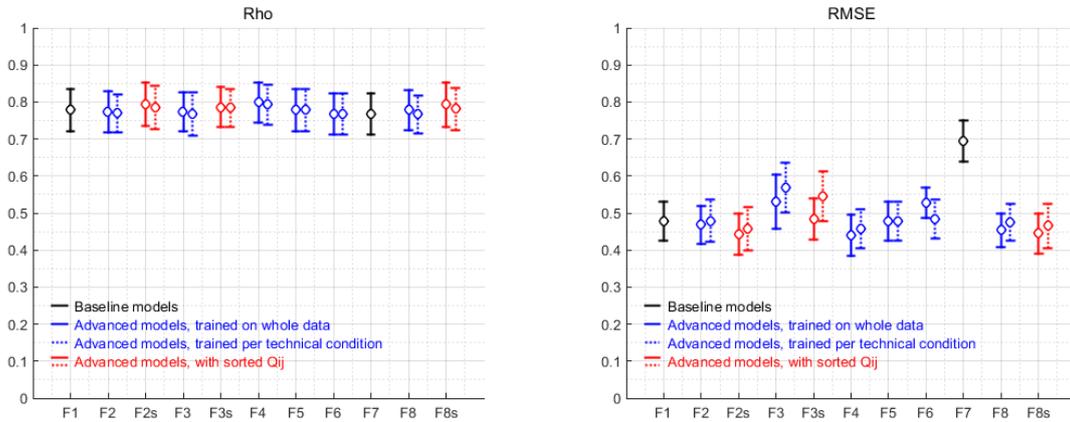
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition FL3sym

Figure H.12: Modeling performance for conditions FM1 (top panel) and FL3sym (bottom panel) of Experiment ACT3.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. F1, F2, F2s, F3s, F4, F5, F8, F8s
				$p = 0.001$
RMSE	per condition	$p = 0.000$	$p = 0.000$	F7 vs. F1, F2, F2s, F4, F, F6, F8, F8s
				$p = 0.008$

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

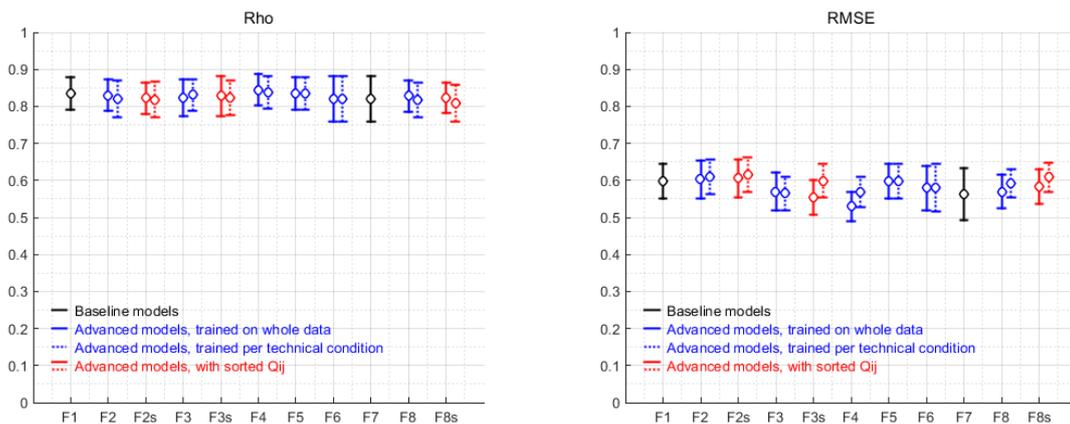
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition FM2

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

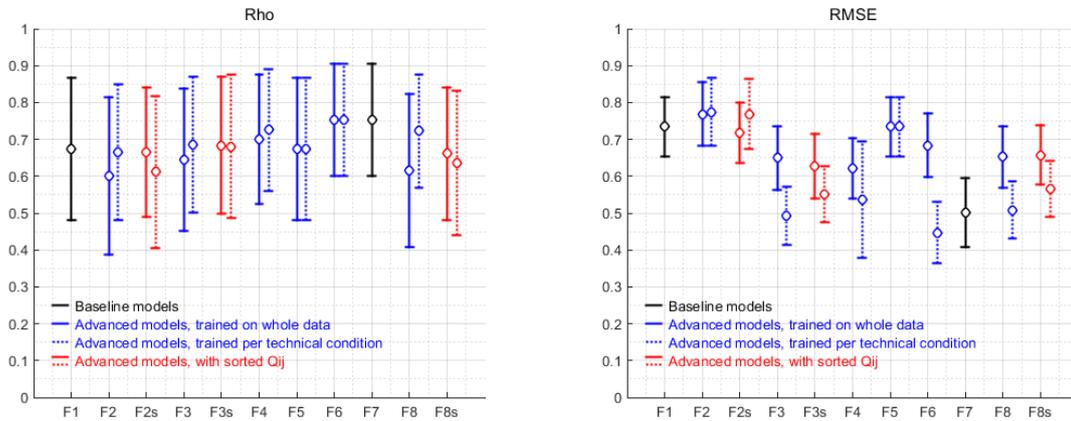
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition FL4sym

Figure H.13: Modeling performance for conditions FM2 (top panel) and FL4sym (bottom panel) of Experiment ACT3.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.001$	$p = 0.000$	F7 vs. F2
			$p = 0.003$	F7 vs. F1, F5
			$p = 0.010$	F7 vs. F2s
			$p = 0.000$	F1, F5 vs. F6
RMSE	per condition	$p = 0.000$	$p = 0.009$	F1, F5 vs. F3
			$p = 0.014$	F1, F5 vs. F7
			$p = 0.022$	F1, F5 vs. F8
			$p = 0.000$	F2 vs. F6
			$p = 0.001$	F2 vs. F3, F7
			$p = 0.002$	F2 vs. F8
			$p = 0.011$	F2 vs. F4
			$p = 0.025$	F2 vs. F3s
			$p = 0.000$	F2s vs. F6
			$p = 0.001$	F2s vs. F3
			$p = 0.002$	F2s vs. F7
			$p = 0.008$	F2s vs. F8
$p = 0.015$	F2s vs. F4			

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

Measure	T-test	Functions	Measure	T-test	Functions
RMSE	$p = 0.004$	F3	RMSE	$p = 0.000$	F6
RMSE	$p = 0.011$	F8			

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 sig. worse than F7.	F3, F6, & F8 (all trained per condition) outperform F1, but not F7.

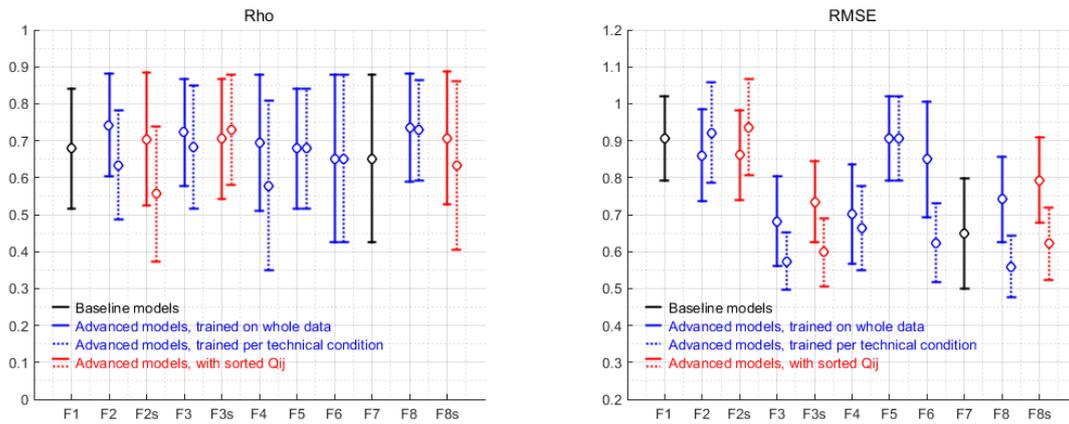
4. Observations and Conclusions

- A) Three advanced models (when trained per condition) outperform the baseline mean, but no one the baseline minimum.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition improves performance for F3, F6, & F8.

Condition TN1

Figure H.14: Modeling performance for condition TN1 of Experiment ACT3.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.012$	$p = 0.152$ (n.s.)	F7 vs. F1, F5
RMSE	per condition	$p = 0.000$	$p = 0.000$	F1, F5 vs. F8
			$p = 0.001$	F1, F5 vs. F3
			$p = 0.004$	F1, F5 vs. F3s
			$p = 0.013$	F1, F5 vs. F8s
			$p = 0.014$	F1, F5 vs. F6
			$p = 0.048$	F1, F5 vs. F7
			$p = 0.000$	F2 vs. F3, F8
			$p = 0.002$	F2 vs. F3s
			$p = 0.006$	F2 vs. F6, F8s
			$p = 0.023$	F2 vs. F7
			$p = 0.044$	F2 vs. F4
			$p = 0.000$	F2s vs. F3, F8
			$p = 0.001$	F2s vs. F3s
			$p = 0.003$	F2s vs. F6, F8s
			$p = 0.012$	F2s vs. F7
			$p = 0.024$	F2s vs. F4

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

Measure	T-test	Functions	Measure	T-test	Functions
RMSE	$p = 0.017$	F6	RMSE	$p = 0.011$	F8
RMSE	$p = 0.055$ (n.s.)	F3s	RMSE	$p = 0.024$	F8s

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 sig. worse than F7.	F3, F3s, F6, & F8 (all trained per condition) outperform F1, but not F7.

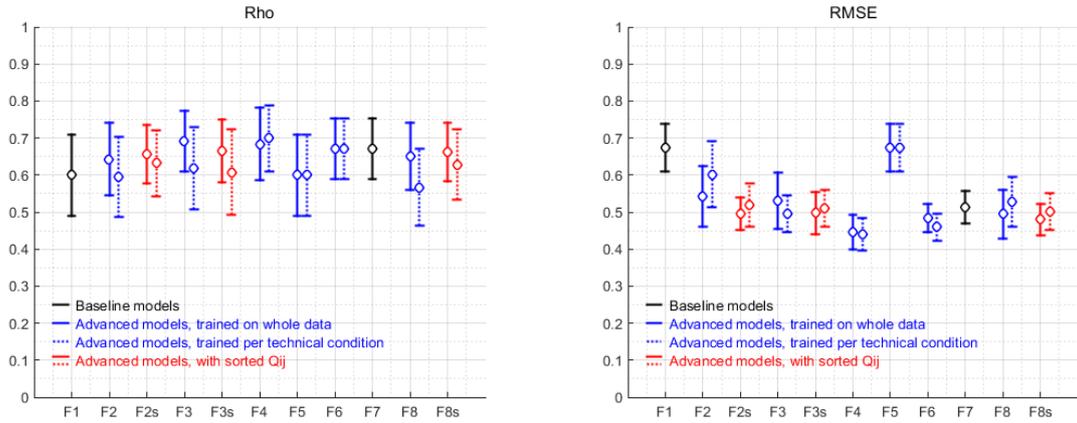
4. Observations and Conclusions

- A) Four advanced models (when trained per condition) outperform the baseline mean, but no one the baseline minimum.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition improves performance for F6, F8, & F8s.

Condition TN2

Figure H.15: Modeling performance for condition TN2 of Experiment ACT3.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F1, F5 vs. F4, F6, F8s $p = 0.001$ F1, F5 vs. F2s, F3s, F8 $p = 0.004$ F1, F5 vs. F7 $p = 0.020$ F1, F5 vs. F3
	per condition	$p = 0.000$	$p = 0.000$	F1, F5 vs. F4, F6 $p = 0.001$ F1, F5 vs. F3, F8s $p = 0.003$ F1, F5 vs. F3s, F7 $p = 0.006$ F1, F5 vs. F2s $p = 0.008$ F1, F5 vs. F8 $p = 0.003$ F2 vs. F4 $p = 0.020$ F2 vs. F6

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

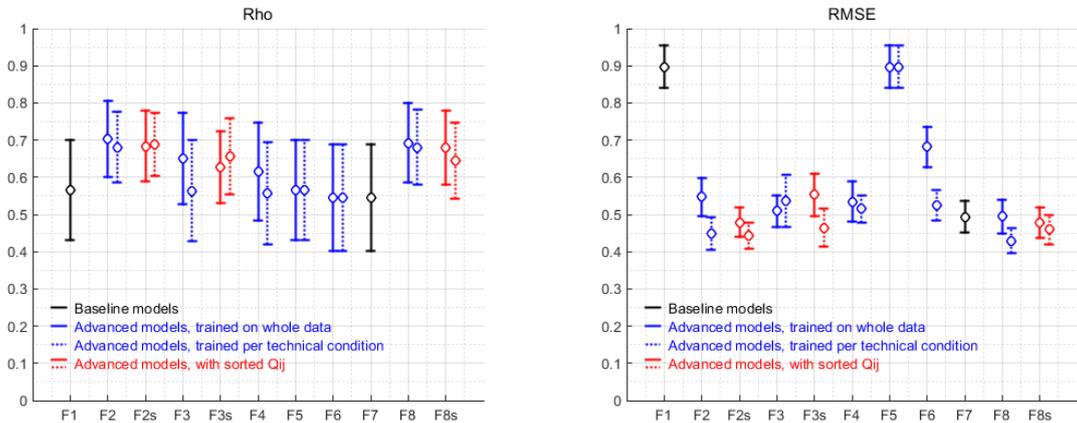
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly worse than F7 (not sig.)	No one outperforms F1.
RMSE	F1 sig. worse than F7.	All except F2 & F5 outperform F1, but not F7.

4. Observations and Conclusions

- A) Seven advanced models outperform the baseline mean, but not the baseline minimum. F4 is the best, though not sig. different from the other six.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition E2

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F1, F5 vs. all other $p = 0.000$ F6 vs. F2s, F3, F7, F8, F8s $p = 0.001$ F6 vs. F4 $p = 0.005$ F6 vs. F2 $p = 0.009$ F6 vs. F3s
	per condition	$p = 0.000$	$p = 0.000$	F1, F5 vs. all other $p = 0.049$ F3 vs. F8

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

RMSE	F2: $p = 0.004$ F3s: $p = 0.021$ F6: $p = 0.000$ F8: $p = 0.022$
------	--

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 sig. worse than F7.	All except F5 outperform F1, but not F7.

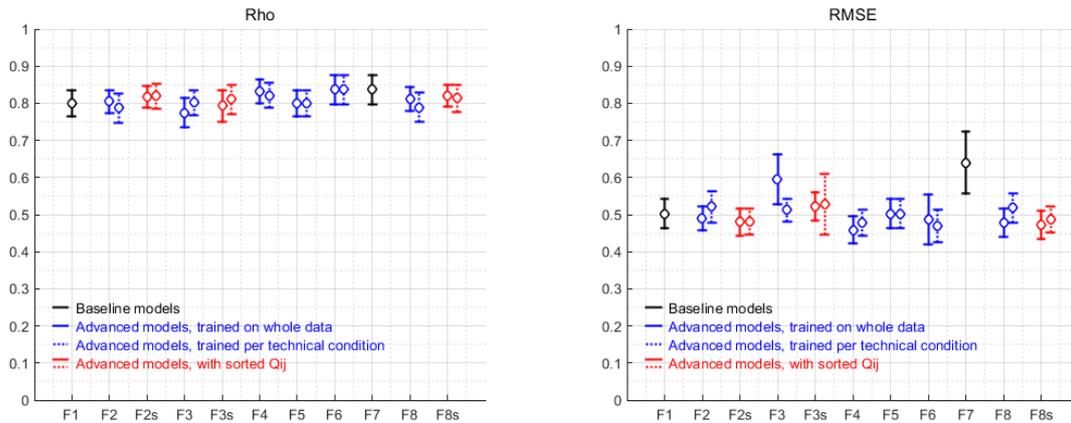
4. Observations and Conclusions

- A) Eight advanced models outperform the baseline mean, but not the baseline minimum. F2s is the best, though not sig. different from the other seven.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition improves performance for F2, F3s, F6 & F8.

Condition E3

Figure H.16: Modeling performance for conditions E2 (top panel) and E3 (bottom panel) of Experiment ACT<sub>3</sub>.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. F2s, F4, F6, F8, F8s
			$p = 0.001$	F7 vs. F2
			$p = 0.003$	F7 vs. F1, F5
			$p = 0.030$	F7 vs. F3s
			$p = 0.004$	F3 vs. F4
RMSE	per condition	$p = 0.000$	$p = 0.018$	F3 vs. F8s
			$p = 0.030$	F3 vs. F8
			$p = 0.040$	F3 vs. F2s
			$p = 0.000$	F7 vs. F2s, F4, F6, F8s
			$p = 0.003$	F7 vs. F1, F5
			$p = 0.016$	F7 vs. F8
			$p = 0.023$	F7 vs. F2
			$p = 0.46$	F7 vs. F3s

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

RMSE F3:  $p = 0.034$

3. Behavior of modeling functions

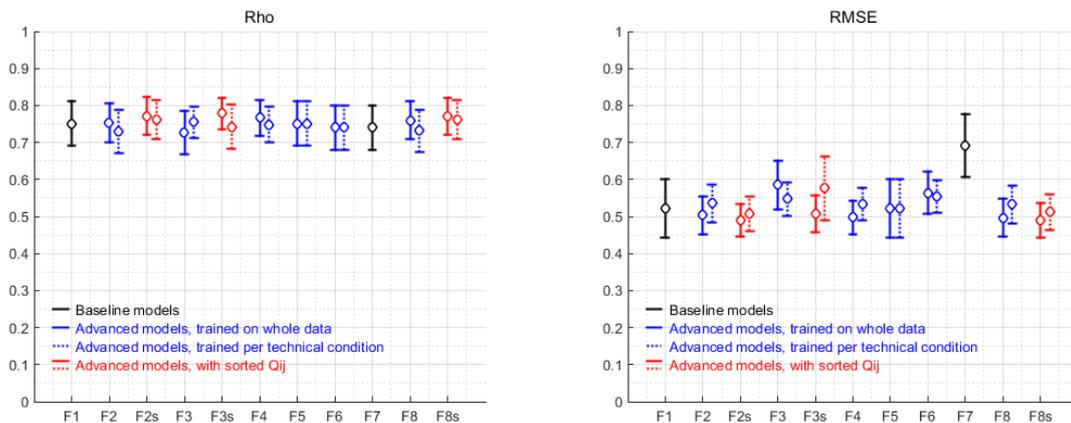
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly worse than F7 (not sig.)	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition improves performance only for F3.

Condition P1

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. F2s, F4, F8, F8s
			$p = 0.001$	F7 vs. F2, F3s
			$p = 0.003$	F7 vs. F1, F5
RMSE	per condition	$p = 0.002$	$p = 0.001$	F7 vs. F2s
			$p = 0.002$	F7 vs. F8s
			$p = 0.004$	F7 vs. F1, F5
			$p = 0.013$	F7 vs. F8
			$p = 0.015$	F7 vs. F4
			$p = 0.016$	F7 vs. F2
			$p = 0.041$	F7 vs. F3

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

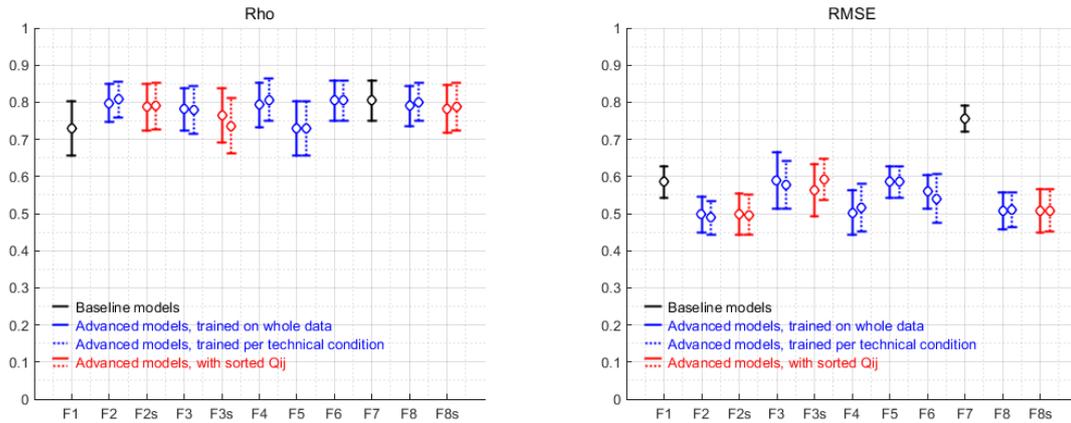
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition P2

Figure H.17: Modeling performance for conditions P1 (top panel) and P2 (bottom panel) of Experiment ACT3.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.001$	F7 vs. F3
			$p = 0.000$	F7 vs. all other
RMSE	per condition	$p = 0.000$	$p = 0.001$	F7 vs. F3s
			$p = 0.000$	F7 vs. all other

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

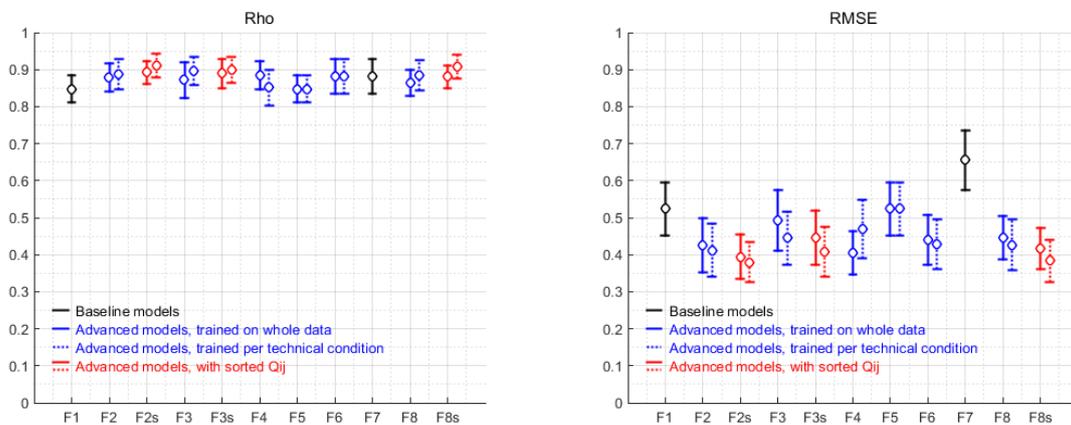
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly worse than F7 (not sig.)	No one outperforms F1, though F2 & F2s are slightly better (not sig.)
RMSE	F1 sig. better than F7.	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition P3

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. F2, F2s, F4, F6, F8s
			$p = 0.001$	F7 vs. F3s, F8
RMSE	per condition	$p = 0.000$	$p = 0.032$	F7 vs. F3
			$p = 0.000$	F7 vs. F2, F2s, F3s, F6, F8, F8s
			$p = 0.001$	F7 vs. F3
			$p = 0.006$	F7 vs. F4

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly worse than F7 (not sig.)	No one outperforms F1.
RMSE	F1 slightly better than F7 (not sig.)	No one outperforms F1, though F2s is slightly better (not sig.)

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

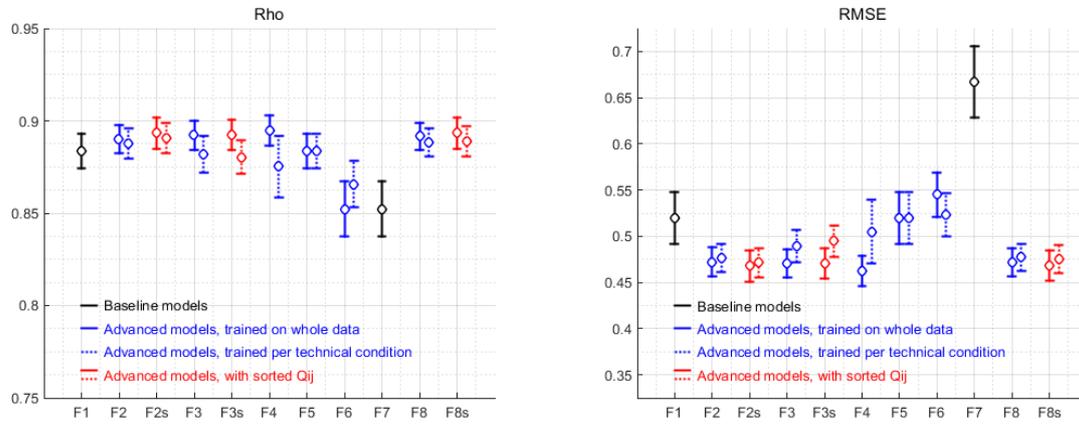
Condition P4

Figure H.18: Modeling performance for conditions P3 (top panel) and P4 (bottom panel) of Experiment ACT3.

## H.4 Experiment LOT<sub>1</sub>

### H.4.1 Performance across technical conditions

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
Rho	whole data	$p = 0.000$	$p = 0.000$	F6, F7 vs. all other
Rho	per condition	$p = 0.000$	$p = 0.000$	F7 vs. F2, F2s, F8, F8s
			$p = 0.001$	F7 vs. F1, F5
			$p = 0.004$	F7 vs. F3
			$p = 0.009$	F7 vs. F3s
			$p = 0.042$	F6 vs. F2s
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. all other
			$p = 0.000$	F6 vs. F2, F2s, F3, F3s, F4, F6, F7, F8, F8s
			$p = 0.012$	F1, F5 vs. F4
			$p = 0.041$	F1, F5 vs. F2s
			$p = 0.043$	F1, F5 vs. F8s
RMSE	per condition	$p = 0.000$	$p = 0.000$	F7 vs. all other

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

Measure	T-test	Functions	Measure	T-test	Functions
Rho	$p = 0.033$	F4	Rho	$p = 0.049$	F3s
RMSE	$p = 0.027$	F4	RMSE	$P = 0.036$	F3s

### 3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 sig. better than F7.	No one outperforms F1.
RMSE	F1 sig. better than F7.	F2s, F4 & F8s sig. outperform F1; F2 & F8 are slightly better than F1 (not sig.)

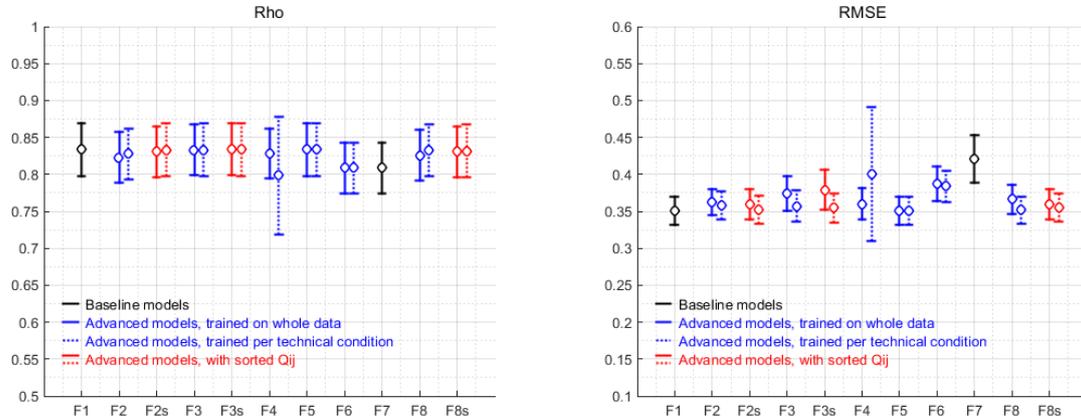
### 4. Observations and Conclusions

- Three advanced models outperform the baseline mean. F4 is the best, though not sig. different from the other two.
- Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- Training per condition does not improve performance, in case of RMSE for F3s & F4 it even reduces performance.

Figure H.19: Modeling performance across all technical conditions of Experiment LOT<sub>1</sub>.

H.4.2 Performance per technical condition

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$ $p = 0.004$ $p = 0.005$ $p = 0.010$	F7 vs. F1, F5 F7 vs. F2s F7 vs. F4, F8s F7 vs. F2
RMSE	per condition	$p = 0.020$	$p = 0.022$ $p = 0.128$ (n.s.)	F7 vs. F8 F7 vs. F1, F5

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

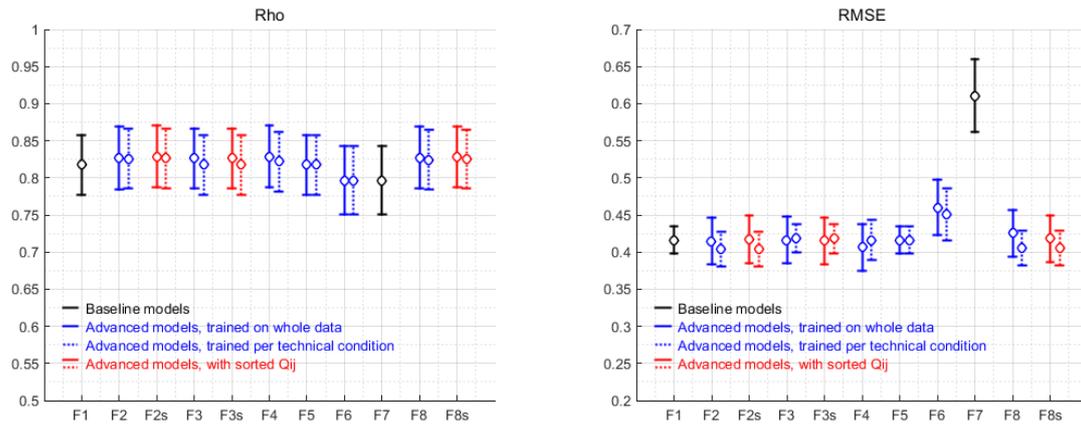
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance, in case of F4 it even leads to unstable performance (large confidence intervals).

Condition Ref5

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. all other
RMSE	per condition	$p = 0.000$	$p = 0.000$	F7 vs. all other

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

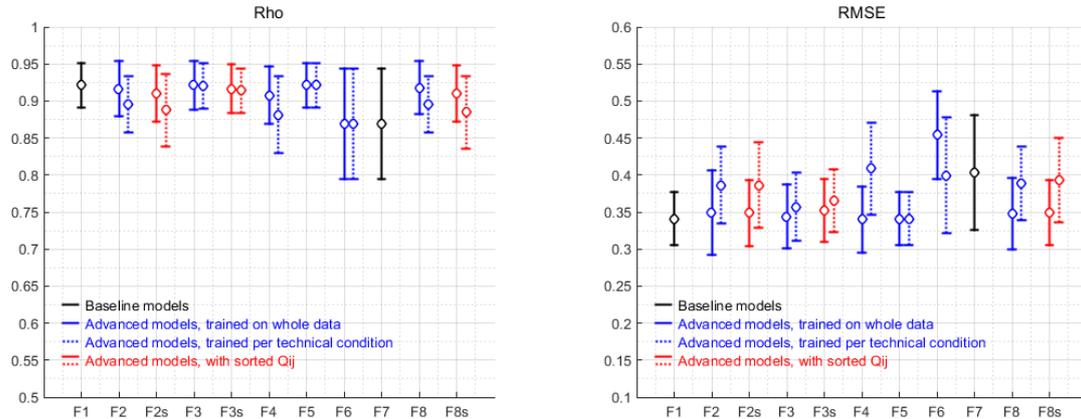
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition B2

Figure H.20: Modeling performance for conditions Ref5 (top panel) and B2 (bottom panel) of Experiment LOT1.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.017$	$p = 0.047$	F4 vs. F6

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

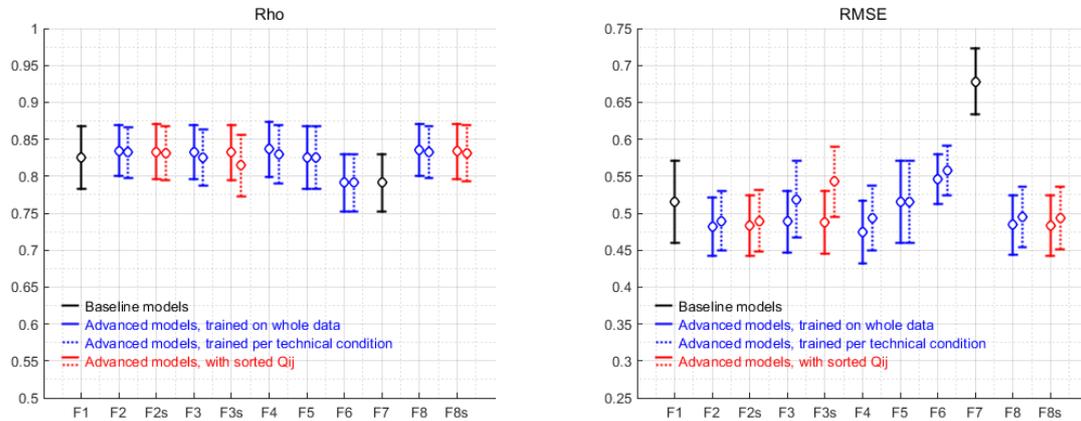
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly better than F7 (not sig.)	No one outperforms F1.
RMSE	F1 slightly better than F7 (not sig.)	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition B2sym

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.001$	F7 vs. F6
			$p = 0.000$	F7 vs. all other
RMSE	per condition	$p = 0.000$	$p = 0.001$	F7 vs. F3s
			$p = 0.007$	F7 vs. F6
			$p = 0.000$	F7 vs. all other

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly better than F7 (not sig.)	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

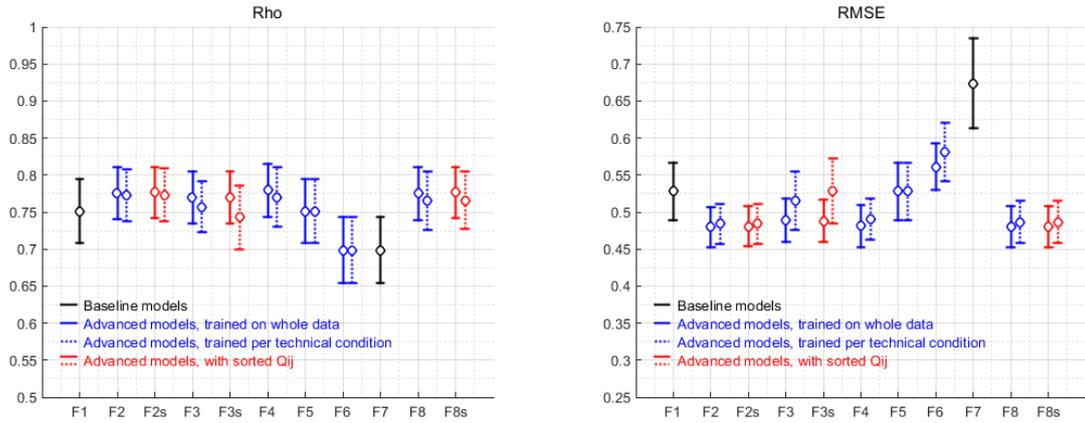
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition P5

Figure H.21: Modeling performance for conditions B2sym (top panel) and P5 (bottom panel) of Experiment LOT1.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

Measure	Training	ANOVA	PostHoc	Functions
Rho	whole data	$p = 0.004$	$p = 0.110$ (n.s.)	F4 vs. F6, F7
Rho	per condition	$p = 0.036$	$p = 0.304$ (n.s.)	F2s vs. F6, F7
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. all other
			$p = 0.034$	F6 vs. F2
			$p = 0.037$	F6 vs. F8
			$p = 0.038$	F6 vs. F8s
			$p = 0.039$	F6 vs. F2s
			$p = 0.042$	F6 vs. F4
RMSE	per condition	$p = 0.000$	$p = 0.018$	F7 vs. F6
			$p = 0.000$	F7 vs. all other
			$p = 0.010$	F6 vs. F2s
			$p = 0.011$	F6 vs. F2
			$p = 0.014$	F6 vs. F8s
			$p = 0.015$	F6 vs. F8
			$p = 0.026$	F6 vs. F4

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

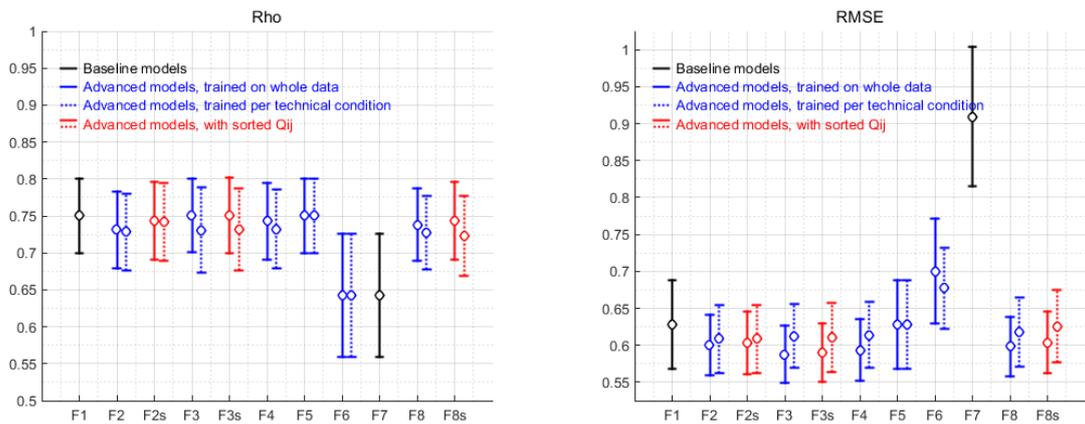
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly better than F7 (not sig.)	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition B2symP5

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
Rho	whole data	$p = 0.016$	$p = 0.297$ (n.s.)	F3s vs. F6, F7
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. all other
RMSE	per condition	$p = 0.000$	$p = 0.000$	F7 vs. all other

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly better than F7 (not sig.)	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

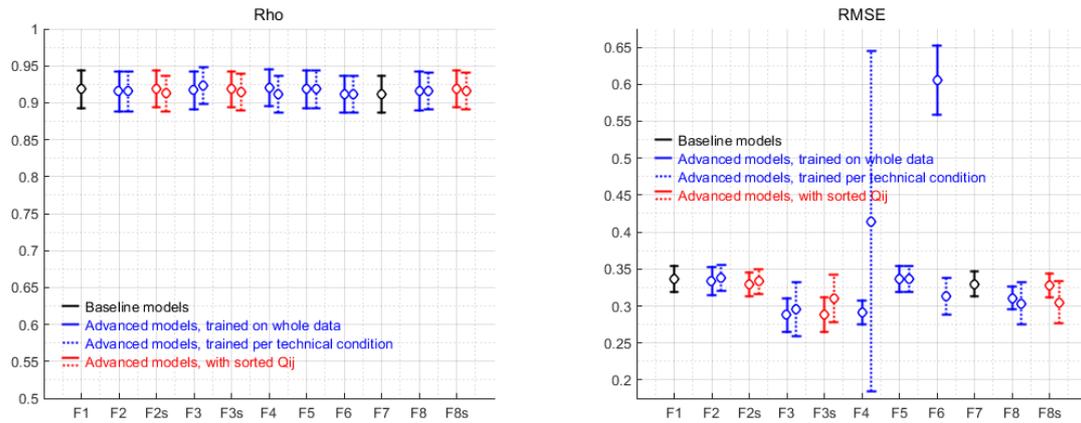
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition E4

Figure H.22: Modeling performance for conditions B2symP5 (top panel) and E4 (bottom panel) of Experiment LOT1.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F6 vs. all other

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

RMSE	F6: $p = 0.000$
------	-----------------

3. Behavior of modeling functions

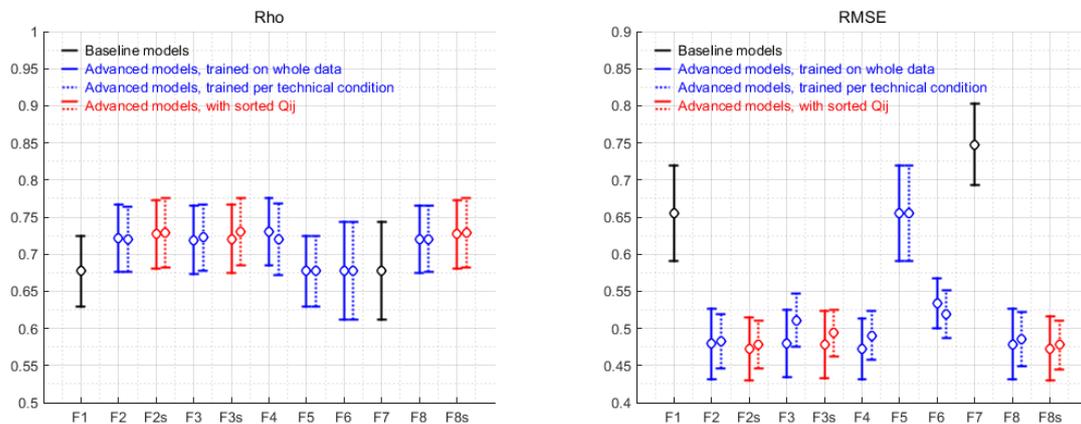
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 essentially equal to F7.	No one outperforms F1, though F3, F3s & F4 are slightly better (not sig.)

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance, in case of RMSE for F4 it even leads to unstable performance (large confidence intervals).

Condition E4sym

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions	
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. F2, F2s, F3, F3s, F4, F6, F8, F8s	
				$p = 0.000$	F1, F5 vs. F2, F2s, F3, F3s, F4, F8, F8s
				$p = 0.017$	F1, F5 vs. F6
RMSE	per condition	$p = 0.000$	$p = 0.000$	F7 vs. F2, F2s, F3, F3s, F4, F6, F8, F8s	
				$p = 0.000$	F1, F5 vs. F2, F2s, F3, F3s, F4, F6, F8, F8s

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 slightly better than F7 (not sig.)	F2, F2s, F3, F3s, F4, F6, F8, & F8s sig. outperform F1.

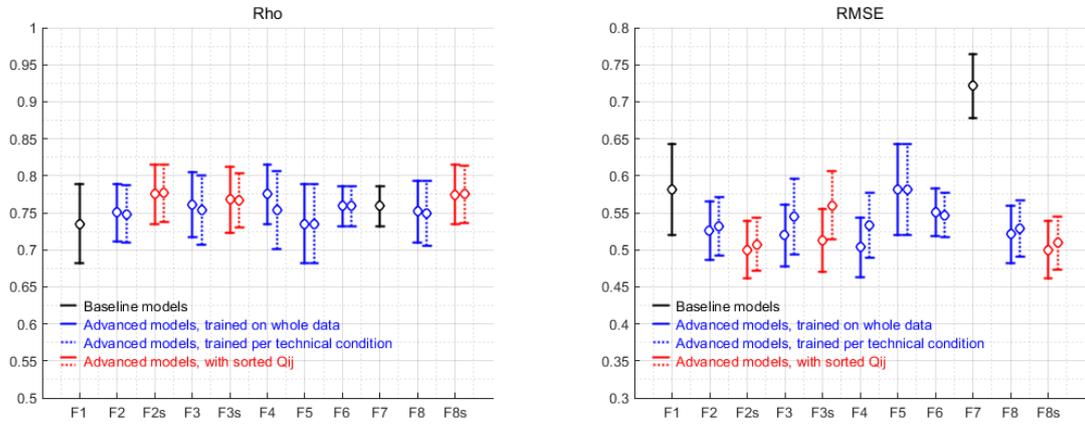
4. Observations and Conclusions

- A) Eight advanced models outperform the baseline mean, seven of them with essentially equal performance.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition R

Figure H.23: Modeling performance for conditions E4sym (top panel) and R (bottom panel) of Experiment LOT1.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. all other
RMSE	per condition	$p = 0.000$	$p = 0.000$	F7 vs. all other

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly worse than F7 (not sig.)	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

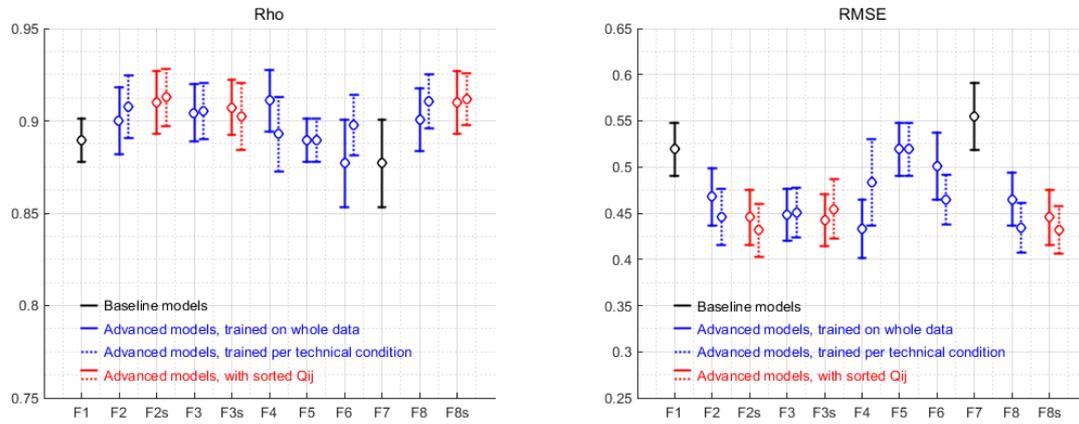
Condition CN

Figure H.24: Modeling performance for condition CN of Experiment LOT1.

## H.5 Experiment LOT2

### H.5.1 Performance across technical conditions

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
Rho	whole data	$p = 0.013$	$p = 0.223$ (n.s.)	F4 vs. F6, F7
Rho	per condition	$p = 0.028$	$p = 0.094$ (n.s.)	F2s vs. F7
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. F2s, F3, F3s, F4, F8s
			$p = 0.002$	F7 vs. F8
			$p = 0.003$	F7 vs. F2
			$p = 0.003$	F1, F5 vs. F4
			$p = 0.016$	F1, F5 vs. F3s
			$p = 0.028$	F1, F5 vs. F2s
			$p = 0.029$	F1, F5 vs. F8s
			$p = 0.046$	F1, F5 vs. F3
RMSE	per condition	$p = 0.000$	$p = 0.000$	F7 vs. F2, F2s, F3, F3s, F8, F8s
			$p = 0.002$	F7 vs. F6
			$p = 0.003$	F1, F5 vs. F2s, F8s
			$p = 0.005$	F1, F5 vs. F8
			$p = 0.037$	F1, F5 vs. F2

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

### 3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2–F6, F8, F8s)
Rho	F1 slightly better than F7 (not sig.)	No one outperforms F1.
RMSE	F1 slightly better than F7 (not sig.)	F2s, F3, F3s, F8s, F2 & F8 trained per condition and F4 trained on whole data outperform F1.

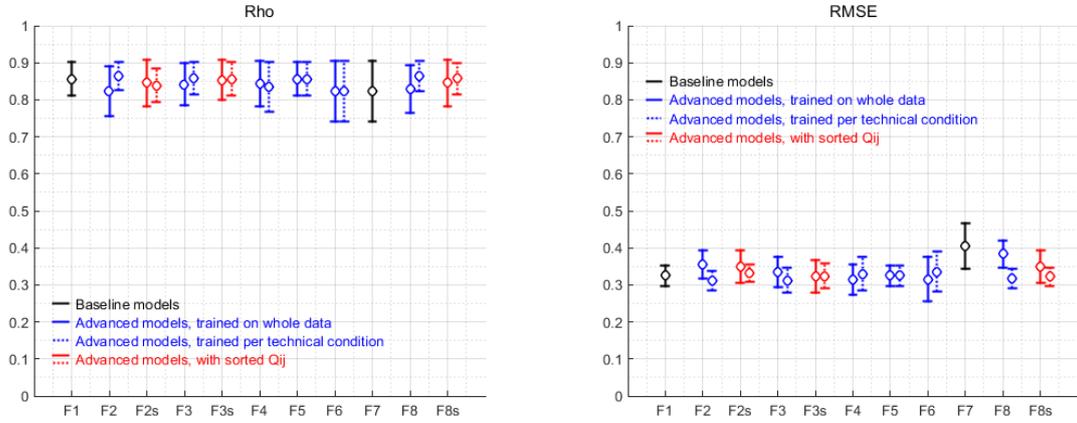
### 4. Observations and Conclusions

- Seven advanced models outperform the baseline mean. F4 (trained on whole data) is the best, though not sig. different from the other six.
- Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- Training per condition does not improve performance.

Figure H.25: Modeling performance across all technical conditions of Experiment LOT2.

H.5.2 Performance per technical condition

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.037$	$p = 0.104$ (n.s.)	F7 vs. F4
RMSE	per condition	$p = 0.025$	$p = 0.013$	F7 vs. F2
			$p = 0.016$	F7 vs. F3
			$p = 0.033$	F7 vs. F8

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for: RMSE F8:  $p = 0.004$

3. Behavior of modeling functions

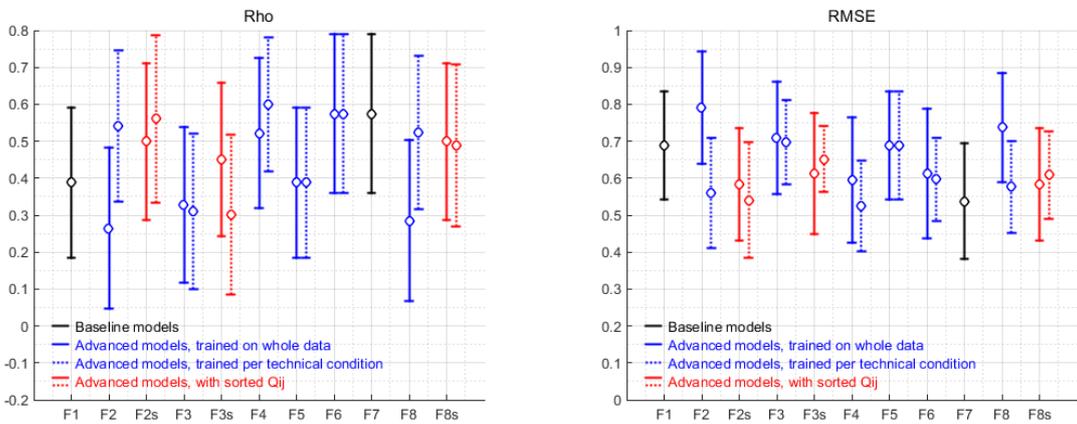
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 slightly better than F7 (not sig.)	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition Ref3

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

Measure	T-test	Functions	Measure	T-test	Functions
Rho	$p = 0.059$ (n.s.)	F2	RMSE	$p = 0.028$	F2

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly worse than F7 (not sig.)	No one outperforms F1.
RMSE	F1 slightly worse than F7 (not sig.)	No one outperforms F1.

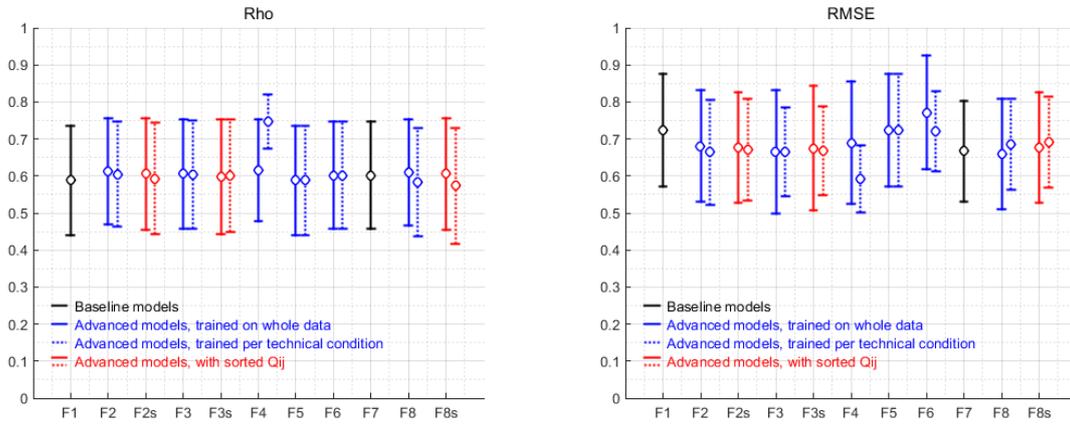
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition improves performance only for F2.

Condition TN1

Figure H.26: Modeling performance for conditions Ref3 (top panel) and TN1 (bottom panel) of Experiment LOT2.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

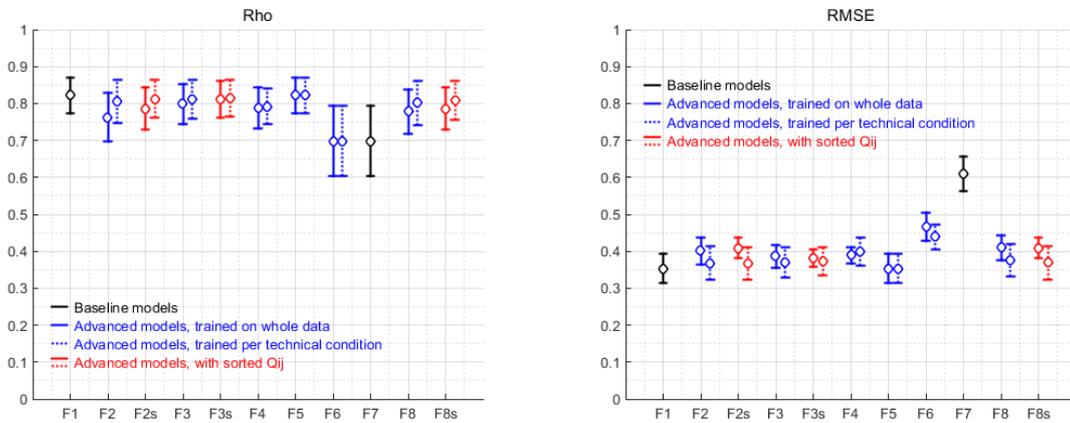
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho, RMSE	F1 essentially equal to F7.	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition TN2

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
Rho	whole data	$p = 0.042$	$p = 0.247$ (n.s.)	F1, F5 vs. F6, F7
Rho	per condition	$p = 0.011$	$p = 0.185$ (n.s.)	F1, F5 vs. F6, F7
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. all other
			$p = 0.000$	F6 vs. F1, F5
			$p = 0.014$	F6 vs. F3s
			$p = 0.033$	F6 vs. F3
RMSE	per condition	$p = 0.000$	$p = 0.000$	F7 vs. all other

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly better than F7 (not sig.)	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

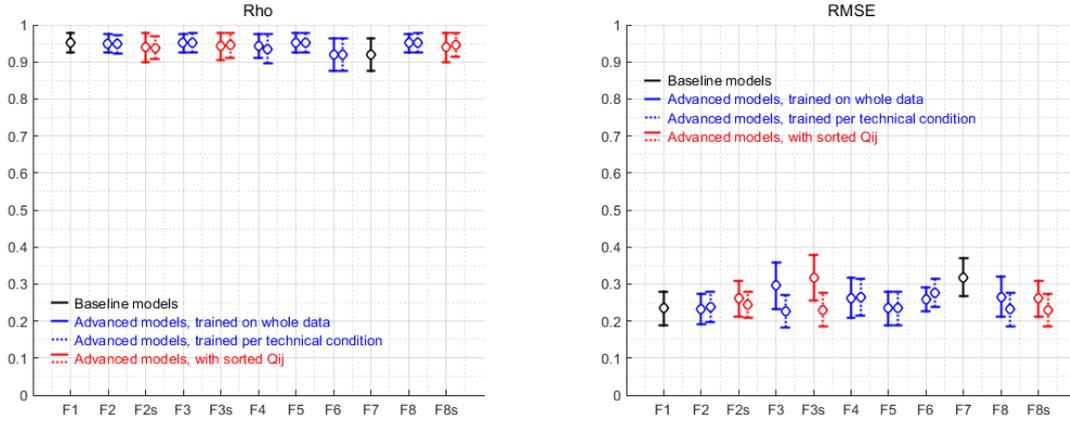
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition B2

Figure H.27: Modeling performance for conditions TN2 (top panel) and B2 (bottom panel) of Experiment LOT2.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

RMSE F3s:  $p = 0.023$

3. Behavior of modeling functions

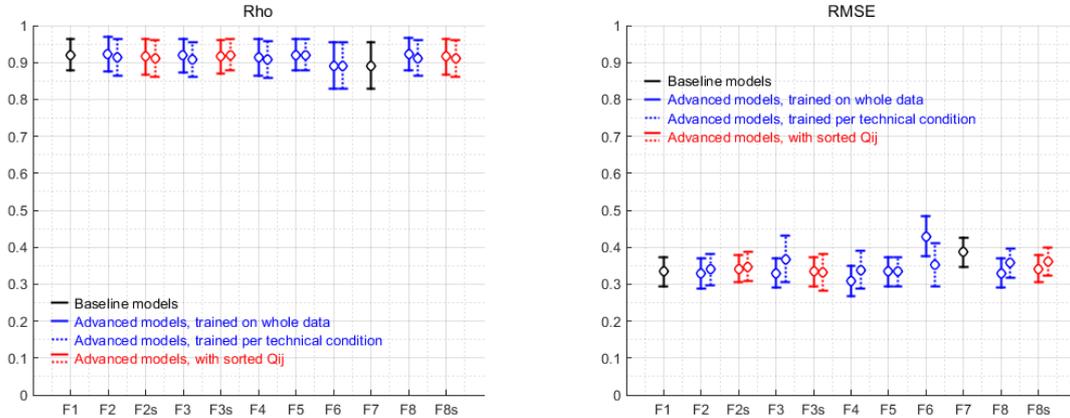
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly better than F7 (not sig.)	No one outperforms F1.
RMSE	F1 slightly better than F7 (not sig.)	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition improves performance only for F3s.

Condition B2sym

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.001$	F6 vs. F4
			$p = 0.016$	F6 vs. F2
			$p = 0.019$	F6 vs. F8
			$p = 0.020$	F6 vs. F3
			$p = 0.034$	F6 vs. F3s
			$p = 0.037$	F6 vs. F1, F5

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

RMSE F6:  $p = 0.048$

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly better than F7 (not sig.)	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

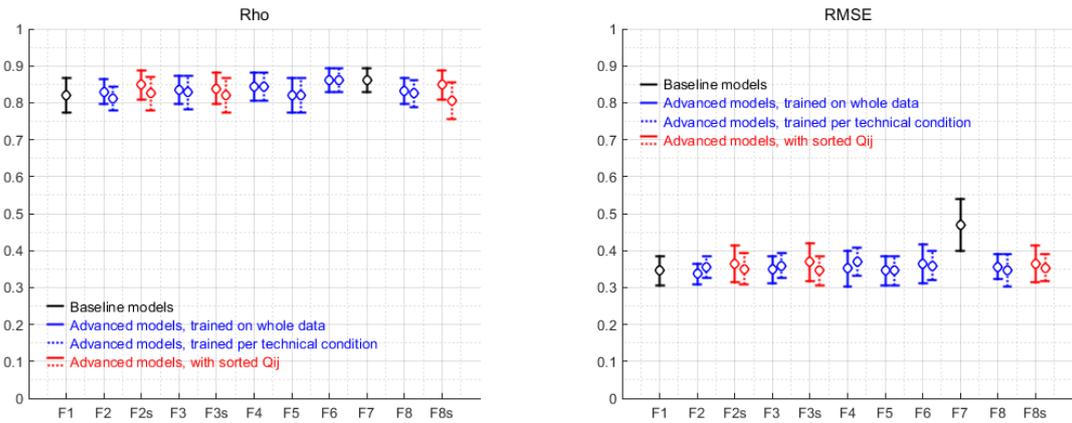
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition improves performance only for F6.

Condition FL4sym

Figure H.28: Modeling performance for conditions B2sym (top panel) and FL4sym (bottom panel) of Experiment LOT2.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.004$	$p = 0.002$	F7 vs. F2
			$p = 0.006$	F7 vs. F1, F5
			$p = 0.007$	F7 vs. F3
			$p = 0.011$	F7 vs. F4
			$p = 0.019$	F7 vs. F8
			$p = 0.042$	F7 vs. F6
RMSE	per condition	$p = 0.001$	$p = 0.045$	F7 vs. F2s
			$p = 0.046$	F7 vs. F8s
			$p = 0.001$	F7 vs. F1, F3s, F5, F8
			$p = 0.002$	F7 vs. F2s
			$p = 0.003$	F7 vs. F8s
			$p = 0.004$	F7 vs. F2
			$p = 0.006$	F7 vs. F6
			$p = 0.007$	F7 vs. F3
			$p = 0.028$	F7 vs. F4

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

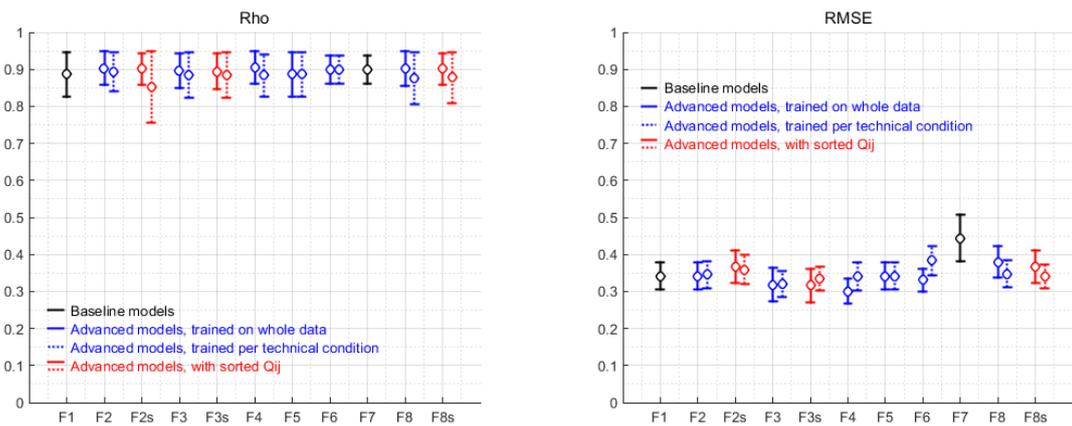
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly better than F7 (not sig.)	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition FM1

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. F4
			$p = 0.001$	F7 vs. F3, F3s
			$p = 0.023$	F7 vs. F1, F5
			$p = 0.024$	F7 vs. F2
RMSE	per condition	$p = 0.001$	$p = 0.000$	F7 vs. F3
			$p = 0.003$	F7 vs. F3s
			$p = 0.007$	F7 vs. F4, F8s
			$p = 0.009$	F7 vs. F1, F5
			$p = 0.014$	F7 vs. F2
			$p = 0.019$	F7 vs. F8

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

RMSE F6:  $p = 0.035$

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly better than F7 (not sig.)	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

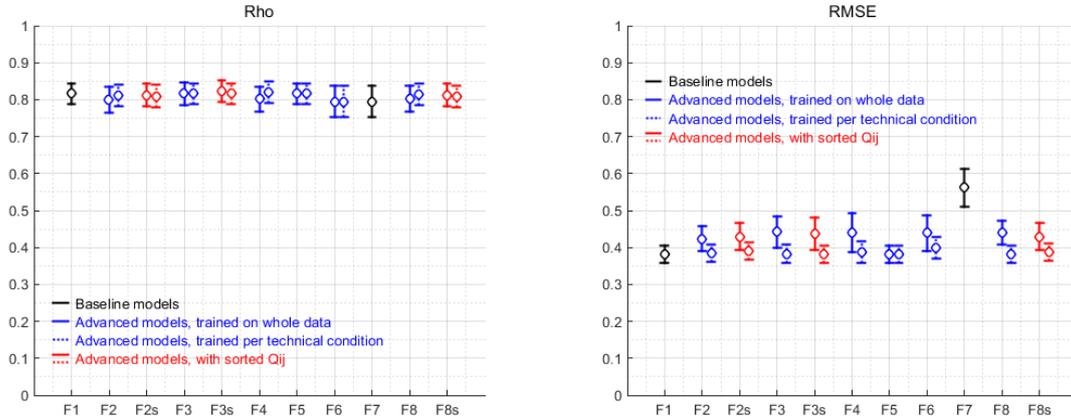
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition FM1sym

Figure H.29: Modeling performance for conditions FM1 (top panel) and FM1sym (bottom panel) of Experiment LOT2.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. F1, F2, F2s, F3s, F5, F6, F8s
RMSE	per condition	$p = 0.000$	$p = 0.001$ $p = 0.000$	F7 vs. F3, F4, F8 F7 vs. all other

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

Measure	T-test	Functions	Measure	T-test	Functions
RMSE	$p = 0.015$	F3	RMSE	$p = 0.022$	F3s
RMSE	$p = 0.004$	F8	RMSE	$p = 0.045$	F8s

3. Behavior of modeling functions

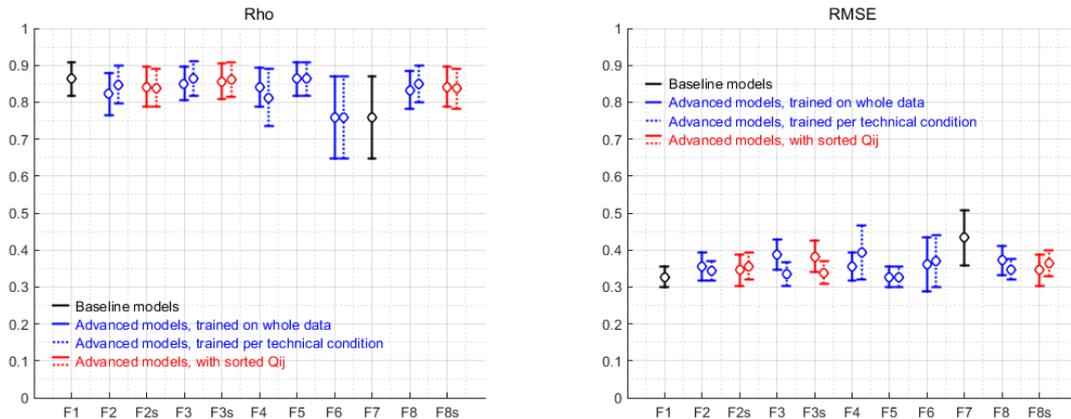
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly better than F7 (not sig.)	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition improves performance for F3, F3s, F8 & F8s.

Condition FM2

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	per condition	$p = 0.027$	$p = 0.047$	F7 vs. F1, F5

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

RMSE	F3: $p = 0.039$
------	-----------------

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho, RMSE	F1 slightly better than F7 (not sig.)	No one outperforms F1.

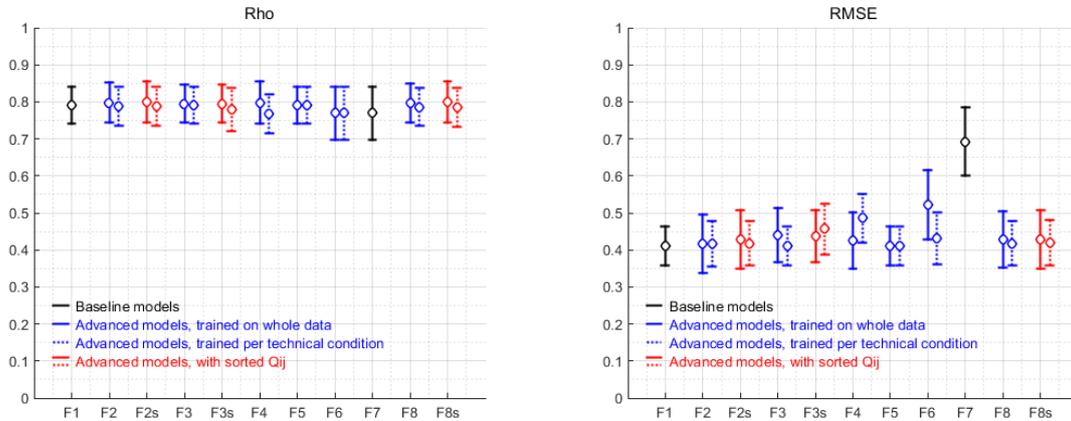
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition improves performance only for F3.

Condition FM2sym

Figure H.30: Modeling performance for conditions FM2 (top panel) and FM2sym (bottom panel) of Experiment LOT2.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. all other except F6
RMSE	per condition	$p = 0.000$	$p = 0.000$	F7 vs. all other

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

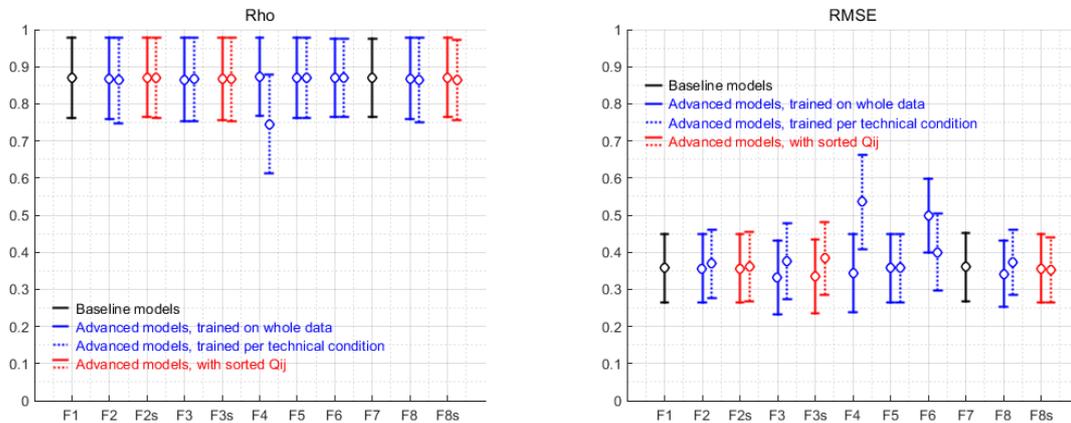
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition FM4

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

RMSE	F4: $p = 0.019$
------	-----------------

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho, RMSE	F1 essentially equal to F7.	No one outperforms F1.

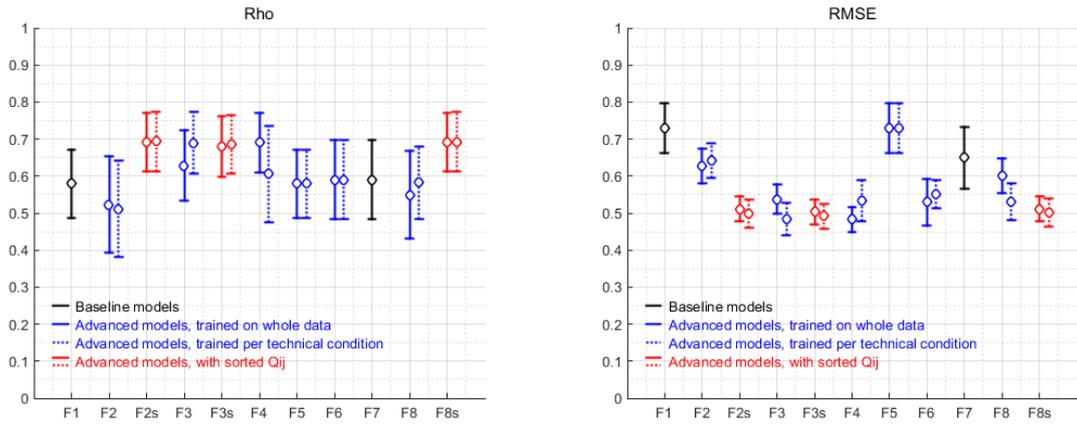
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance, in case of RMSE for F4 it even reduces performance.

Condition FM4sym

Figure H.31: Modeling performance for conditions FM4 (top panel) and FM4sym (bottom panel) of Experiment LOT2.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

Measure	Training	ANOVA	PostHoc	Functions				
RMSE	whole data	$p = 0.000$	$p = 0.000$	F1, F5 vs. F2s, F3, F3s, F4, F6, F8s				
				$p = 0.020$	F1, F5 vs. F8			
				$p = 0.000$	F7 vs. F4			
				$p = 0.003$	F7 vs. F3s			
				$p = 0.007$	F7 vs. F2s, F8s			
				$p = 0.046$	F7 vs. F6			
				$p = 0.003$	F2 vs. F4			
				$p = 0.030$	F2 vs. F3s			
				RMSE	per condition	$p = 0.000$	$p = 0.000$	F1, F5 vs. F2s, F3, F3s, F4, F6, F8, F8s
								$p = 0.000$
$p = 0.001$	F7 vs. F3s							
$p = 0.002$	F7 vs. F2s, F8s							
$p = 0.001$	F2 vs. F3							
$p = 0.002$	F2 vs. F3s							
$p = 0.006$	F2 vs. F2s, F8s							

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

RMSE F8:  $p = 0.035$

3. Behavior of modeling functions

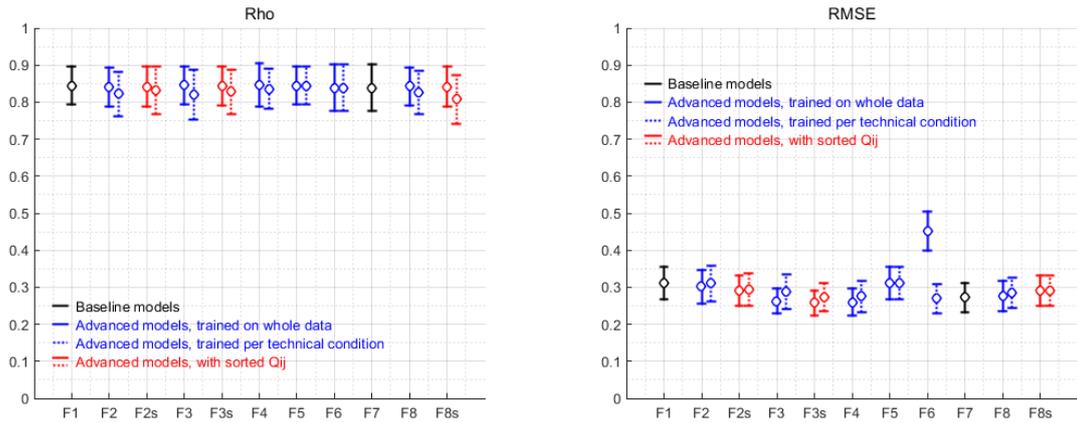
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly better than F7 (not sig.)	No one outperforms F1.
RMSE	F1 sig. better than F7.	F2s, F3, F3s, F4, F6, F8s and F8 trained per condition outperform F1. F2s, F3s, F8s and F4 & F6 trained on whole data and F3 trained per condition outperform F7.

4. Observations and Conclusions

- A) Seven advanced models outperform the baseline mean, six of them the baseline minimum.
- B) Sorting-based assignment of  $Q_{ij}$  outperforms rule-based assignment for one function (F2s better than F2).
- C) Training per condition improves performance only for F8.

Condition FM4P3

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F6 vs. all other

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

RMSE F6:  $p = 0.000$

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 slightly worse than F7 (not sig.)	No one outperforms F1.

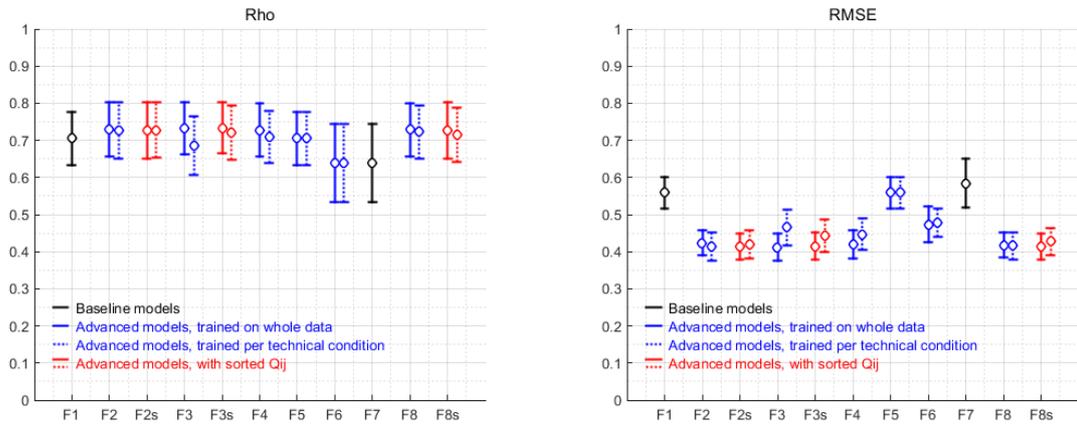
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition improves performance only for F6.

Condition FM4P3sym

Figure H.32: Modeling performance for conditions FM4P3 (top panel) and FM4P3sym (bottom panel) of Experiment LOT2.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F1, F5 vs. F2, F2s, F3, F3s, F4, F8, F8s
				$p = 0.000$
RMSE	per condition	$p = 0.000$	$p = 0.007$	F7 vs. F6
			$p = 0.000$	F1, F5 vs. F2, F2s, F8
			$p = 0.001$	F1, F5 vs. F8s
			$p = 0.008$	F1, F5 vs. F3s
			$p = 0.011$	F1, F5 vs. F4
			$p = 0.000$	F7 vs. F2, F2s, F3s, F4, F8, F8s
			$p = 0.005$	F7 vs. F3
$p = 0.020$	F7 vs. F6			

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

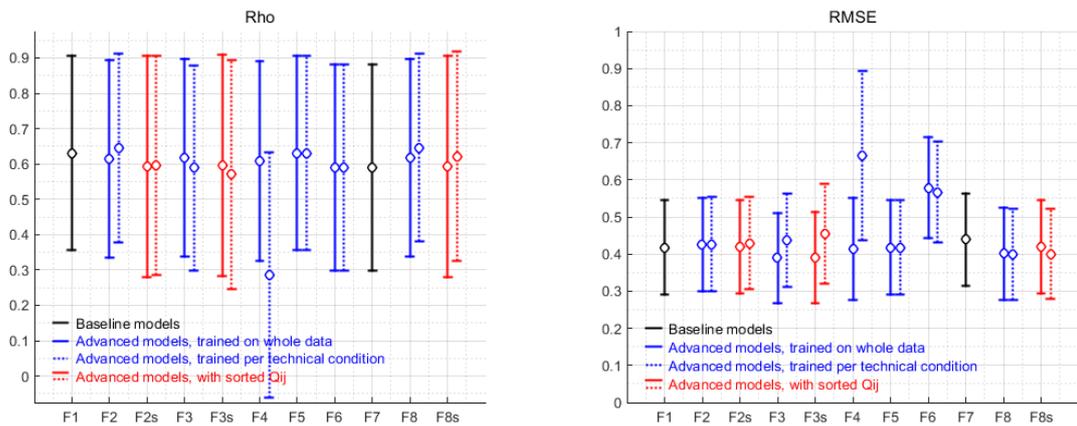
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly better than F7 (not sig.)	No one outperforms F1.
RMSE	F1 essentially equal to F7.	F2, F2s, F3, F4, F8, F8s and F3s trained on whole data outperform F1.

4. Observations and Conclusions

- A) Seven advanced models sig. outperform the baseline mean, all with essentially equal performance.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition E2

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho, RMSE	F1 essentially equal to F7.	No one outperforms F1.

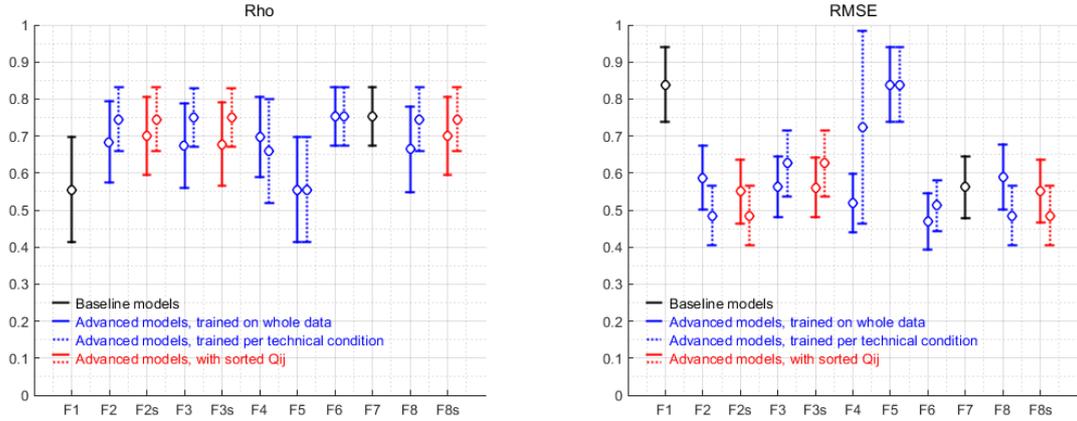
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance, for F4 it even tends to reduce performance.

Condition E2sym

Figure H.33: Modeling performance for conditions E2 (top panel) and E2sym (bottom panel) of Experiment LOT2.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

Measure	Training	ANOVA	PostHoc	Functions
Rho	per condition	$p = 0.006$	$p = 0.233$ (n.s.)	F1, F5 vs. F6, F7
RMSE	whole data	$p = 0.000$	$p = 0.000$	F1, F5 vs. F2s, F3, F3s, F4, F6, F7, F8s
RMSE	per condition	$p = 0.000$	$p = 0.002$	F1, F5 vs. F2, F8
			$p = 0.000$	F1, F5 vs. F2, F2s, F8, F8s
			$p = 0.002$	F1, F5 vs. F6
			$p = 0.021$	F1, F5 vs. F7

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

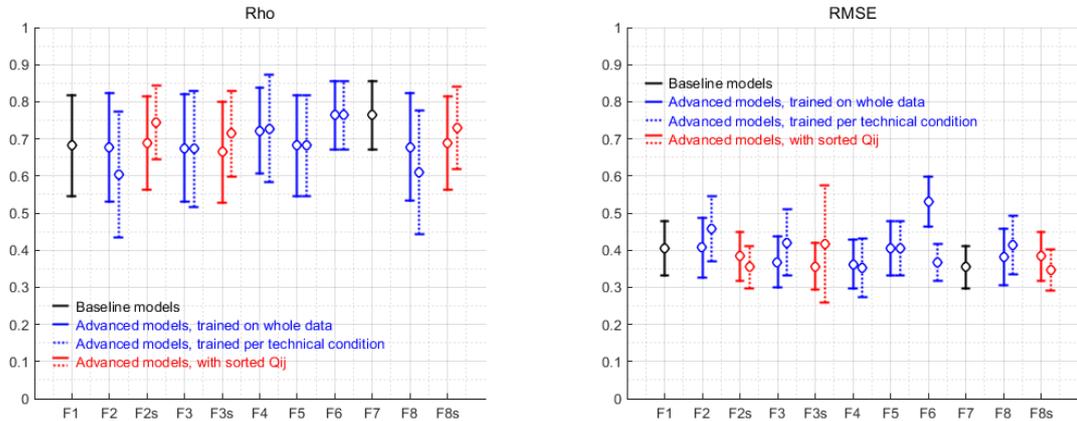
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly worse than F7 (not sig.)	No one outperforms F1 nor F7.
RMSE	F1 sig. worse than F7.	No one outperforms F1 nor F7.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance, in case of RMSE for F4 it even leads to unstable performance (large confidence intervals).

Condition E3

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.018$	$p = 0.011$	F6 vs. F7
			$p = 0.013$	F6 vs. F3s
			$p = 0.020$	F6 vs. F4
			$p = 0.032$	F6 vs. F3

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

RMSE	F6: $p = 0.000$
------	-----------------

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho, RMSE	F1 slightly worse than F7 (not sig.)	No one outperforms F1 nor F7.

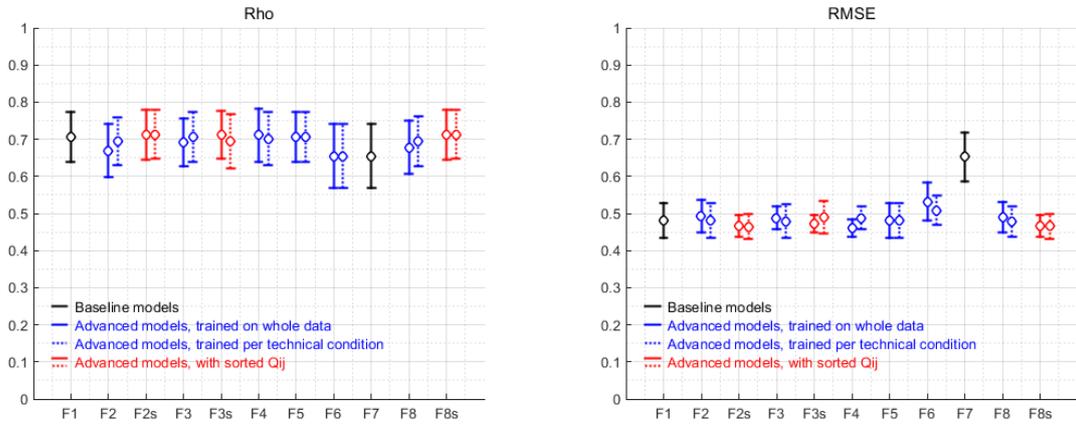
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition improves performance only for F6, and leads to unstable performance for F3s (large confidence intervals).

Condition E3sym

Figure H.34: Modeling performance for conditions E3 (top panel) and E3sym (bottom panel) of Experiment LOT2.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.001$	F7 vs. F6
			$p = 0.000$	F7 vs. all other
RMSE	per condition	$p = 0.000$	$p = 0.000$	F7 vs. all other

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

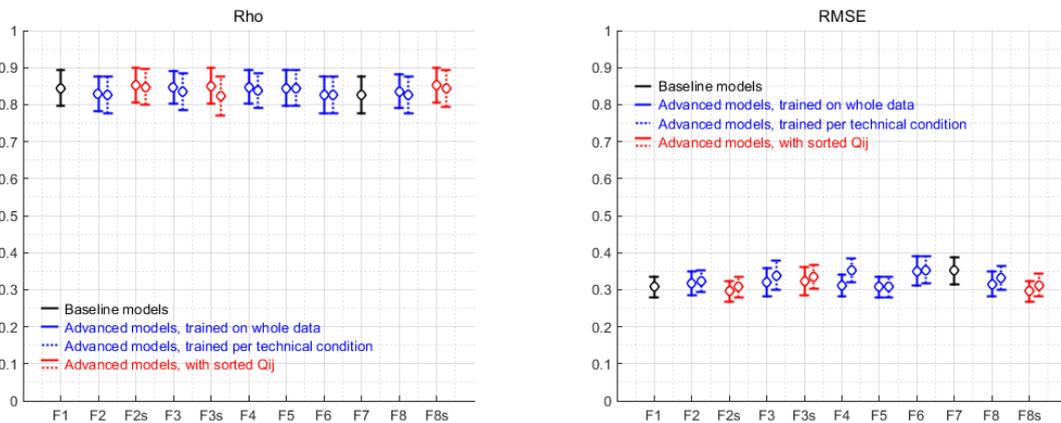
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition P1

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 slightly better than F7 (not sig.)	No one outperforms F1.

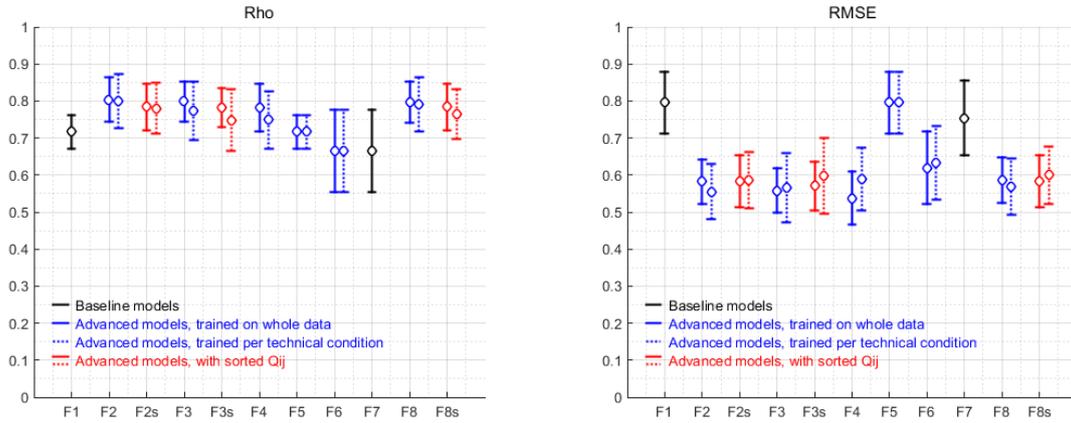
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition P1sym

Figure H.35: Modeling performance for conditions P1 (top panel) and P1sym (bottom panel) of Experiment LOT2.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

Measure	Training	ANOVA	PostHoc	Functions
Rho	whole data	$p = 0.006$	$p = 0.179$ (n.s.)	F2 vs. F6, F7
RMSE	whole data	$p = 0.000$	$p = 0.000$ $p = 0.001$ $p = 0.003$ $p = 0.004$ $p = 0.038$ $p = 0.002$ $p = 0.011$ $p = 0.027$	F1, F5 vs. F3, F4 F1, F5 vs. F3s F1, F5 vs. F2, F2s, F8s F1, F5 vs. F8 F1, F5 vs. F6 F7 vs. F4 F7 vs. F3 F7 vs. F3s
RMSE	per condition	$p = 0.000$	$p = 0.003$ $p = 0.006$ $p = 0.008$ $p = 0.025$ $p = 0.027$ $p = 0.048$ $p = 0.048$	F1, F5 vs. F2 F1, F5 vs. F3 F1, F5 vs. F8 F1, F5 vs. F2s F1, F5 vs. F4 F1, F5 vs. F3s F7 vs. F2

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

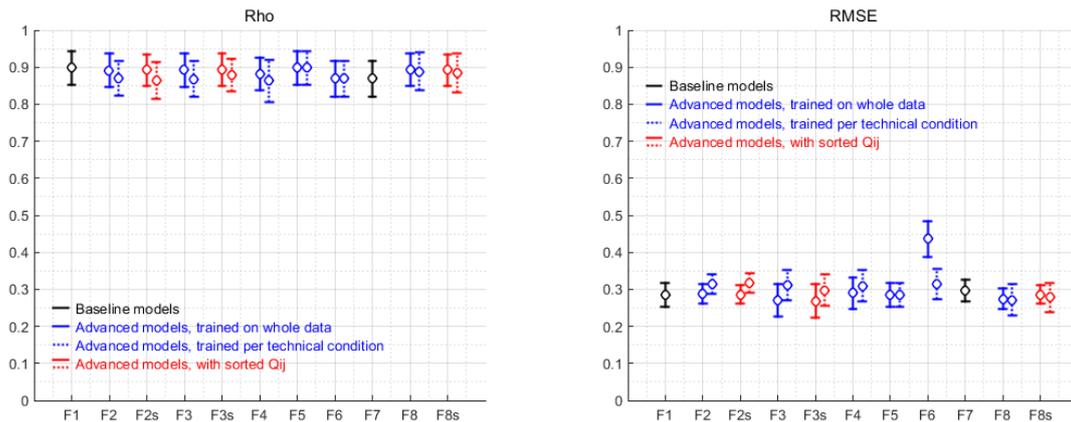
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly better than F7 (not sig.)	No one outperforms F1.
RMSE	F1 slightly worse than F7 (not sig.)	F2, F2s, F3, F3s, F4, F8, F8s and F6 trained on whole data outperform F1. F3, F3s & F4 trained on whole data and F2 trained per condition outperform F7.

4. Observations and Conclusions

- A) Eight advanced models outperform the baseline mean, four the baseline minimum.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition P3

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F6 vs. all other

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

RMSE	F6: $p = 0.000$
------	-----------------

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho, RMSE	F1 essentially equal to F7.	No one outperforms F1.

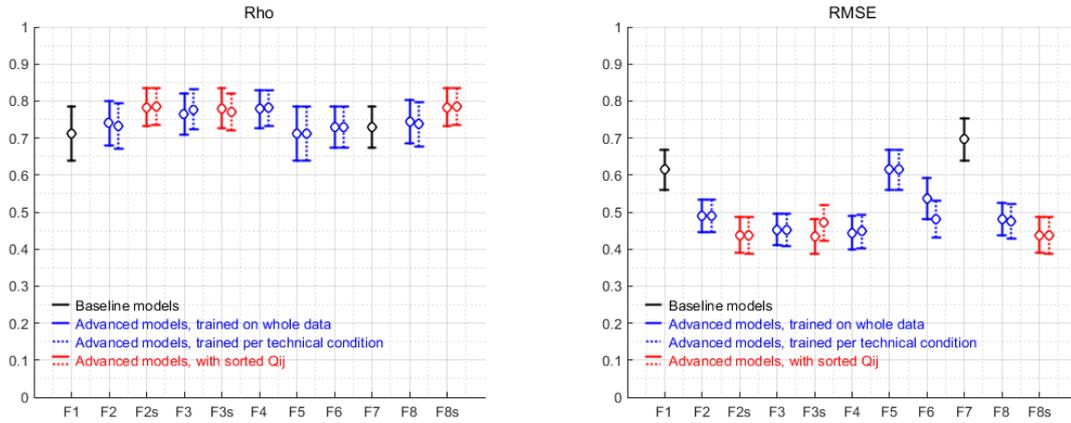
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition improves performance only for F6.

Condition P3sym

Figure H.36: Modeling performance for conditions P3 (top panel) and P3sym (bottom panel) of Experiment LOT2.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

Measure	Training	ANOVA	PostHoc	Functions	
RMSE	whole data	$p = 0.000$	$p = 0.000$	F1, F5 vs. F2s, F3, F3s, F4, F8s	
				$p = 0.005$	F1, F5 vs. F8
				$p = 0.015$	F1, F5 vs. F2
				$p = 0.000$	F7 vs. F2, F2s, F3, F3s, F4, F6, F8, F8s
RMSE	per condition	$p = 0.000$	$p = 0.000$	F1, F5 vs. F2s, F3, F4, F8s	
				$p = 0.002$	F1, F5 vs. F3s
				$p = 0.003$	F1, F5 vs. F8
				$p = 0.006$	F1, F5 vs. F6
				$p = 0.016$	F1, F5 vs. F2
				$p = 0.000$	F7 vs. F2, F2s, F3, F3s, F4, F6, F8, F8s

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

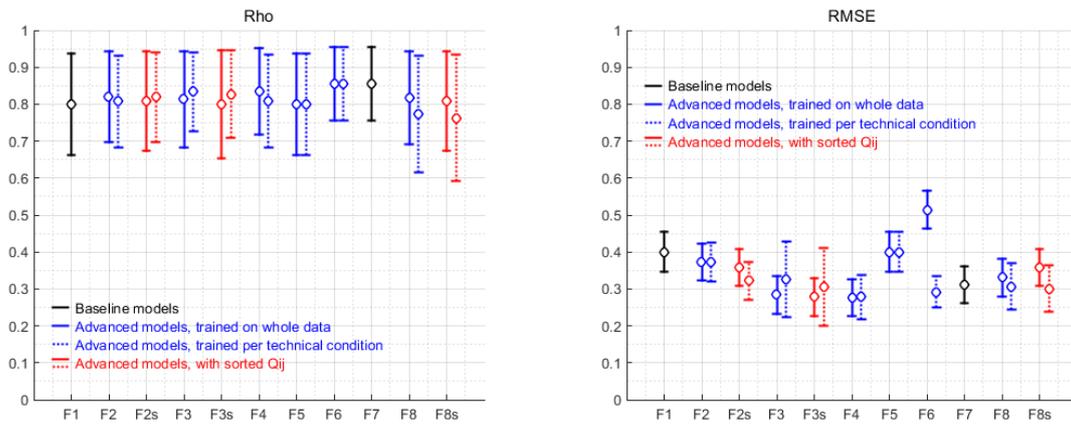
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 slightly better than F7 (not sig.)	All except F5 outperform F1.

4. Observations and Conclusions

- A) Eight advanced models outperform the baseline mean, all with not sig. different performance.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition P4

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions	
RMSE	whole data	$p = 0.000$	$p = 0.000$	F6 vs. F3, F3s, F7, F8	
				$p = 0.001$	F6 vs. F2s, F8s
				$p = 0.003$	F6 vs. F2
				$p = 0.023$	F1, F5 vs. F4
				$p = 0.029$	F1, F5 vs. F3s

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

RMSE	F6: $p = 0.000$
------	-----------------

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly worse than F7 (not sig.)	No one outperforms F1 nor F7.
RMSE	F1 slightly worse than F7 (not sig.)	F3s & F4 trained on whole data outperform F1.

4. Observations and Conclusions

- A) Two advanced models (trained on whole data) outperform the baseline mean, but not the minimum.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition improves performance only for F6.

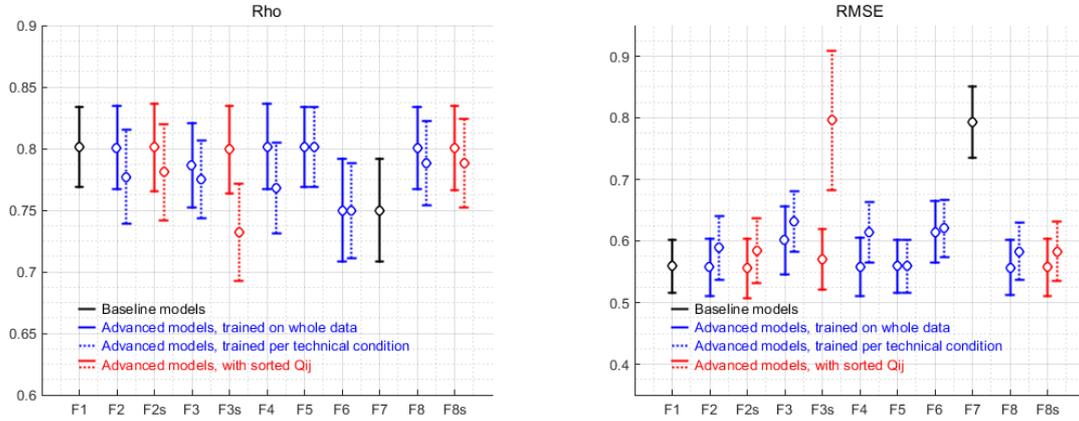
Condition P4sym

Figure H.37: Modeling performance for conditions P4 (top panel) and P4sym (bottom panel) of Experiment LOT2.

## H.6 Experiment AVCT<sub>1</sub>

### H.6.1 Performance across technical conditions

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F<sub>1</sub> - F<sub>8</sub>s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. all other
RMSE	per condition	$p = 0.000$	$p = 0.000$	F <sub>3s</sub> , F <sub>7</sub> vs. F <sub>1</sub> , F <sub>2</sub> , F <sub>2s</sub> , F <sub>4</sub> , F <sub>5</sub> , F <sub>8</sub> , F <sub>8s</sub>
			$p = 0.001$	F <sub>3s</sub> , F <sub>7</sub> vs. F <sub>6</sub>
			$p = 0.003$	F <sub>3s</sub> , F <sub>7</sub> vs. F <sub>3</sub>

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

Measure	T-test	Functions	Measure	T-test	Functions
Rho	$p = 0.012$	F <sub>3s</sub>	RMSE	$p = 0.001$	F <sub>3s</sub>

### 3. Behavior of modeling functions

Measure	Baseline (F <sub>1</sub> & F <sub>7</sub> )	Advanced (F <sub>2</sub> -F <sub>6</sub> , F <sub>8</sub> , F <sub>8s</sub> )
Rho	F <sub>1</sub> slightly better than F <sub>7</sub> (not sig.)	No one outperforms F <sub>1</sub> .
RMSE	F <sub>1</sub> sig. better than F <sub>7</sub> .	No one outperforms F <sub>1</sub> .

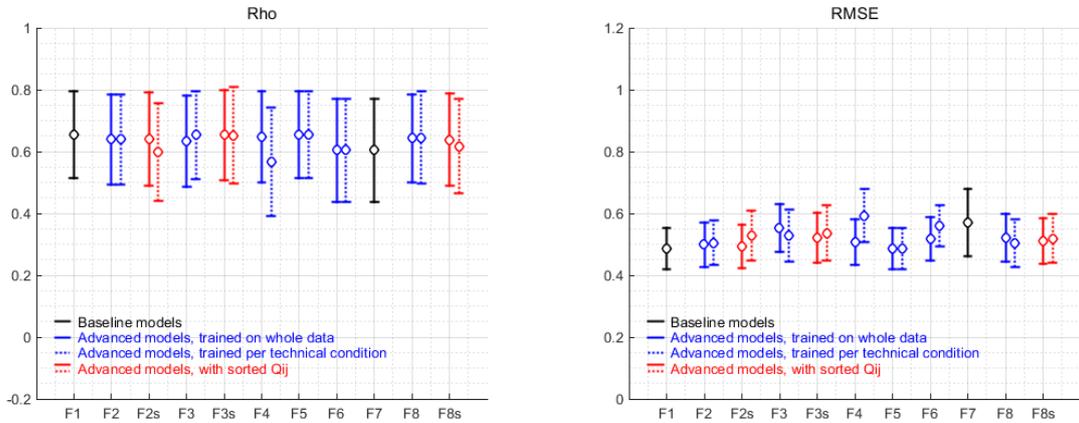
### 4. Observations and Conclusions

- No advanced model outperforms the baseline mean.
- Equal performance for sorting-based (F<sub>2s</sub>, F<sub>3s</sub>, F<sub>8s</sub>) and rule-based (F<sub>2</sub>, F<sub>3</sub>, F<sub>8</sub>) assignment of  $Q_{ij}$ .
- Training per condition does not improve performance, for F<sub>3s</sub> it even reduces performance.

Figure H.38: Modeling performance across all technical conditions of Experiment AVCT<sub>1</sub>.

H.6.2 Performance per technical condition

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

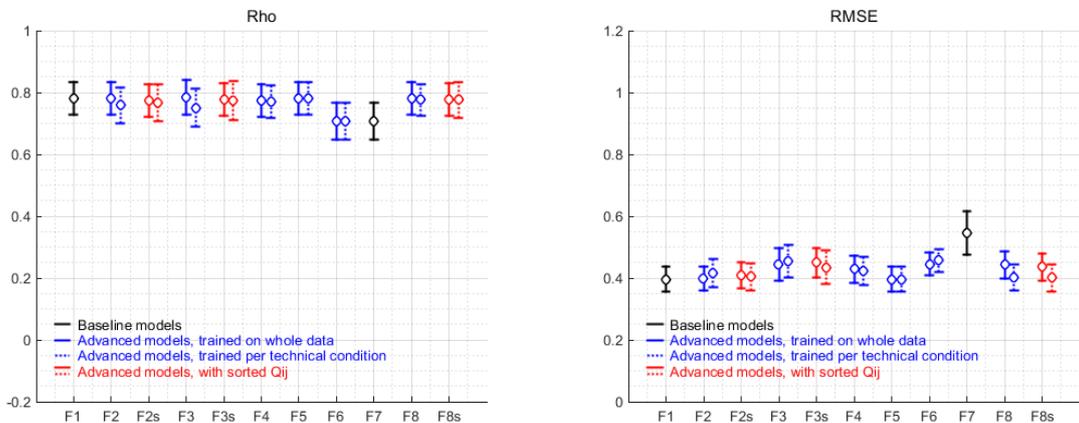
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 slightly better than F7 (not sig.)	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition AVRef

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. F1, F2, F5
			$p = 0.001$	F7 vs. F2s
			$p = 0.015$	F7 vs. F4
			$p = 0.031$	F7 vs. F8s
			$p = 0.001$	F7 vs. F2s, F8, F8s
RMSE	per condition	$p = 0.000$	$p = 0.000$	F7 vs. F1, F5
			$p = 0.001$	F7 vs. F2s, F8, F8s
			$p = 0.005$	F7 vs. F2
			$p = 0.010$	F7 vs. F4
			$p = 0.041$	F7 vs. F3s

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly better than F7 (not sig.)	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

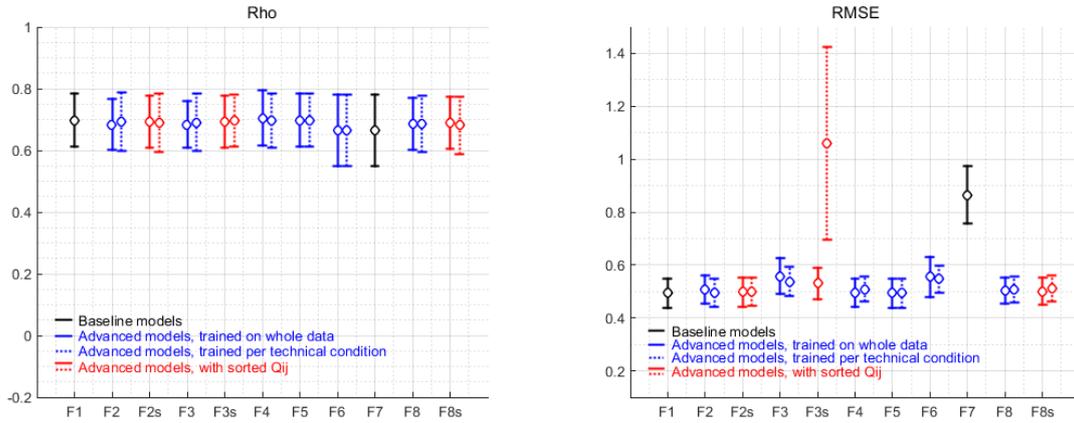
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition VF1

Figure H.39: Modeling performance for conditions AVRef (top panel) and VF1 (bottom panel) of Experiment AVCT1.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. all other
RMSE	per condition	$p = 0.000$	$p = 0.000$	F3s vs. all other
			$p = 0.001$	F7 vs. F1, F2, F2s, F5
			$p = 0.002$	F7 vs. F4, F8, F8s
			$p = 0.007$	F7 vs. F3
			$p = 0.010$	F7 vs. F6

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

RMSE F3s:  $p = 0.005$

3. Behavior of modeling functions

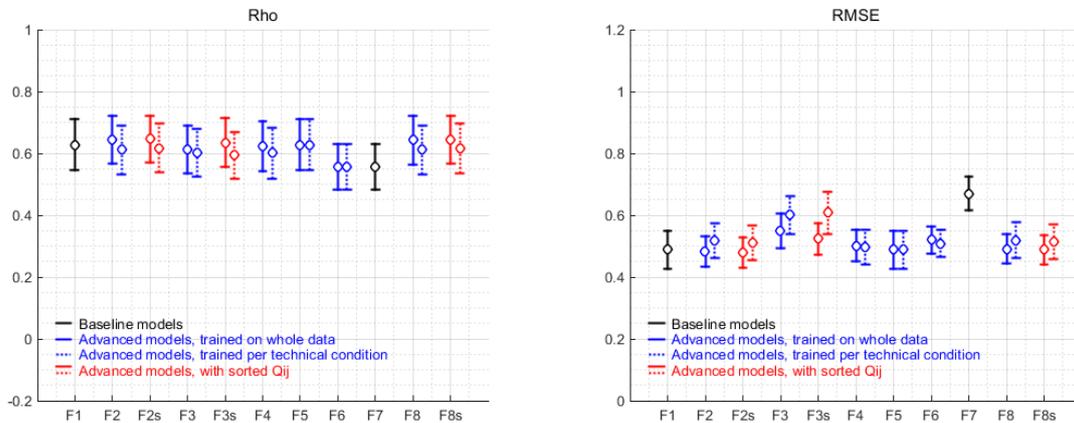
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly better than F7 (not sig.)	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance, in case of RMSE for F3s it even reduces performance.

Condition VR

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. F1, F2, F2s, F4, F5, F8, F8s
			$p = 0.002$	F7 vs. F6
			$p = 0.003$	F7 vs. F3s
			$p = 0.040$	F7 vs. F3
RMSE	per condition	$p = 0.000$	$p = 0.000$	F7 vs. F1, F5
			$p = 0.001$	F7 vs. F4
			$p = 0.003$	F7 vs. F6
			$p = 0.004$	F7 vs. F2s
			$p = 0.005$	F7 vs. F8s
			$p = 0.007$	F7 vs. F2
			$p = 0.008$	F7 vs. F8

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

RMSE F3s:  $p = 0.047$

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly better than F7 (not sig.)	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

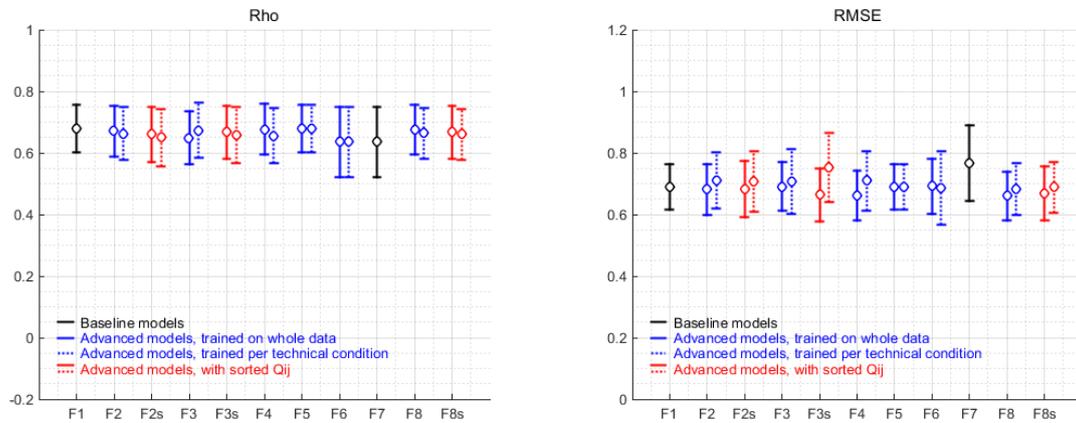
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance, in case of RMSE for F3s it even reduces performance.

Condition FM2

Figure H.40: Modeling performance for conditions VR (top panel) and FM2 (bottom panel) of Experiment AVCT1.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

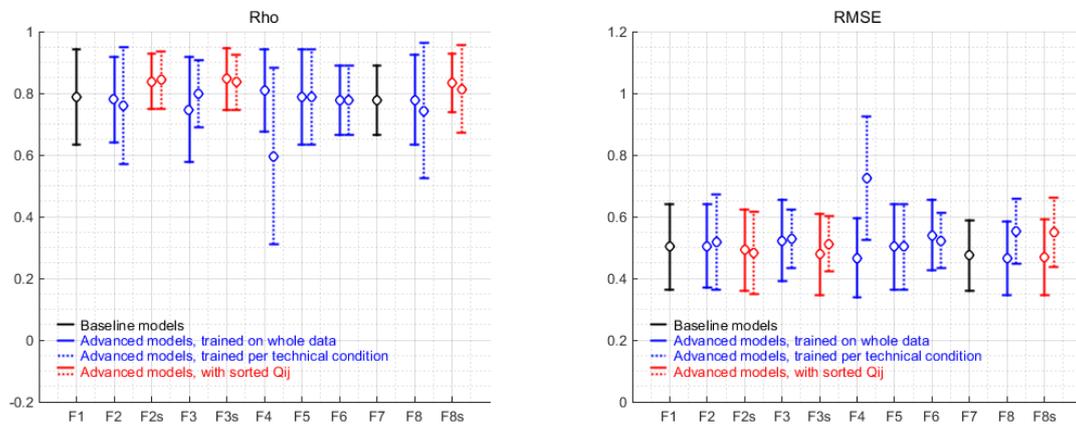
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition E2

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

RMSE F4:  $p = 0.029$

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho, RMSE	F1 essentially equal to F7.	No one outperforms F1.

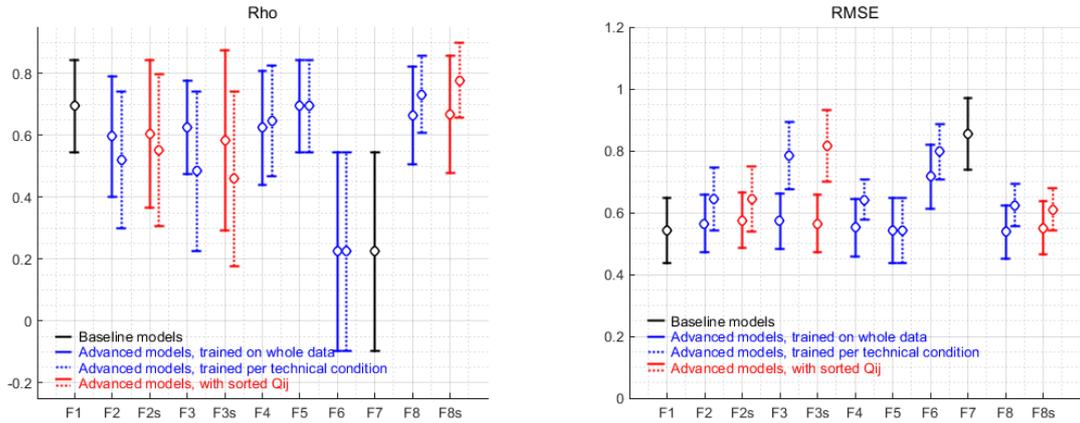
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance, for F4 it leads to reduced and unstable performance (large confidence intervals).

Condition FM2-VF1

Figure H.41: Modeling performance for conditions E2 (top panel) and FM2-VF1 (bottom panel) of Experiment AVCT1.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
Rho	whole data	$p = 0.004$	$p = 0.080$ (n.s.)	F1, F5 vs. F6, F7
Rho	per condition	$p = 0.001$	$p = 0.032$ $p = 0.044$	F6, F7 vs. F8s
RMSE	whole data	$p = 0.000$	$p = 0.000$ $p = 0.001$ $p = 0.002$	F7 vs. F1, F4, F5, F8, F8s F7 vs. F2, F3, F3s F7 vs. F2s
RMSE	per condition	$p = 0.000$	$p = 0.000$ $p = 0.003$ $p = 0.009$ $p = 0.018$ $p = 0.014$ $p = 0.031$	F1, F5 vs. F7 F1, F5 vs. F3s F1, F5 vs. F6 F1, F5 vs. F3 F7 vs. F8s F7 vs. F8

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

Measure	T-test	Functions	Measure	T-test	Functions
RMSE	$p = 0.003$	F3	RMSE	$p = 0.001$	F3s

3. Behavior of modeling functions

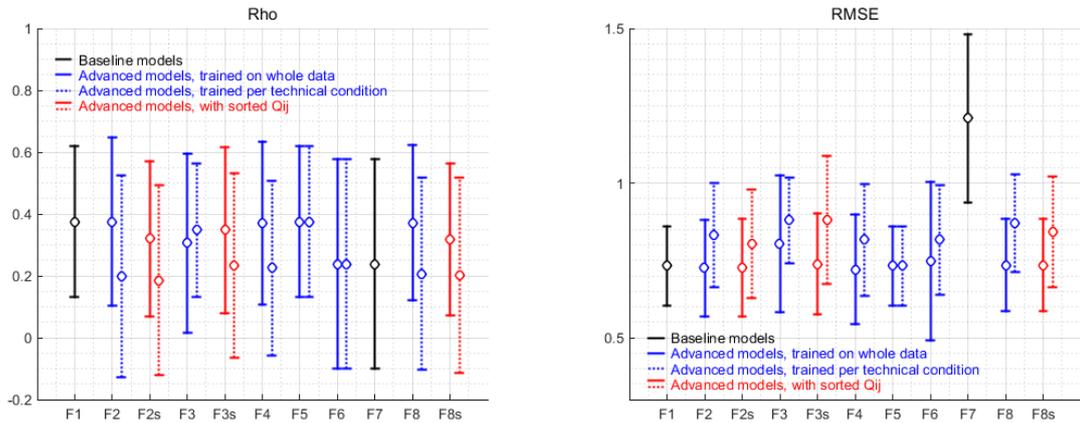
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly better than F7 (not sig.)	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance, in case of RMSE for F3 & F3s it even reduces performance.

Condition E2-VF1

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.005$	$p = 0.006$ $p = 0.008$ $p = 0.009$ $p = 0.010$ $p = 0.011$	F7 vs. F4 F7 vs. F2, F2s F7 vs. F1, F5 F7 vs. F8, F8s F7 vs. F3s
RMSE	per condition	$p = 0.016$	$p = 0.015$ $p = 0.006$ $p = 0.049$	F7 vs. F6 F7 vs. F1, F5 F7 vs. F2s

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

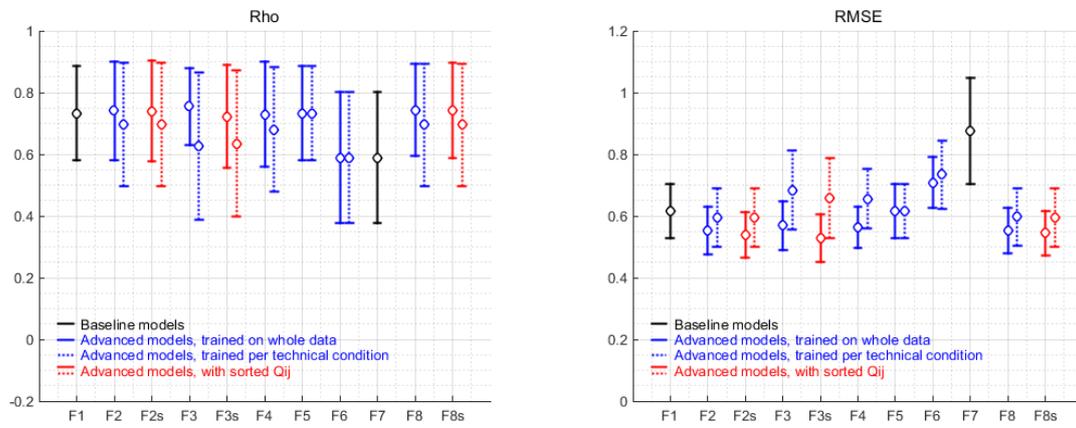
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition FM2-VR

Figure H.42: Modeling performance for conditions E2-VF1 (top panel) and FM2-VR (bottom panel) of Experiment AVCT1.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions	
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. F2, F2s, F3, F3s, F4, F8, F8s	
				$p = 0.002$	F7 vs. F1, F5
				$p = 0.013$	F7 vs. F2, F2s
				$p = 0.014$	F7 vs. F8s
RMSE	per condition	$p = 0.006$	$p = 0.015$	F7 vs. F8	
			$p = 0.036$	F7 vs. F, F5	

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly better than F7 (not sig.)	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance, for F3 & F3s it even leads to unstable performance (large confidence intervals).

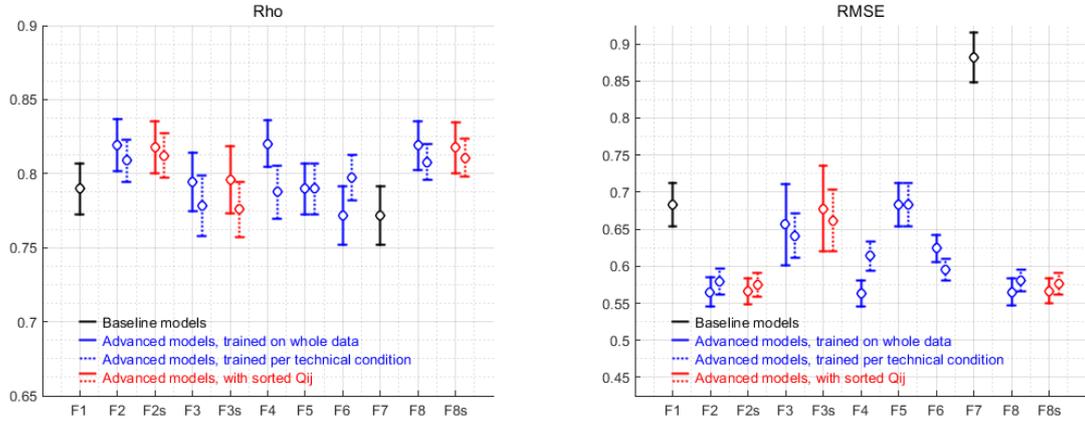
Condition E2-VR

Figure H.43: Modeling performance for condition E2-VR of Experiment AVCT1.

## H.7 Experiment AVCT<sub>2</sub>

### H.7.1 Performance across technical conditions

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F<sub>1</sub> - F<sub>8s</sub>: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
Rho	whole data	P = 0.000	p = 0.007	F6, F7 vs. F4
			p = 0.010	F6, F7 vs. F8
			p = 0.011	F6, F7 vs. F2
			p = 0.015	F6, F7 vs. F2s
Rho	per condition	p = 0.000	p = 0.017	F6, F7 vs. F8s
			p = 0.024	F7 vs. F2s
			p = 0.037	F7 vs. F8
RMSE	whole data	p = 0.000	p = 0.000	F7 vs. all other
			p = 0.000	F1, F3s, F5 vs. F2, F2s, F4, F8, F8s
			p = 0.001	F3 vs. F4
			p = 0.002	F3 vs. F2, F8
RMSE	per condition	p = 0.000	p = 0.003	F3 vs. F2s, F8s
			p = 0.000	F7 vs. all other
			p = 0.000	F1, F5 vs. F2, F2s, F6, F8, F8s
			p = 0.005	F1, F5 vs. F4
			p = 0.000	F3s vs. F2, F2s, F8, F8s
			p = 0.009	F3s vs. F6
			p = 0.008	F3 vs. F2s
			p = 0.011	F3 vs. F8s
p = 0.022	F3 vs. F2			
			p = 0.028	F3 vs. F8

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

Measure	T-test	Functions	Measure	T-test	Functions
Rho	p = 0.007	F4	Rho	p = 0.041	F6
RMSE	P = 0.000	F4	RMSE	p = 0.016	F6

### 3. Behavior of modeling functions

Measure	Baseline (F <sub>1</sub> & F <sub>7</sub> )	Advanced (F <sub>2</sub> -F <sub>6</sub> , F <sub>8</sub> , F <sub>8s</sub> )
Rho	F <sub>1</sub> slightly better than F <sub>7</sub> (not sig.)	No one outperforms F <sub>1</sub> . F <sub>2</sub> , F <sub>2s</sub> , F <sub>8</sub> , F <sub>8s</sub> and F <sub>4</sub> trained on whole data perform slightly better (not sig.)
RMSE	F <sub>1</sub> sig. better than F <sub>7</sub> .	F <sub>2</sub> , F <sub>2s</sub> , F <sub>4</sub> , F <sub>8</sub> , F <sub>8s</sub> and F <sub>6</sub> trained per condition outperform F <sub>1</sub> .

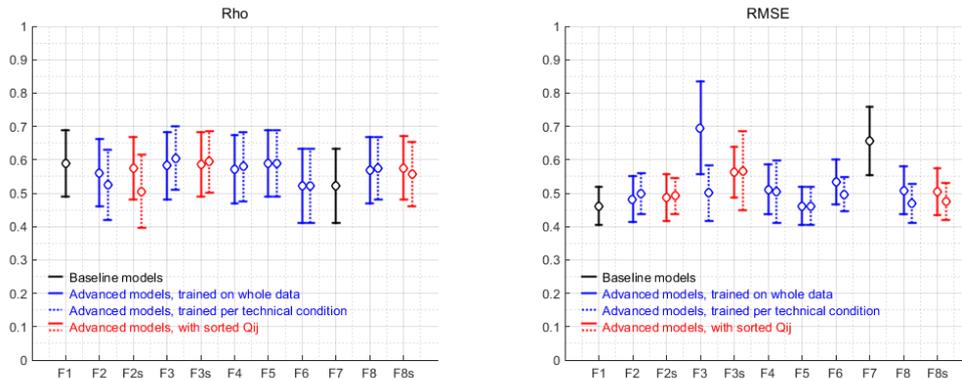
### 4. Observations and Conclusions

- Six advanced models outperform the baseline mean.
- Equal performance for sorting-based (F<sub>2s</sub>, F<sub>3s</sub>, F<sub>8s</sub>) and rule-based (F<sub>2</sub>, F<sub>3</sub>, F<sub>8</sub>) assignment of Q<sub>ij</sub>.
- Training per condition improves performance for F<sub>6</sub> and reduces performance for F<sub>4</sub>.

Figure H.44: Modeling performance across all technical conditions of Experiment AVCT<sub>2</sub>.

H.7.2 Performance per technical condition

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.026$	F7 vs. F1, F5
			$p = 0.002$	F3 vs. F1, F5
			$p = 0.008$	F3 vs. F2
			$p = 0.010$	F3 vs. F2s
RMSE	per condition	$p = 0.007$	$p = 0.034$	F3 vs. F8s
			$p = 0.042$	F3 vs. F8
			$p = 0.010$	F7 vs. F1, F5
			$p = 0.017$	F7 vs. F8
			$p = 0.026$	F7 vs. F8s

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for: RMSE F3:  $p = 0.017$

3. Behavior of modeling functions

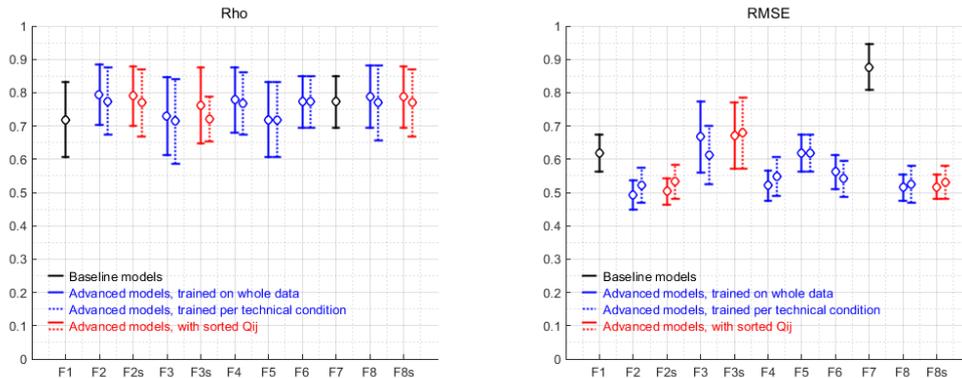
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition improves performance only for F3.

Condition AVRef

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. all other
			$p = 0.003$	F3 vs. F2
			$p = 0.009$	F3 vs. F2s
			$p = 0.024$	F3 vs. F8
			$p = 0.029$	F3 vs. F8s
			$p = 0.041$	F3 vs. F4
			$p = 0.002$	F3s vs. F2
			$p = 0.006$	F3s vs. F2s
			$p = 0.016$	F3s vs. F8
			$p = 0.019$	F3s vs. F8s
RMSE	per condition	$p = 0.000$	$p = 0.028$	F3s vs. F4
			$p = 0.001$	F7 vs. F3s
			$p = 0.000$	F7 vs. all other
			$p = 0.029$	F3s vs. F2
			$p = 0.039$	F3s vs. F8

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly better than F7 (not sig.)	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1. F2, F2s, F4, F6, F8 & F8s perform slightly better (not sig.)

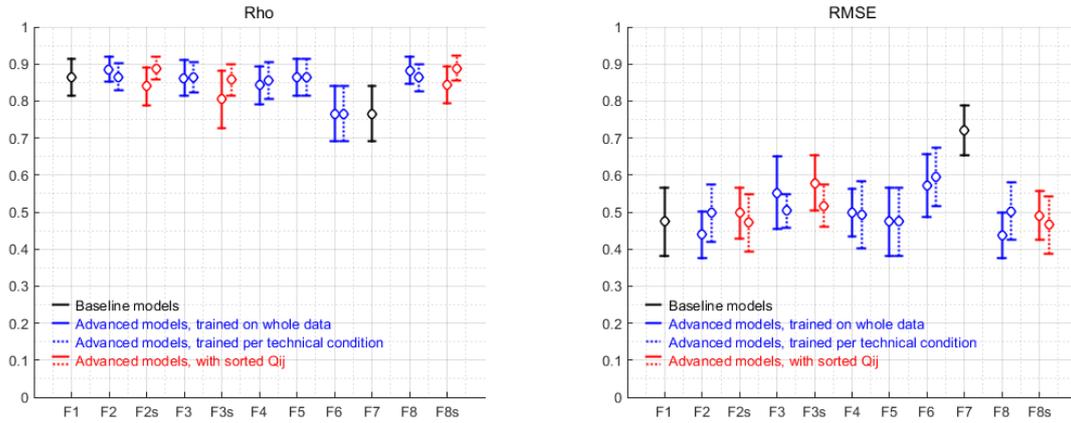
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean. Six models are slightly better (not sig.)
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance, in case of RMSE for F4 it even reduces performance.

Condition VF2

Figure H.45: Modeling performance for conditions AVRef (top panel) and VF2 (bottom panel) of Experiment AVCT2.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
Rho	whole data	$p = 0.009$	$p = 0.100$ (n.s.)	F6, F7 vs. F2
Rho	per condition	$p = 0.001$	$p = 0.016$	F6, F7 vs. F2s, F8s
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. F1, F2, F5, F8
			$p = 0.001$	F7 vs. F2s, F8s
			$p = 0.002$	F7 vs. F4
RMSE	per condition	$p = 0.000$	$p = 0.000$	F7 vs. F1, F2s, F5, F8s
			$p = 0.001$	F7 vs. F4
			$p = 0.002$	F7 vs. F2, F3, F8
			$p = 0.007$	F7 vs. F3s

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

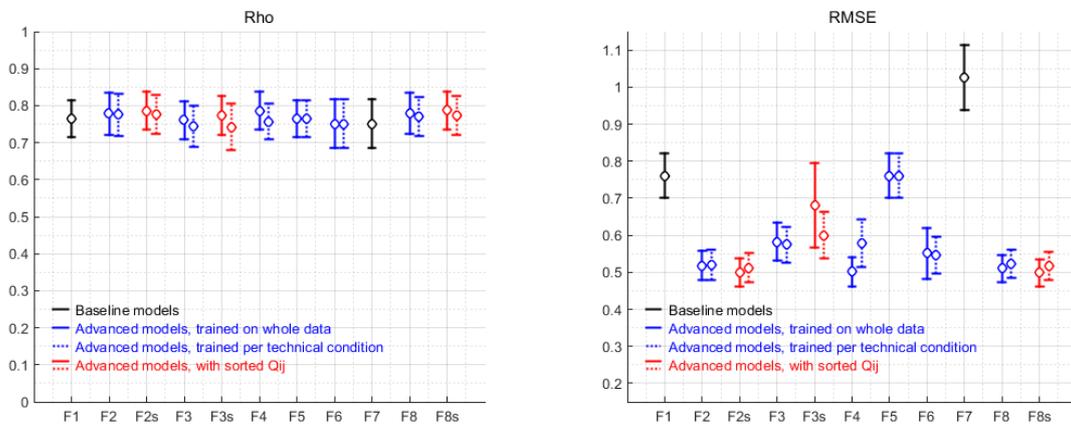
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly better than F7 (not sig.)	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition VF2sym

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. all other
			$p = 0.000$	F1, F5 vs. F2, F2s, F4, F6, F8, F8s
			$p = 0.002$	F1, F5 vs. F3
			$p = 0.001$	F3s vs. F2s, F8s
			$p = 0.002$	F3s vs. F4
			$p = 0.004$	F3s vs. F8
			$p = 0.008$	F3s vs. F2
RMSE	per condition	$p = 0.000$	$p = 0.000$	F7 vs. all other
			$p = 0.000$	F1, F5 vs. F2, F2s, F3, F4, F6, F8, F8s
			$p = 0.002$	F1, F5 vs. F3s

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 sig. better than F7.	F2, F2s, F3, F6, F8, F8s and F3s trained per condition outperform F1.

4. Observations and Conclusions

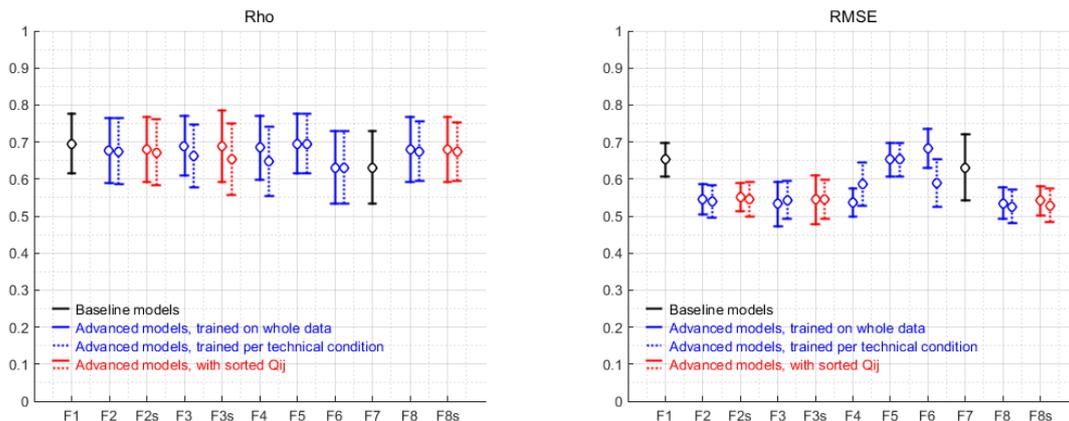
- A) Seven advanced models outperform the baseline mean. F4 (trained on whole data) is the best, though not sig. different from the other six.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance, in case of RMSE for F4 it even reduces performance.

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for: RMSE F4:  $p = 0.040$

Condition VB

Figure H.46: Modeling performance for conditions VF2sym (top panel) and VB (bottom panel) of Experiment AVCT2.

**1. Visualizing the results:** Errorbar plots showing the mean values & 95% confidence intervals.



**2.1 Testing for significant differences between functions F1 - F8s:** Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.002$	F6 vs. F3, F8
			$p = 0.003$	F6 vs. F4
			$p = 0.005$	F6 vs. F8s
			$p = 0.007$	F6 vs. F3s
			$p = 0.008$	F6 vs. F2
RMSE	per condition	$p = 0.000$	$p = 0.015$	F6 vs. F2s
			$p = 0.039$	F1, F5 vs. F8

**2.2 Testing for differences between two training modes:** T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

RMSE F6:  $p = 0.023$

**3. Behavior of modeling functions**

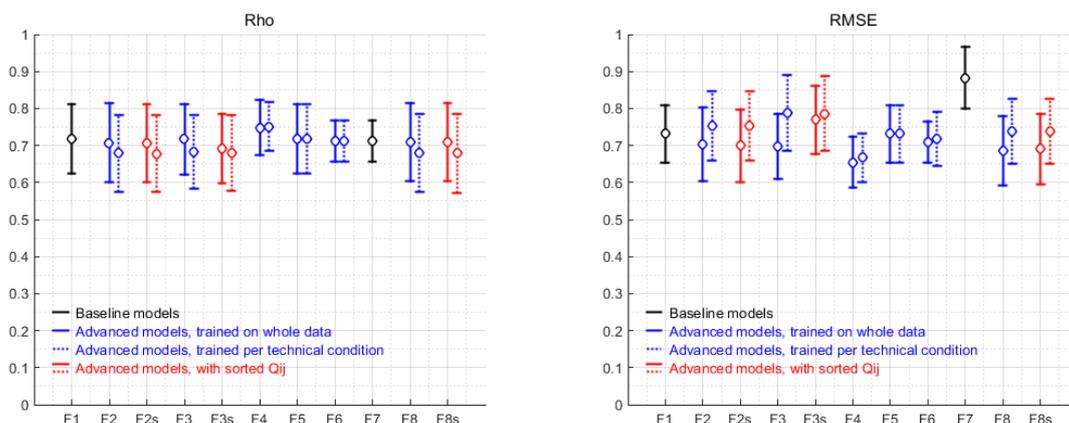
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 essentially equal to F7.	F8 trained per condition outperforms F1 but not F7. F2, F2s, F4 & F8s perform slightly better than F1 (not sig.)

**4. Observations and Conclusions**

- A) One advanced model (f8 trained on whole data) sig. outperforms the baseline mean but not the baseline minimum.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition improves performance only for F6.

**Condition VBsym**

**1. Visualizing the results:** Errorbar plots showing the mean values & 95% confidence intervals.



**2.1 Testing for significant differences between functions F1 - F8s:** Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.020$	$p = 0.006$	F7 vs. F4
			$p = 0.046$	F7 vs. F8
RMSE	per condition	$p = 0.086$ (n.s.)	$p = 0.017$	F7 vs. F4

**2.2 Testing for differences between two training modes:** T-tests, significance level  $p \leq 0.05$ . No significant differences found.

**3. Behavior of modeling functions**

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 slightly better than F7.	No one outperforms F1. (not sig.)

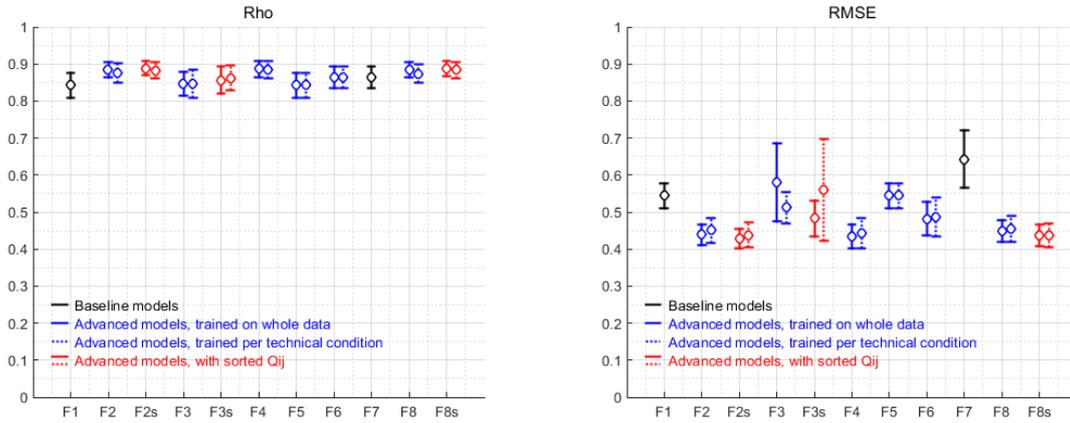
**4. Observations and Conclusions**

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

**Condition E2**

Figure H.47: Modeling performance for conditions VBsym (top panel) and E2 (bottom panel) of Experiment AVCT2.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
Rho	whole data	$p = 0.029$	$p = 0.516$ (n.s.)	F1, F5 vs. F2s
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. F2, F2s, F3s, F4, F6, F8, F8s
			$p = 0.040$	F1, F5 vs. F2s
			$p = 0.001$	F3 vs. F2s, F4
			$p = 0.002$	F3 vs. F8s
			$p = 0.003$	F3 vs. F2
			$p = 0.008$	F3 vs. F8
RMSE	per condition	$p = 0.000$	$p = 0.000$	F7 vs. F2, F2s, F4, F8, F8s
			$p = 0.007$	F7 vs. F6

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

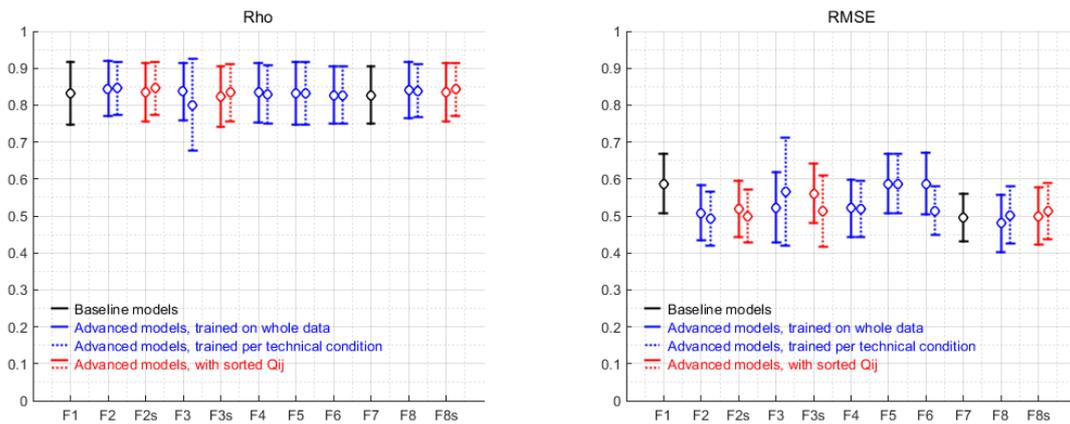
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly worse than F7 (not sig.)	No one outperforms F1.
RMSE	F1 sig. better than F7.	F2s outperforms F1. F2, F4, F8 & F8s perform slightly better than F1 (not sig.)

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition stabilizes performance of F3 (smaller confidence interval), but destabilizes performance of F3s (large confidence interval).

Condition FM4

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 slightly worse than F7 (not sig.)	No one outperforms F1 nor F7.

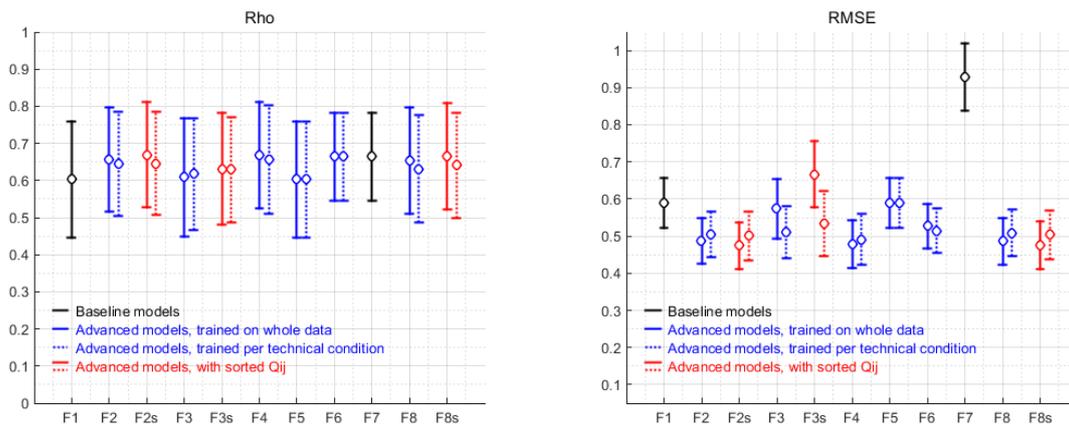
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean nor minimum.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition FM4sym

Figure H.48: Modeling performance for conditions FM4 (top panel) and FM4sym (bottom panel) of Experiment AVCT2.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. all other
			$p = 0.005$	F3s vs. F2s, F8s
			$p = 0.007$	F3s vs. F4
			$p = 0.012$	F3s vs. F8
RMSE	per condition	$p = 0.000$	$p = 0.013$	F3s vs. F2
			$p = 0.000$	F7 vs. all other

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

RMSE F3s:  $p = 0.032$

3. Behavior of modeling functions

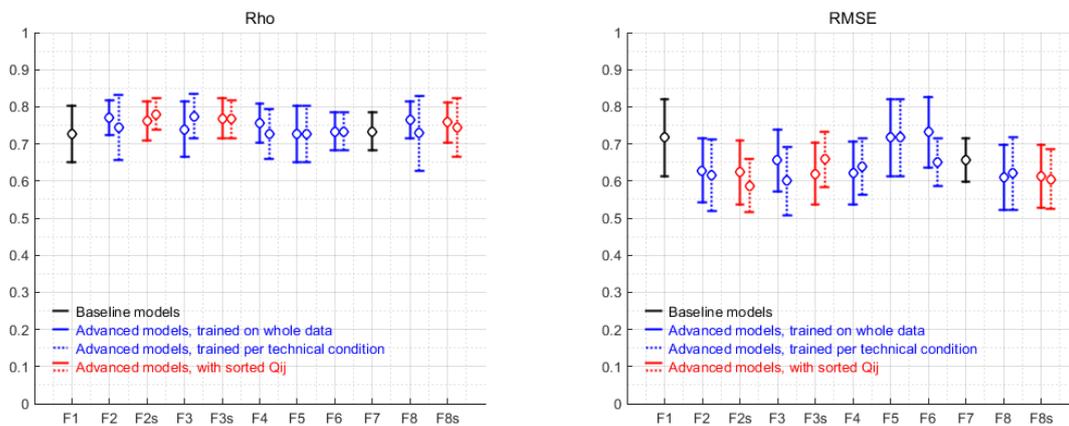
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1. F2, F2s, F4, F8, & F8s perform slightly better than F1 (not sig.)

4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition FM4P3

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . No significant differences found.

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 slightly worse than F7 (not sig.)	No one outperforms F1 nor F7.

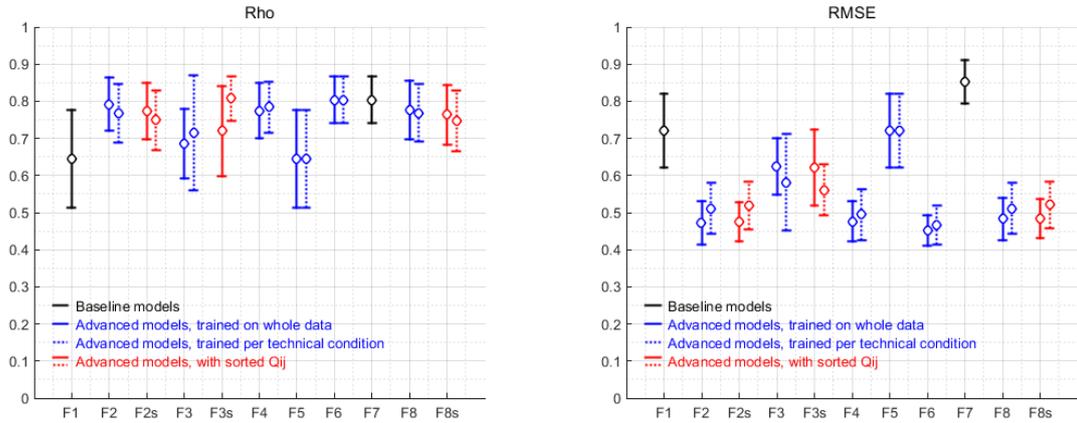
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean nor minimum.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance, for F3 and F3s it even leads to unstable performance (large confidence intervals).

Condition FM4P3sym

Figure H.49: Modeling performance for conditions FM4P3 (top panel) and FM4P3sym (bottom panel) of Experiment AVCT2.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions	
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. F2, F2s, F3, F3s, F4, F6, F8, F8s	
				F1, F5 vs. F2, F2s, F4, F6, F8, F8s	
RMSE	per condition	$p = 0.000$	$p = 0.000$	F7 vs. F2, F2s, F3, F3s, F4, F6, F8, F8s	
				F1, F5 vs. F6	
				$p = 0.002$	F1, F5 vs. F4
				$p = 0.007$	F1, F5 vs. F2, F8
				$p = 0.013$	F1, F5 vs. F2s
			$p = 0.014$	F1, F5 vs. F8s	

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

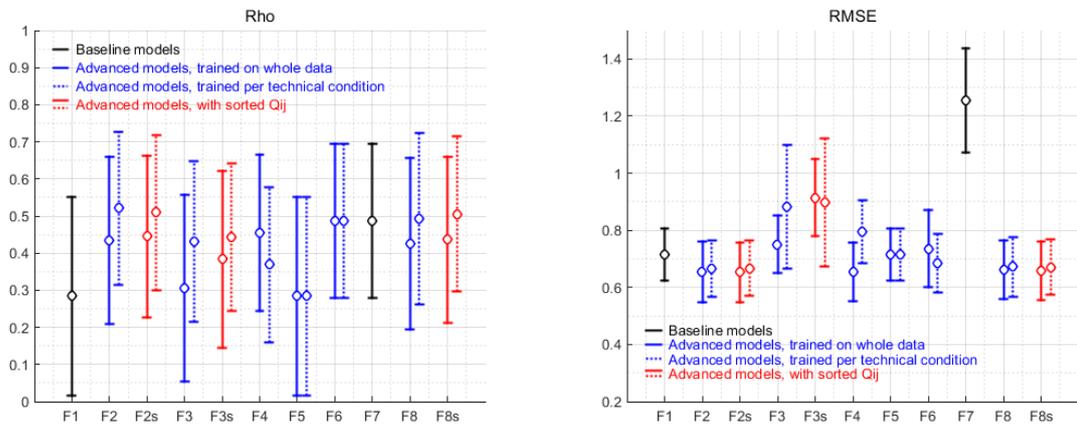
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly worse than F7 (not sig.)	No one outperforms F1 nor F7.
RMSE	F1 slightly better than F7 (not sig.)	F2, F2s, F4, F6, F8 & F8s outperform F1.

4. Observations and Conclusions

- A) Six advanced models outperform the baseline mean. F6 is the best, though not sig. different from the other five.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition FM4-AVP3

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions	
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. F1, F2, F2s, F3, F4, F5, F6, F8, F8s	
				F7 vs. F3s	
RMSE	per condition	$p = 0.000$	$p = 0.000$	F7 vs. F1, F2, F2s, F4, F5, F6, F8, F8s	
				$p = 0.005$	F7 vs. F3
				$p = 0.009$	F7 vs. F3s

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly worse than F7 (not sig.)	No one outperforms F1 or F7.
RMSE	F1 sig. better than F7.	No one outperforms F1.

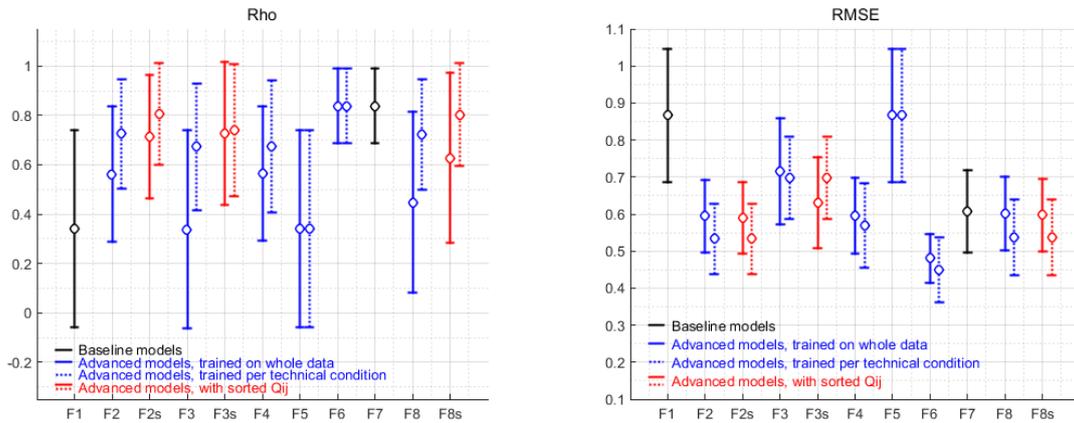
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance.

Condition FM4-VF2

Figure H.50: Modeling performance for conditions FM4-AVP3 (top panel) and FM4-VF2 (bottom panel) of Experiment AVCT2.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F1, F5 vs. F6
RMSE	per condition	$p = 0.000$	$p = 0.000$	F1, F5 vs. F6
			$p = 0.004$	F1, F5 vs. F2, F2s
			$p = 0.005$	F1, F5 vs. F8, F8s
			$p = 0.020$	F1, F5 vs. F4

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

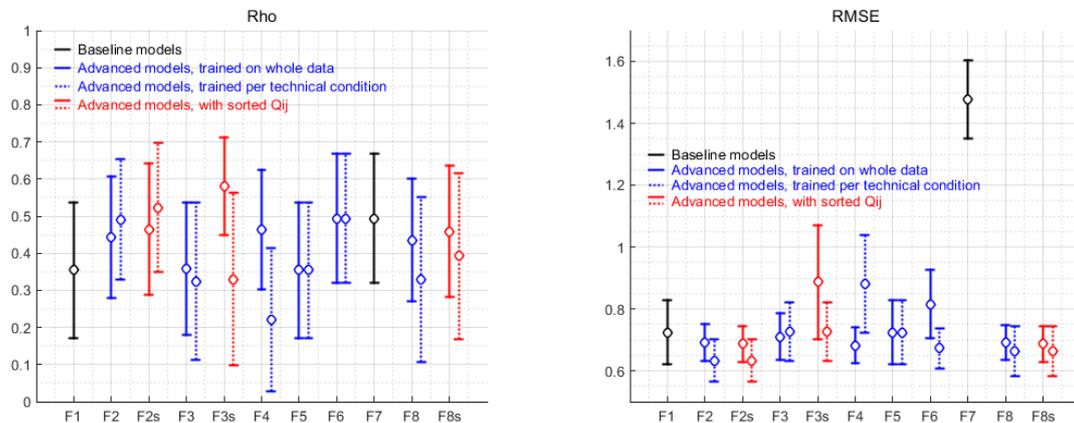
Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly worse than F7 (not sig.)	No one outperforms F1.
RMSE	F1 slightly worse than F7 (not sig.)	F2, F2s, F4, F8 & F8s (all trained per condition) and F6 outperform F1, but no one F7.

4. Observations and Conclusions

- A) Six advanced model outperform the baseline mean, but not the baseline minimum.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance, but it stabilizes Rho for F3 (smaller confidence interval).

Condition FM4P3-VF2

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. all other
RMSE	per condition	$p = 0.000$	$p = 0.000$	F7 vs. all other
			$p = 0.013$	F4 vs. F2, F2s

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . Significant differences found for:

Measure	T-test	Function	Measure	T-test	Functions
Rho	$p = 0.050$ (n.s)	F4	Rho	$p = 0.029$	F3s
RMSE	$p = 0.017$	F4	RMSE	$p = 0.025$	F6

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 slightly worse than F7 (not sig.)	No one outperforms F1.
RMSE	F1 sig. better than F7.	No one outperforms F1.

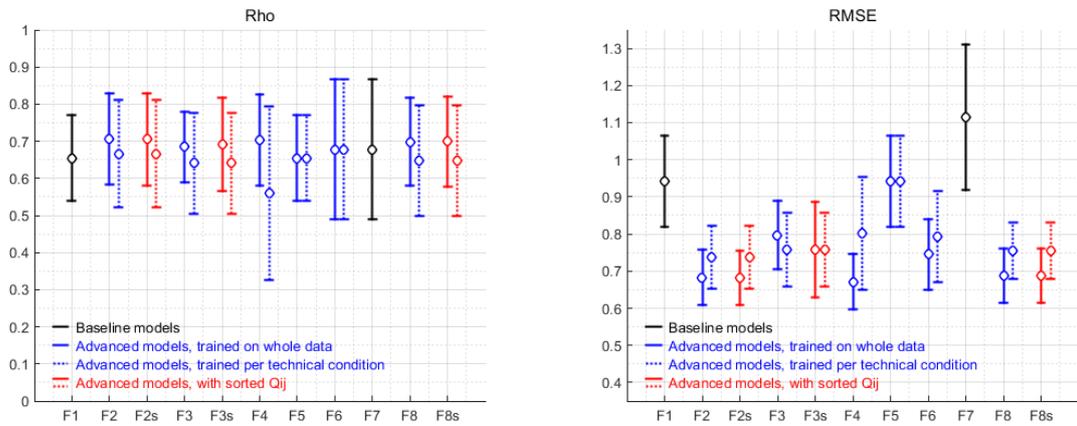
4. Observations and Conclusions

- A) No advanced model outperforms the baseline mean.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition improves performance for F6 (RMSE), reduces performance for F4 (RMES) and has a mixed effect on F3 (better RMSE, worse Rho).

Condition FM4-VB

Figure H.51: Modeling performance for conditions FM4P3-VF2 (top panel) and FM4-VB (bottom panel) of Experiment AVCT2.

1. Visualizing the results: Errorbar plots showing the mean values & 95% confidence intervals.



2.1 Testing for significant differences between functions F1 - F8s: Oneway ANOVAs & PostHoc tests with Sidak correction, tests conducted separately for two training modes (whole data, per condition), significance level  $p \leq 0.05$ . Significant differences found for:

Measure	Training	ANOVA	PostHoc	Functions					
RMSE	whole data	$p = 0.000$	$p = 0.000$	F7 vs. F2, F2s, F3s, F4, F6, F8, F8s					
				$p = 0.001$	F7 vs. F3				
				$p = 0.015$	F1, F5 vs. F4				
				$p = 0.026$	F1, F5 vs. F2s				
				$p = 0.027$	F1, F5 vs. F2				
				$p = 0.032$	F1, F5 vs. F8s				
				$p = 0.034$	F1, F5 vs. F8				
				RMSE	per condition	$p = 0.000$	$p = 0.000$	F7 vs. F2, F2s	
								$p = 0.001$	F7 vs. F3, F3s, F8, F8s
								$p = 0.004$	F7 vs. F6
$p = 0.006$	F7 vs. F4								

2.2 Testing for differences between two training modes: T-tests, significance level  $p \leq 0.05$ . No significant differences found.

3. Behavior of modeling functions

Measure	Baseline (F1 & F7)	Advanced (F2-F6, F8, F8s)
Rho	F1 essentially equal to F7.	No one outperforms F1.
RMSE	F1 slightly better than F7 (not sig.)	F2, F2s, F4, F8 & F8s (all trained on whole data) outperform F1.

4. Observations and Conclusions

- A) Five advanced models outperform the baseline mean. F4 is the best, though not sig. different from the other four.
- B) Equal performance for sorting-based (F2s, F3s, F8s) and rule-based (F2, F3, F8) assignment of  $Q_{ij}$ .
- C) Training per condition does not improve performance, for F4 it even leads to unstable performance (large confidence intervals).

Condition FM4P3-VB

Figure H.52: Modeling performance for condition FM4P3-VB of Experiment AVCT2.

# Bibliography

- 3GPP. *TS 26.071 version 12.0.0 Release 12 - Mandatory speech CODEC speech processing functions; AMR speech CODEC, General description*. Standard. Sophia Antipolis Cedex, France: 3rd Generation Partnership Project, 2014.
- Adel, Mohamed et al. "Improved E-Model for Monitoring Quality of Multi-Party VoIP communications". In: *Proceedings of the IEEE Globecom Workshops*. Atlanta, USA, 2013, pp. 1180–1185.
- Anderson, Anne H. et al. "The HCRC Map Task Corpus". In: *Language and Speech* 34.4 (1991), pp. 351–366.
- Anderson, Anne H. et al. "Virtual team meetings: An analysis of communication and context". In: *Computers in Human Behavior* 23 (2007), pp. 2558–2580.
- Apperley, Mark and Masood Masoodian. "An Experimental Evaluation of Video Support for Shared Work-Space Interaction". In: *Proceedings of CHI 1995*. 1995.
- Bachl, Maximilian. "Impact of packet loss on localization and subjective quality in 3D-telephone calls". Bachelor Thesis. Quality and Usability Lab, Technische Universität Berlin, Germany, 2014.
- Baldis, Jessica J. "Effects of Spatial Audio on Memory, Comprehension, and Preference during Desktop Conferences". In: *Proceedings of the ACM CHI 2001 Human Factors in Computing Systems Conference*. Ed. by Michel Beaudouin-Lafon and Robert J. K. Jacob. Vol. 3. 1. 2001, pp. 166–173.
- Belmudez, Benjamin. *Assessment and Prediction of Audiovisual Quality for Videotelephony*. Springer, 2009.
- Berndtsson, Gunilla, Mats Folkesson, and Valentin Kulyk. "Subjective quality assessment of video conferences and telemeetings". In: *19th International Packet Video Workshop*. 2012.
- Blaskovich, Jennifer L. "Exploring the Effect of Distance: An Experimental Investigation of Virtual Collaboration, Social Loafing, and Group Decisions". In: *Journal of Information Systems* 22.1 (2008), pp. 27–46.
- Blauert, Jens. *Spatial Hearing – The Psychophysics of Human Sound Localization – Revised edition*. Harvard, MA, USA: The MIT Press, 1997.
- Blauert, Jens, ed. *The Technology of Binaural Listening*. Berlin, Germany: Springer, 2013.
- Bodden, Markus and Ute Jekosch. *Entwicklung und Durchführung von Tests mit Versuchspersonen zur Verifizierung von Modellen zur Berech-*

- nung der Sprachübertragungsqualität. Project Report (unpublished). Institute of Communication Acoustics, Ruhr-University Bochum, Germany, 1996.
- Bovik, Al, ed. *The Essential Guide to Image Processing*. London, UK: Elsevier Inc., 2009.
- Braden, R. *Requirements for Internet Hosts – Communication Layers*. International Standard. Internet Engineering Task Force (IETF), 1989. URL: <https://tools.ietf.org/rfc/rfc1122.txt>.
- Bradner, Erin and Gloria Mark. "Why Distance Matters: Effects on Cooperation, Persuasion and Deception". In: *Proceedings of CSCW 2002*. 2002.
- Brady, Paul T. "A statistical analysis of on-off patterns in 16 conversations". In: *Bell Syst. Tech. J.* 47.1 (1968), pp. 73–99.
- Brady, Paul T. "A technique for investigating on-off patterns of speech". In: *Bell Syst. Tech. J.* 44.1 (1965), pp. 1–22.
- Brady, Paul T. "Effects of transmission delay on conversational behavior on echo-free telephone circuits". In: *Bell Syst. Tech. J.* 50.1 (1971), pp. 115–134.
- Brosig, Jeannette, Joachim Weimann, and Axel Ockenfels. "The Effect of Communication Media on Cooperation". In: *German Economic Review* 4.2 (2003), pp. 217–241.
- Bruenken, Roland, Jan L. Plass, and Detlev Leutner. "Direct Measurement of Cognitive Load in Multimedia Learning". In: *Educational Psychologist* 38.1 (2003), pp. 53–61.
- Burgoon, Judee K. et al. "Trust and Deception in Mediated Communication". In: *Proceedings of the 36th Hawaii International Conference on System Sciences*. 2003.
- Buxton, William A. S. "Telepresence: integrating shared task and person spaces". In: *Proceedings of Graphics Interface '92*. Amsterdam, The Netherlands, Oct. 1992, pp. 123–129.
- Carletta, Jean et al. *HCRC dialogue structure coding manual*. Report. University of Edinburgh, UK, 1996. URL: <http://www.lancaster.ac.uk/fass/projects/eagles/maptask.htm>.
- Carletta, Jean et al. "The reliability of a dialogue structure coding scheme". In: *Computational Linguistics* 23.1 (1997), pp. 13–31.
- Champagne, Catherine et al. "A bootstrap method for assessing classification accuracy and confidence for agricultural land use mapping in Canada". In: *International Journal of Applied Earth Observation and Geoinformation* 29 (2014), pp. 44–52.
- Chen, Milton. "Conveying Conversational Cues Through Video". PhD Thesis. Stanford University, 2003.
- Chiu, Stephen. "Extracting fuzzy rules from data for function approximation and pattern classification". In: *Fuzzy Information Engineering: A Guided Tour of Applications*. Ed. by Didier Dubois, Henri Prade, and Ronald R. Yager. John Wiley&Sons, 1997.
- Chu, Wai C. *Speech Coding Algorithms - Foundation and Evolution of Standardized Coders*. Hoboken, New Jersey, USA: Wiley, 2014.
- Clark, Herbert H. and Susan E. Brennan. "Grounding in Communication". In: *Perspectives on socially shared cognition*. Ed. by Lauren

- B. Resnick, John M. Levine, and Stephanie D. Teasley. American Psychological Association, 1991, pp. 127–149.
- Connell, Joanie B. et al. "Effects of communication medium on interpersonal perceptions". In: *Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work*. ACM. 2001, pp. 117–124.
- Crede, Marcus and Janet A. Sniezek. "Group judgment processes and outcomes in video-conferencing versus face-to-face groups". In: *International Journal of Human-Computer Studies* 59 (2003), pp. 875–897.
- Daly-Jones, Owen, Andrew Monk, and Leon Watts. "Some advantages of video conferencing over high-quality audio conferencing: fluency and awareness of attentional focus". In: *International Journal of Human-Computer Studies* 49 (1998), pp. 21–58.
- Doerry, Eckehard. "An Empirical Comparison of Copresent and Technologically-Mediated Interaction based on Communicative Breakdown". PhD Thesis. Department of Computer and Information Science, Graduate School of the University of Oregon, 1995.
- Doherty-Sneddon, Gwyneth et al. "Face-to-Face and Video-Mediated Communication: A Comparison of Dialogue Structure and Task Performance". In: *Journal of Experimental Psychology: Applied* 3.2 (1997), pp. 105–125.
- Efron, Bradley. "Bootstrap Methods: Another Look at the Jackknife". In: *The Annals of Statistics* 7.1 (1979), pp. 1–26.
- Eg, Ragnhild et al. "Audiovisual robustness: exploring perceptual tolerance to asynchrony and quality distortion". In: *Multimedia Tools and Applications* 74 (2015), pp. 345–365.
- Egger, Sebastian, Raimund Schatz, and Stefan Scherer. "It takes two to tango – assessing the impact of delay on conversational interactivity on perceived speech quality". In: *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech 2010)*. Makuhari, Japan, 2010, pp. 1321–1324.
- Engdahl, Tomi. *Powering Microphones*. retrieved on: August 13, 2015. 2012. URL: [http://www.epanorama.net/circuits/microphone\\_powering.html](http://www.epanorama.net/circuits/microphone_powering.html).
- Erickson, Thomas et al. "Telepresence in Virtual Conferences: An Empirical Comparison of Distance Collaboration Technologies". In: *Proceedings of CSCW 2010*. 2010.
- European Network on Quality of Experience in Multimedia Systems and Services. *Qualinet White Paper on Definitions of Quality of Experience*. White Paper. Version Version 1.2. Patrick Le Callet, Sebastian Möller and Andrew Perkis (Eds.) Lausanne, Switzerland: (COST Action IC 1003), Mar. 2013.
- Field, Andy. *Discovering Statistics using SPSS*. 3rd ed. SAGE publications, 2009.
- Firestone, Scott, Thiya Ramalingam, and Steve Fry. *Voice and Video Conferencing Fundamentals*. Cisco Press, 2007.

- Fjermestad, Jerry. "An analysis of communication mode in group support systems research". In: *Decision Support Systems* 37 (2004), pp. 239–263.
- Fjermestad, Jerry and Starr Roxanne Hiltz. "An Assessment of Group Support Systems Experimental Research: Methodology and Results". In: *Journal of Management Information Systems* 15.3 (1999), pp. 7–149.
- Fjermestad, Jerry and Starr Roxanne Hiltz. "Case and field studies of group support systems: an empirical assessment". In: *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*. IEEE. 2000, pp. 10–19.
- Fjermestad, Jerry and Starr Roxanne Hiltz. "Experimental studies of group decision support systems: an assessment of variables studied and methodology". In: *Proceedings of the 30th Annual Hawaii International Conference on System Sciences*. Vol. 2. IEEE. 1997, pp. 45–65.
- Fussell, Susan R. and Nicholas I. Benimoff. "Social and Cognitive Processes in Interpersonal Communication: Implications for advanced telecommunicatinos technologies". In: *Human Factors* 37 (1995), pp. 228–250.
- Fussell, Susan R., Robert E. Kraut, and Jane Siegel. "Coordination of Communication: Effects of Shared Visual Context on Collaborative Work". In: *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. 2000.
- Fussell, Susan R., Leslie D. Setlock, and Robert E. Kraut. "Effects of Head-Mounted and Scene-Oriented Video Systems on Remote Collaboration on Physical Tasks". In: *Proceedings of the Conference on human factors in computing systems CHI 03*. 2003.
- Garcia, Marie-Neige. *Parametric Packet-based Audiovisual Quality model for IPTV services*. Springer, 2014.
- Geier, Matthias, Jens Ahrens, and Sascha Spors. "The soundscape renderer: A unified spatial audio reproduction framework for arbitrary rendering methods". In: *Proceedings of the AES 124th Convention*. Amsterdam, The Netherlands, 2008.
- Gergle, Darren, Robert E. Kraut, and Susan R. Fussell. "Language Efficiency and Visual Technology – Minimizing Collaborative Effort with Visual Information". In: *Journal of Language and Social Psychology* 23.4 (Dec. 2004), pp. 1–27.
- Gierlich, Hans W. "Wideband Speech Communication – The Quality Parameters as Perceived by the User". In: *Proceedings of Forum Acusticum 2005*. 2005.
- Grayson, David M. and Adrew F. Monk. "Are You Looking at Me? Eye Contact and Desktop Video Conferencing". In: *ACM Transactions on Computer-Human Interaction* 10.3 (Sept. 2003), pp. 221–243.
- Gros, Laetitia and Gontran Filandre. *A study on tasks for assessing audiovisual quality of videoconferencing systems in multipoint conversation tests*. ITU-T Contribution COM 12 - C 259 - E. Geneva, Switzerland: International Telecommunication Union, Oct. 2011.

- Group, Moving Pictures Expert. *MPEG Standards*. retrieved on August 19, 2015. International Organization for Standardization & International Electrotechnical Commission (ISO/IEC). 2013. URL: <http://mpeg.chiariglione.org/standards>.
- Guéguin, Marie et al. "On the evaluation of the conversational speech quality in telecommunications". In: *Eurasip Journal on Applied Signal Processing* (2008), pp. 185–248.
- Hammer, Florian, Peter Reichl, and Alexander Raake. "Elements of interactivity in telephone conversations". In: *Proceedings of the 8th International Conference on Spoken Language Processing (Interspeech 2004)*. Jeju Island, Korea, 2004, pp. 1741–1744.
- Handgraaf, Michel J. J. et al. "Web-conferencing as a viable method for group decision research". In: *Judgment and Decision Making* 7.5 (2012), pp. 659–668.
- Hanzo, Lajos, Peter J. Cherriman, and Jürgen Streit. *Video Compression and Communications – From Basics to H.261, H.263, H.264, MPEG4 for DVB and HSDPA-Style Adaptive Turbo-Transceivers*. Chichester, West Sussex, UK: Wiley, 2007.
- Hoeldtke, Katrin and Alexander Raake. "Conversation analysis of multi-party conferencing and its relation to perceived quality". In: *Proceedings of the IEEE International Conference on Communications ICC*. Kyoto, Japan, June 2011.
- Hyder, Mansoor, Michael Haun, and Christian Hoene. "Placing the participants of a spatial audio conference call". In: *Proceedings of the 7th IEEE Consumer Communications and Networking Conference (CCNC)*. IEEE. 2010, pp. 1–7.
- ISO/IEC. *ISO/IEC 11172-3:1993 - Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s – Part 3: Audio*. International Standard. International Organization for Standardization & International Electrotechnical Commission, 1993.
- ISO/IEC. *ISO/IEC 13818-7:1995 - Information technology – Generic coding of moving pictures and associated audio information – Part 2: Video*. International Standard. International Organization for Standardization & International Electrotechnical Commission, 2013.
- ISO/IEC. *ISO/IEC 13818-7:1995 - Information technology – Generic coding of moving pictures and associated audio information – Part 3: Audio*. International Standard. International Organization for Standardization & International Electrotechnical Commission, 1995.
- ISO/IEC. *ISO/IEC 13818-7:2006 - Information technology – Generic coding of moving pictures and associated audio information – Part 7: Advanced Audio Coding (AAC)*. International Standard. International Organization for Standardization & International Electrotechnical Commission, 2006.
- ISO/IEC. *ISO/IEC 14496-2:2004 - Information technology – Coding of audio-visual objects – Part 2: Visual*. International Standard. International Organization for Standardization & International Electrotechnical Commission, 2004.

- ISO/IEC. *ISO/IEC 7498-1:1994 - Information technology – Open Systems Interconnection – Basic Reference Model: The Basic Model*. International Standard. International Organization for Standardization & International Electrotechnical Commission, 1994.
- ITU-R. *Recommendation BS.1116-1 - Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*. International Standard. Geneva, Switzerland: International Telecommunication Union, 1997.
- ITU-R. *Recommendation BS.1284-1 - General methods for the subjective assessment of sound quality*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2003.
- ITU-R. *Recommendation BS.1285 - Pre-selection methods for the subjective assessment of small impairments in audio systems*. International Standard. Geneva, Switzerland: International Telecommunication Union, 1997.
- ITU-R. *Recommendation BS.1387-1 - Method for objective measurements of perceived audio quality*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2001.
- ITU-R. *Recommendation BS.1534-1 - Method for the subjective assessment of intermediate quality levels of coding systems*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2003.
- ITU-R. *Recommendation BT.1788 - Methodology for the subjective assessment of video quality in multimedia applications*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2007.
- ITU-R. *Recommendation BT.500-13 - Methodology for the subjective assessment of the quality of television pictures*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2012.
- ITU-R. *Recommendation BT.601-7 - Studio encoding parameters of digital television for standard 4:3 and wide screen 16:9 aspect ratios*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2011.
- ITU-R. *Recommendation BT.710-4 - Subjective assessment methods for image quality in high-definition television*. International Standard. Geneva, Switzerland: International Telecommunication Union, 1998.
- ITU-T. *Handbook on Telephony*. Geneva, Switzerland: International Telecommunication Union, 1992.
- ITU-T. *ITU-R Recommendations – BS-Series: Broadcasting service (sound)*. retrieved on August 19, 2015. International Telecommunication Union. 2015. URL: <https://www.itu.int/rec/R-REC-BS/en>.
- ITU-T. *ITU-R Recommendations – BT-Series: Broadcasting service (television)*. retrieved on August 19, 2015. International Telecommunication Union. 2015. URL: <https://www.itu.int/rec/R-REC-BT/en>.
- ITU-T. *ITU-T Recommendations by series*. retrieved on August 19, 2015. International Telecommunication Union. 2015. URL: <http://www.itu.int/ITU-T/recommendations/index.aspx>.
- ITU-T. *ITU-T Recommendations – P-Series: Terminals and subjective and objective assessment methods*. retrieved on August 19, 2015. Interna-

- tional Telecommunication Union. 2015. URL: <https://www.itu.int/rec/T-REC-P/en>.
- ITU-T. *P-Series Supplement P.Sup26 - Scenarios for the subjective quality evaluation of audio and audiovisual multiparty telemeetings*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2012.
- ITU-T. *Question 10/12 - Conferencing and telemeeting assessment*. retrieved on August 19, 2015. International Telecommunication Union. 2015. URL: <http://www.itu.int/en/ITU-T/studygroups/2013-2016/12/pages/q10.aspx>.
- ITU-T. *Recommendation E.800 - Definitions of terms related to quality of service*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2008.
- ITU-T. *Recommendation G.1011 - Reference guide to quality of experience assessment methodologies*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2015.
- ITU-T. *Recommendation G.107 - The E-model: a computational model for use in transmission planning*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2014.
- ITU-T. *Recommendation G.1070 - Opinion model for video-telephony applications*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2012.
- ITU-T. *Recommendation G.107.1 - Wideband E-model*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2015.
- ITU-T. *Recommendation G.113 - Transmission impairments due to speech processing*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2007.
- ITU-T. *Recommendation G.114 - One-way transmission time*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2003.
- ITU-T. *Recommendation G.191 - Software tools for speech and audio coding standardization*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2010.
- ITU-T. *Recommendation G.711 - Pulse Code Modulation (PCM) of voice frequencies*. International Standard. Geneva, Switzerland: International Telecommunication Union, 1988.
- ITU-T. *Recommendation G.722 - 7 kHz audio-coding within 64 kbit/s*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2012.
- ITU-T. *Recommendation G.722.2 - Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2003.
- ITU-T. *Recommendation G.726 - 40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM)*. International Standard. Geneva, Switzerland: International Telecommunication Union, 1990.
- ITU-T. *Recommendation G.729 - Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)*. Interna-

- tional Standard. Geneva, Switzerland: International Telecommunication Union, 2012.
- ITU-T. *Recommendation H.263 - Video coding for low bit rate communication*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2005.
- ITU-T. *Recommendation H.264 - Advanced video coding for generic audiovisual services*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2014.
- ITU-T. *Recommendation H.265 - High efficiency video coding*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2015.
- ITU-T. *Recommendation J.144 - Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2004.
- ITU-T. *Recommendation J.246 - Perceptual visual quality measurement techniques for multimedia services over digital cable television networks in the presence of a reduced bandwidth reference*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2008.
- ITU-T. *Recommendation J.247 - Objective perceptual multimedia video quality measurement in the presence of a full reference*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2008.
- ITU-T. *Recommendation J.341 - Objective perceptual multimedia video quality measurement of HDTV for digital cable television in the presence of a full reference*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2011.
- ITU-T. *Recommendation P.1201 - Parametric non-intrusive assessment of audiovisual media streaming quality*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2012.
- ITU-T. *Recommendation P.1202 - Parametric non-intrusive bitstream assessment of video media streaming quality*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2012.
- ITU-T. *Recommendation P.1301 - Subjective quality evaluation of audio and audiovisual telemeetings*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2012.
- ITU-T. *Recommendation P.562 - Analysis and interpretation of INMD voice-service measurements*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2004.
- ITU-T. *Recommendation P.563 - Single-ended method for objective speech quality assessment in narrow-band telephony applications*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2004.
- ITU-T. *Recommendation P.564 - Conformance testing for voice over IP transmission quality assessment models*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2007.
- ITU-T. *Recommendation P.800 - Methods for objective and subjective assessment of quality*. International Standard. Geneva, Switzerland: International Telecommunication Union, 1996.

- ITU-T. *Recommendation P.805 - Subjective evaluation of conversational quality*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2007.
- ITU-T. *Recommendation P.806 - A Subjective Quality Test Methodology using Multiple Rating Scales*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2014.
- ITU-T. *Recommendation P.833 - Methodology for derivation of equipment impairment factors from subjective listening-only tests*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2001.
- ITU-T. *Recommendation P.833.1 - Methodology for the derivation of equipment impairment factors from subjective listening-only tests for wideband speech codecs*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2009.
- ITU-T. *Recommendation P.835 - Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2003.
- ITU-T. *Recommendation P.851 - Subjective quality evaluation of telephone services based on spoken dialogue systems*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2003.
- ITU-T. *Recommendation P.862 - Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2001.
- ITU-T. *Recommendation P.862.2 - Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2007.
- ITU-T. *Recommendation P.863 - Perceptual objective listening quality assessment*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2014.
- ITU-T. *Recommendation P.880 - Continuous evaluation of time-varying speech quality*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2004.
- ITU-T. *Recommendation P.910 - Subjective video quality assessment methods for multimedia applications*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2008.
- ITU-T. *Recommendation P.911 - Subjective audiovisual quality assessment methods for multimedia applications*. International Standard. Geneva, Switzerland: International Telecommunication Union, 1998.
- ITU-T. *Recommendation P.920 - Interactive test methods for audiovisual communications*. International Standard. Geneva, Switzerland: International Telecommunication Union, 2000.
- Jackson, Matthew et al. "Impact of video frame rate on communicative behaviour in two and four party groups". In: *Proceedings of the 2000 ACM conference on Computer Supported Cooperative Work*. ACM. 2000, pp. 11–20.

- Jekosch, Ute. "Sprache hören und beurteilen: Ein Ansatz zur Grundlegung der Sprachqualitätsbeurteilung". Habilitation thesis. Universität/Gesamthochschule Essen, Germany, 2003.
- Jekosch, Ute. *Voice and Speech Quality Perception – Assessment and Evaluation*. Berlin, Germany: Springer, 2005.
- Kilgore, Ryan, Mark Chignell, and Paul Smith. "Spatialized audio-conferencing: what are the benefits?" In: *Proceedings of the 2003 conference of the Centre for Advanced Studies on Collaborative research CASCON '03*. IBM Press, 2003, pp. 135–144.
- Knapp, Mark L. and Judith A. Hall. *Nonverbal communication in human interaction*. 7th edition. Boston, USA: Wadsworth, Cengage Learning, 2010.
- Korhonen, Jari. "New methods for robust audio streaming in a wireless environment". PhD Thesis. Tampere University of Technology, Finland, 2006.
- Köster, Friedemann and Sebastian Möller. "Perceptual Speech Quality Dimensions in a Conversational Situation". In: *Proceedings of the 16th Annual Conference of the Speech Communication Association (Interspeech)*. Dresden, Germany, Sept. 2015, pp. 2544–2548.
- Köster, Friedemann et al. "Comparison between the Discrete ACR Scale and an Extended Continuous Scale for the Quality Assessment of Transmitted Speech". In: *Fortschritte der Akustik (DAGA2015) - 41. Jahrestagung für Akustik*. Nürnberg, Germany, Mar. 2015.
- Kraut, Robert E., Susan R. Fussell, and Jane Siegel. "Visual Information as a Conversational Resource in Collaborative Physical Tasks". In: *Human-Computer Interaction* 18 (2003), pp. 13–49.
- Kyriacou, E.C., C.S. Pattichis, and M.S. Pattichis. "An overview of recent health care support systems for eEmergency and mHealth applications". In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2009)*. IEEE. 2009, pp. 1246–1249.
- Lamblin, Claude. "Algebraic code-excited linear prediction speech coding method". Pat. US 5717825 A (US). France Telecom. Feb. 10, 1995.
- Light, Richard J. and David B. Pillemer. *Summing Up*. Cambridge, MA, USA: Harvard University Press, 1984.
- Löber, Andreas, Sibylle Grimm, and Gerhard Schwabe. "Audio vs. chat: Can media speed explain the differences in productivity?" In: *Proceedings of the European Conference on Information Systems*. 2006, pp. 2172–2183.
- Löber, Andreas, Gerhard Schwabe, and Sibylle Grimm. "Audio vs. chat: The effects of group size on media choice". In: *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*. IEEE. 2007, pp. 41–41.
- Masoodian, Masood. "Human-to-Human Communication Support for Computer-Based Shared Workspace Collaboration". PhD Thesis. Department of Computer Science, The University of Waikato, Hamilton, New Zealand, 1996.

- Masoodian, Masood and Mark Apperley. "User perceptions of human-to-human communication modes in CSCW environments". In: *Proceedings of ED-MEDIA'95*. 1995, pp. 17–21.
- Masoodian, Masood, Mark Apperley, and Lesley Frederickson. "Video support for shared work-space interaction: an empirical study". In: *Interacting with Computers* 7.3 (1995), pp. 237–253.
- Matarazzo, Giacinto. "The effects of video quality and task on remote collaboration: a laboratory experiment". In: *17th International Symposium on Human Factors in Telecommunication*. Copenhagen, Denmark, May 1999.
- McGrath, Joseph Edward. *Groups: Interaction and performance*. Prentice-Hall Englewood Cliffs, NJ, 1984.
- McLeod, Poppy Lauretta et al. "The Eyes Have It: Minority Influence in Face-To-Face and Computer-Mediated Group Discussion". In: *Journal of Applied Psychology* 82.5 (1997), pp. 706–718.
- Merriam-Webster. *Dictionary and Thesaurus*. retrieved on August 19, 2015. Merriam-Webster Incorporated, Springfield, MA. 2015. URL: <http://www.merriam-webster.com/>.
- Möller, Sebastian. *Assessment and Prediction of Speech Quality in Telecommunications*. Kluwer Academic publishers, 2000.
- Möller, Sebastian. *Quality of Telephone-Based Spoken Dialogue Systems*. New York, USA: Springer, 2005.
- Möller, Sebastian and Alexander Raake. "Motivation and Introduction". In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 3–9.
- Möller, Sebastian, Marcel Wältermann, and Marie-Neige Garcia. "Features of Quality of Experience". In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 73–84.
- Möller, Sebastian et al. "A Taxonomy of Quality of Service and Quality of Experience of Multimodal Human-Machine-Interaction". In: *Proceedings of the International Workshop on Quality of Multimedia Experience QoMEX*. 2009.
- Möller, Sebastian et al. "Speech Quality Estimation". In: *IEEE Signal Processing Magazine* 28.6 (Nov. 2011), pp. 18–28.
- Moore, Brian C. J. *Introduction to the Psychology of Hearing*. 6th ed. Leiden, The Netherlands: Brill, 2013.
- Nahumi, Dror. "Conferencing arrangement for compressed information signals". Pat. US 5390177 A (US). At&T Corp. Feb. 14, 1993.
- Nakamura, Junichi, ed. *Image Sensors and Signal Processing for Digital Still Cameras*. Boca Raton, Florida, USA: Taylor & Francis Group, 2006.
- Naylor, Patrick A. and Nikolay D. Gaubitch, eds. *Speech Dereverberation*. Berlin, Germany: Springer, 2010.
- Nguyen, David and John Canny. "MultiView: Spatially Faithful Group Video Conferencing". In: *Proceedings of CHI 2005*. 2005.

- Norman, Donald A. "Some Observations on Mental Models". In: *Mental Models*. Ed. by Dedre Genter and Albert L. Stevens. Psychology Press, 1983, pp. 7–14.
- O'Brien, Heather L. and Elaine G. Toms. "What is user engagement? A conceptual framework for defining user engagement with technology". In: *Journal of the American Society for Information Science & Technology* 59.6 (2008), pp. 938–955.
- O'Conaill, Brid, Steve Whittaker, and Sylvia Wilbur. "Conversations Over Video Conferences: An Evaluation of the Spoken Aspects of Video-Mediated Communication". In: *Human-Computer Interaction* 8 (1993), pp. 389–428.
- Olson, Gary M. and Judith S. Olson. "Distance Matters". In: *Human-Computer Interaction* 15 (2000), pp. 139–178.
- O'Malley, Claire et al. "Comparison of face-to-face and video-mediated interaction". In: *Interacting with Computers* 8.2 (1996), pp. 177–192.
- Paepcke, Andreas et al. "Yelling in the hall: using sidetone to address a problem with mobile remote presence systems". In: *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 2011, pp. 107–116.
- Park, Kyoung Shin. "Enhancing Cooperative Work in Amplified Collaboration Environments". PhD Thesis. Graduate College of the University of Illinois at Chicago, 2003.
- Pashler, Harold E. *The psychology of attention*. Cambridge, MA, USA: The MIT Press, 1998.
- Perkins, C., O. Hodson, and V. Hardman. "A survey of packet loss recovery techniques for streaming audio". In: *IEEE Network Magazine* 11.5 (1998), pp. 40–48.
- Pinson, Margaret and Stephen Wolf. *Techniques for evaluating objective video quality models using overlapping subjective data sets*. Technical Report TR-09-457. US Department of Commerce, National Telecommunications and Information Administration (NTIA), 2008.
- Pulkki, Ville and Matti Karjalainen. *Communication Acoustics: An Introduction to Speech, Audio and Psychoacoustics*. Chichester, West Sussex, UK: The MIT Press, 2015.
- Raake, Alexander. *Speech Quality of VoIP – Assessment and Prediction*. Chichester, West Sussex, UK: Wiley, 2006.
- Raake, Alexander and Sebastian Egger. "Quality and Quality of Experience". In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 11–34.
- Raake, Alexander and Claudia Schlegel. *3CT - 3-party Conversation Test scenarios for conferencing assessment*. 2010. DOI: 10.5281/zenodo.16129.
- Raake, Alexander et al. "IP-Based Mobile and Fixed Network Audio-visual Media Services". In: *IEEE Signal Processing Magazine* 28.6 (Nov. 2011), pp. 68–79.
- Raake, Alexander et al. "Listening and conversational quality of spatial audio conferencing". In: *Proceedings of the AES 40th International Conference*. Tokyo, Oct. 2010.

- Ramirez, Javier, Juan Manuel Görriz, and José Carlos Segura. "Voice Activity Detection - Fundamentals and Speech Recognition System Robustness". In: *Robust Speech Recognition and Understanding*. Ed. by Michael Grimm and Kristian Kroschel. In-Tech Education and Publishing, 2007.
- Rawat, Paresch and Jyoti Singhai. "Review of Motion Estimation and Video Stabilization techniques For hand held mobile video". In: *Signal and Image Processing : An International Journal (SIPIJ)* 2.2 (June 2011), pp. 159–168.
- Reiter, Ulrich et al. "Factors Influencing Quality of Experience". In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 55–72.
- Richardson, Iain E. G. *H.264 and MPEG-4 Video Compression – Video Coding for Next-generation Multimedia*. Chichester, West Sussex, UK: Wiley, 2003.
- Roegiers, David. "Dynamical Aspects of Spatial-Audio in Multi-participant Teleconferences". Master Thesis. Universiteit Gent, Belgium, 2013.
- Rothbucher, Martin et al. "3D Audio Conference System with Backward Compatible Conference Server using HRTF Synthesis". In: *Journal of Multimedia Processing and Technologies* 2 (4 Dec. 2011), pp. 159–175.
- Sacks, Harvey, Emanuel A. Schegloff, and Gail Jefferson. "A Simplest Systematics for the Organisation of Turn-Taking for Conversation". In: *Language* 50 (4 Dec. 1974), pp. 696–735.
- Sanford, Alison, Anne H. Anderson, and Jim Mullin. "Audio channel constraints in video-mediated communication". In: *Interacting with Computers* 16 (2004), pp. 1069–1094.
- Schiffner, Falk and Janto Skowronek. *3seqCT - 3-party Sequential Conversation Test scenarios for conferencing assessment*. 2013. DOI: 10.5281/zenodo.16137.
- Schnotz, Wolfgang and Christian Kuerschner. "A Reconsideration of Cognitive Load Theory". In: *Educational Psychology Review* 19.4 (2007), pp. 469–508.
- Schoenberg, Katrin. "The Quality of Mediated-Conversations under Transmission Delay". PhD Thesis. Technische Universität Berlin, Germany, 2016.
- Schoenberg, Katrin, Alexander Raake, and Pierre Lebreton. "On interaction behaviour in telephone conversations under transmission delay". In: *Proceedings of the 6th International Workshop on Quality of Multimedia Experience QoMEX*. Sept. 2014, pp. 31–36.
- Schoenberg, Katrin et al. "On interaction behaviour in telephone conversations under transmission delay". In: *Speech Communication* 63 (2014), pp. 1–14.
- Sellen, Abigail J. "Speech Patterns in Video-Mediated Conversations". In: *Proceedings of CHI 1992*. 1992.
- Silzle, Andreas. "Quality Taxonomies for Auditory Virtual Environments". In: *Proceedings of the 122nd AES Convention*. May 2007.

- Singh, Kundan, Gautam Nair, and Henning Schulzrinne. "Centralized conferencing using SIP". In: *Internet Telephony Workshop*. Vol. 7. 2001, pp. 57–63.
- Skowronek, Janto. *3SCT - 3-party Short Conversation Test scenarios for conferencing assessment (Version 01)*. 2013. DOI: 10.5281/zenodo.16134.
- Skowronek, Janto. *Document describing scenarios*. Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, Apr. 2013.
- Skowronek, Janto. *Document summarizing model review*. Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, Apr. 2013.
- Skowronek, Janto. *Documentation on system setup*. Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, June 2013.
- Skowronek, Janto. *Final Project Report – Quality of Multiparty Audio-Video Communication*. Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, Dec. 2014.
- Skowronek, Janto. *Focus groups on quality perception and attribution*. Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, Nov. 2014.
- Skowronek, Janto. *Improved model for audio-only communication*. Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, Dec. 2013.
- Skowronek, Janto. *Improved model for audiovisual communication*. Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, Nov. 2014.
- Skowronek, Janto. *Initial model for audio-only communication based on available data*. Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, June 2013.
- Skowronek, Janto. *Initial model for audiovisual communication*. Project Report "Quality of Multiparty Audio-Video Communication" (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, May 2014.
- Skowronek, Janto. "Internetumfrage zur Wahrnehmung heutiger Telefonkonferenzen im geschäftlichen Umfeld". In: *Fortschritte der Akustik (DAGA2012) - 38. Deutsche Jahrestagung für Akustik*. Darmstadt: Deutsche Gesellschaft für Akustik, 2012, pp. 899–900.
- Skowronek, Janto. *Pilot test for audiovisual communication*. Project Report "Quality of Multiparty Audio-Video Communication" (Un-

- published). Assessment of IP-based Applications, Technische Universität Berlin, Germany, Feb. 2014.
- Skowronek, Janto. *Second test for audiovisual communication*. Project Report “Quality of Multiparty Audio-Video Communication” (Unpublished). Assessment of IP-based Applications, Technische Universität Berlin, Germany, Oct. 2014.
- Skowronek, Janto. *Topics for Question 18/12 in view of insights from a field survey*. ITU-T Contribution COM 12 - C 285 - E. Geneva, Switzerland: International Telecommunication Union, Oct. 2011.
- Skowronek, Janto. *xCT - Scalable Multiparty Conversation Test scenarios for conferencing assessment (Version 01)*. 2015. DOI: 10.5281/zenodo.16138.
- Skowronek, Janto, Julian Herlinghaus, and Alexander Raake. “Quality Assessment of Asymmetric Multiparty Telephone Conferences: A Systematic Method from Technical Degradations to Perceived Impairments”. In: *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*. International Speech Communication Association. Lyon, France, Aug. 2013, pp. 2604–2608.
- Skowronek, Janto, Martin McKinney, and Steven van de Par. “A demonstrator for automatic music mood estimation”. In: *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*. 2007.
- Skowronek, Janto and Alexander Raake. “Assessment of Cognitive Load, Speech Communication Quality and Quality of Experience for spatial and non-spatial audio conferencing calls”. In: *Speech Communication* (2015), pp. 154–175. DOI: 10.1016.
- Skowronek, Janto and Alexander Raake. “Conceptual model of multiparty conferencing and telemeeting quality”. In: *Proceedings of the 7th International Workshop on Quality of Multimedia Experience (QoMEX 2015)*. Pilo, Greece, May 2015, pp. 1–6.
- Skowronek, Janto and Alexander Raake. *Development of scalable conversation test scenarios for multi-party telephone conferences*. ITU-T Contribution COM 12 - C187 - E. Geneva, Switzerland: International Telecommunication Union, Jan. 2011.
- Skowronek, Janto and Alexander Raake. “Einfluss von Bandbreite und räumlicher Sprachwiedergabe auf die kognitive Anstrengung bei Telefonkonferenzen in Abhängigkeit von der Teilnehmeranzahl”. In: *Fortschritte der Akustik (DAGA2011) - 37. Deutsche Jahrestagung für Akustik*. Deutsche Gesellschaft für Akustik. Düsseldorf, Germany, Mar. 2011, pp. 873–874.
- Skowronek, Janto and Alexander Raake. “Investigating the effect of number of interlocutors on the quality of experience for multi-party audio conferencing”. In: *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech2011)*. International Speech Communication Association. Florence, Italy, Aug. 2011, pp. 829–832.
- Skowronek, Janto and Alexander Raake. *Listening test stimuli using scalable multiparty conversation test scenarios*. ITU-T Contribution

- COM 12 - C 286 - E. Geneva, Switzerland: International Telecommunication Union, Oct. 2011.
- Skowronek, Janto and Alexander Raake. *Number of interlocutors and QoE in multiparty telephone conferences*. ITU-T Contribution COM 12 - C 287 - E. Geneva, Switzerland: International Telecommunication Union, Oct. 2011.
- Skowronek, Janto and Alexander Raake. *Report on a study on asymmetric multiparty telephone conferences for a future update of P.1301*. ITU-T Contribution COM 12 - C134 - E. Geneva, Switzerland: International Telecommunication Union, Dec. 2013.
- Skowronek, Janto and Alexander Raake. *Update on number of Interlocutors and QoE in multiparty telephone conferences*. ITU-T Contribution COM 12 - C035 - E. Geneva, Switzerland: International Telecommunication Union, Mar. 2013.
- Skowronek, Janto and Falk Schiffner. *3SCT - 3-party Short Conversation Test scenarios for conferencing assessment (Version 02)*. 2015. DOI: 10.5281/zenodo.16135.
- Skowronek, Janto, Falk Schiffner, and Alexander Raake. "On the influence of involvement on the quality of multiparty conferencing". In: *4th International Workshop on Perceptual Quality of Systems (PQS 2013)*. International Speech Communication Association. Vienna, Austria, Sept. 2013, pp. 141–146.
- Skowronek, Janto and Katrin Schoenenberg. *3CNG - 3-party Celebrity Name Guessing task for Conferencing Assessment*. 2017. DOI: 10.5281/zenodo.345834.
- Skowronek, Janto, Katrin Schoenenberg, and Gunilla Berndtsson. "Multimedia Conferencing and Telemeetings". In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 217–228.
- Skowronek, Janto, Katrin Schoenenberg, and Alexander Raake. "Experience with and insights about the new ITU-T standard on quality assessment of conferencing systems". In: *Proceedings of the international conference of acoustics (AIA-DAGA 2013)*. Merano, Mar. 2013, pp. 444–447.
- Skowronek, Janto and Maxim Spur. *3SurvivalCT - 3-party Survival task Conversation Test scenarios for conferencing assessment*. 2017. DOI: 10.5281/zenodo.345837.
- Skowronek, Janto, Anne Weigel, and Alexander Raake. "Quality of Multiparty Telephone Conferences from the Perspective of a Passive Listener". In: *Fortschritte der Akustik - 41. Jahrestagung für Akustik (DAGA)*. Nürnberg, Germany, Mar. 2015.
- Skowronek, Janto et al. *Initial insights for P.AMT based on reviewing existing recommendations*. ITU-T Contribution COM 12 - C 284 - E. Geneva, Switzerland: International Telecommunication Union, Oct. 2011.
- Skowronek, Janto et al. "Method and Apparatus for Computing the Perceived Quality of a Multiparty Audio or Audiovisual Telecommunication Service or System". Patent application WO/2016/041593 (EP). Deutsche Telekom AG. Mar. 2016.

- Skowronek, Janto et al. *Proposed main body and normative annex for draft recommendation P.AMT*. ITU-T Contribution COM 12 - C318 - E. Geneva, Switzerland: International Telecommunication Union, May 2012.
- Skowronek, Janto et al. *Proposed non-normative appendices for draft recommendation P.AMT*. ITU-T Contribution COM 12 - C319 - E. Geneva, Switzerland: International Telecommunication Union, May 2012.
- Skowronek, Janto et al. "Speech recordings for systematic assessment of multi-party conferencing". In: *Proceedings of Forum Acusticum 2011*. Aalborg, Denmark: European Acoustical Society, June 2011.
- Smith, Paxton J., Peter Kabal, and Rafi Rabipour. "Speaker Selection for Tandem-Free Operation VoIP Conference Bridges". In: *Proceedings of IEEE Workshop on Speech Coding*. IEEE. Tsukuba, Japan, Oct. 2002, pp. 120–122.
- Smith, Paxton J. et al. "Tandem-free VoIP conferencing: A bridge to next-generation networks". In: *IEEE Communications Magazine* 41.5 (2003), pp. 136–145.
- Sonnenwald, Diane H., Kelly L. Maglaughlin, and Mary C. Whitton. "Designing to support situation awareness across distances: an example from a scientific collaboratory". In: *Information Processing and Management* 40 (2004), pp. 989–1011.
- Sonnenwald, Diane H., Mary C. Whitton, and Kelly L. Maglaughlin. "Evaluating a scientific collaboratory: Results of a controlled experiment". In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 10.2 (2003), pp. 150–176.
- Spur, Maxim. "Implementation of a Spatial VoIP Conferencing Demonstrator". Master Thesis. Quality and Usability Lab, Technische Universität Berlin, Germany, 2015.
- Spur, Maxim. "Influences of Loudness Differences in Multiparty Conferencing Quality". Bachelor Thesis. Assessment of IP-based Applications, Technische Universität Berlin, Germany, 2012.
- Straus, Susan G. "Testing a Typology of Tasks: An Empirical Validation of McGrath's (1984) Group Task Circumplex". In: *Small Group Research* 30 (1999), pp. 166–187. DOI: 0.1177/104649649903000202.
- Straus, Susan G. and Joseph E. McGrath. "Does the Medium Matter? The Interaction of Task Type and Technology on Group Performance and Member Reactions". In: *Journal of Applied Psychology* 79.1 (1994), pp. 87–97.
- Suh, Kil Soo. "Impact of communication medium on task performance and satisfaction: an examination of media-richness theory". In: *Information & Management* 35 (1999), pp. 295–312.
- Tajariol, Federico, Jean-Michel Adam, and Michel Dubois. "Seeing the Face and Observing the Actions: the Effects of Nonverbal Cues on Mediated Tutoring Dialogue". In: *Intelligent tutoring systems*. Vol. 5091. Lecture Notes in Computer Science. Springer, 2008, pp. 480–489.
- Tanakian, Mohammad Javad, Mehdi Rezaei, and Farahnaz Mohanna. "Digital Video Stabilization System by Adaptive Fuzzy Kalman

- Filtering". In: *Information Systems and Telecommunication* 1.4 (Oct. 2013), pp. 223–232.
- Thompson, Lori Foster and Michael D. Covert. "Stepping Up to the Challenge: A Critical Examination of Face-to-Face and Computer-Mediated Team Decision Making". In: *Group Dynamics: Theory, Research, and Practice* 6.1 (2002), pp. 52–64.
- Thompson, Lori Foster and Michael D. Covert. "Teamwork Online: The Effects of Computer Conferencing on Perceived Confusion, Satisfaction, and Postdiscussion Accuracy". In: *Group Dynamics: Theory, Research, and Practice* 7.2 (2003), pp. 135–151.
- Thomsen, G. and Y. Jani. "Internet telephony: Going like crazy". In: *IEEE Spectrum* 37.5 (2000), pp. 52–58.
- Vaalgamaa, Markus and Benjamin Belmudez. "Audiovisual Communication". In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 195–212.
- Vähätalo, Antti and Ingemar Johansson. "Voice activity detection for GSM adaptive multi-rate codec". In: *Proceedings of IEEE Workshop on Speech Coding*. 1999, pp. 55–57.
- Vallin, Jean-Marc et al. "High-Quality, Low-Delay Music Coding in the OPUS Codec". In: *Proceedings of the 135th AES Convention*. New York, USA, Oct. 2013.
- Varela, Martín, Lea Skorin-Kapov, and Touradj Ebrahimi. "Quality of Service Versus Quality of Experience". In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 85–96.
- Vary, Peter and Rainer Martin. *Digital Speech Transmission – Enhancement, coding and Error Concealment*. Chichester, West Sussex, UK: Wiley, 2006.
- Vatakis, Argiro and Charles Spence. "Audiovisual synchrony perception for speech and music assessed using a temporal order judgment task". In: *Neuroscience Letters* 393 (2006), pp. 40–44.
- Veinott, Elizabeth S. et al. "Video Helps Remote Work: Speakers Who Need to Negotiate Common Ground Benefit from Seeing Each Other". In: *Proceedings of CHI 1999*. 1999.
- Vos, Koen, Søren Skak Jensen, and Karsten Vandborg Sørensen. *SILK speech codec*. International Standard. Internet Engineering Task Force (IETF), 2010. URL: <https://tools.ietf.org/id/draft-vos-silk-02.txt>.
- Vos, Koen et al. "Voice Coding with OPUS". In: *Proceedings of the 135th AES Convention*. New York, USA, Oct. 2013.
- Wältermann, Marcel. *Dimension-based Quality Modelling of Transmitted Speech*. Springer, 2013.
- Wältermann, Marcel, Alexander Raake, and Sebastian Möller. *Comparison between the discrete ACR scale and an extended continuous scale for the quality assessment of transmitted speech*. ITU-T Contribution COM 12 - C 39 - E. Geneva, Switzerland: International Telecommunication Union, Feb. 2009.

- Weigel, Anne. "Beurteilung der Qualität von Telefonkonferenzen mit asymmetrischen Verbindungen aus der Perspektive eines passiven Zuhörers". Bachelor Thesis. Assessment of IP-based Applications, Technische Universität Berlin, Germany, 2014.
- Weiss, Benjamin et al. "Temporal Development of Quality of Experience". In: *Quality of Experience - Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, 2014, pp. 133–147.
- Werkhoven, Peter J., Jan Maarten Schraagen, and Patrick A.J. Punte. "Seeing is believing: communication performance under isotropic teleconferencing conditions". In: *Displays* 22 (2001), pp. 137–149.
- Whittaker, Steve and Brid O'Conaill. "An Evaluation of Video Mediated Communication". In: *INTERACT '93 and CHI '93 Conference Companion on Human Factors in Computing Systems*. ACM. 1993, pp. 73–74.
- Wiegand, Thomas et al. "Overview of the H.264/AVC Video Coding Standard". In: *IEEE Transactions On Circuits And Systems For Video Technology* 7 (July 2003), pp. 560–576.
- Wu, Wanmin et al. "Quality of Experience in Distributed Interactive Multimedia Environments: Toward a Theoretical Framework". In: *Proceedings of the 17th ACM international conference on Multimedia*. 2009.
- Wu, Yu et al. "vSkyConf: Cloud-assisted multi-party mobile video conferencing". In: *Proceedings of the second ACM SIGCOMM workshop on Mobile cloud computing*. ACM. 2013, pp. 33–38.
- Zielinski, Slawomir, Francis Rumsey, and Soren Bech. "On Some Biases Encountered in Modern Audio Quality Listening Tests – A Review". In: *Journal of the Audio Engineering Society* 56.6 (2008), pp. 427–451.
- Zölzer, Udo. *Digital Audio Signal Processing*. 2nd ed. Chichester, West Sussex, UK: Wiley, 2006.
- Zoubier, A.M. and B. Boashash. "The Bootstrap and its Application in Signal Processing". In: *IEEE Signal Processing Magazine* (Jan. 1998), pp. 56–76.