The Quality of Mediated-Conversations under Transmission Delay

vorgelegt von
Dipl.-Psych.
Katrin Schoenenberg
geb. in Groß-Gerau

von der Fakultät IV - Elektrotechnik und Informatik
der Technischen Universität Berlin
Assessment of IP-Based Application
zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften
Dr.-Ing.

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender: Prof. Dr.-Ing. Sebastian Möller
Gutachter: Prof. Dr.-Ing. Alexander Raake
Gutachter: Prof. PhD Søren Bech
Gutachter: Prof. PhD Peter Svensson

Tag der wissenschaftlichen Aussprache: 13.12.2015

Berlin 2016

Abstract

The influence of technical impairments on the perception of people using communication media has so far usually been considered from a single perspective. Most studies examine either the engineering side of communication systems or look at psychological aspects of users. The following work is incorporating these two views to gain a more complete picture of what happens if mediated communication is impaired. In particular, the focus lies on the influence of one specific degradation, pure transmission delay as experienced by the user.

For this purpose, a framework is postulated representing a more complete view on the measurement of the quality in mediated conversations. The five components discussed that constitute the **Quality of Mediated Conversations** are: the Conversational Quality, the Mediated Interaction, the Experiencer, the Interaction Partners and the Circumstances.

By means of laboratory experiments, the influence of the impairment pure transmission delay is investigated with relation to the proposed components. The focus here is on the components Conversational Quality, Mediated Interaction and the perception of the Interaction Partners. Seven experiments are conducted including two- and three-party groups. In six of these experiments, interlocutors are communicating over audio-only connections of three different speech bandwidths. The seventh experiment constitutes a first attempt to include audio-video Quality of Mediated Conversations into the framework as well.

The results show that the Conversational Quality which can be considered as a measure of the perceived technical quality is related to the transmission delay present in the connection. This relationship is strongly dependent on the type of conversation and the intended interaction speed of the interlocutors.

For this reason, a new parameter representing the delay sensitivity of a mediated conversation is introduced to enhance the prediction of the E-Model, an instrumental model predicting the Conversational Quality. The corresponding parameter is fitted based on the acquired data, and recommendations on values that could be used in a network-planning context are derived. Furthermore, a relationship of the delay sensitivity parameter and a measurement operationalising the component Mediated Interaction is shown.

Beyond the considerations on Conversational Quality, the component Mediated Interaction is investigated further. Nonverbal interaction is examined and the newly developed metrics are found to be strongly effected by delay. The investigation of verbal interaction supports this outcome and confirms the assumption that communication under transmission delay is more difficult. More precisely, interlocutors experience severe changes during the course of conversation also when no change in the perceived Conversational Quality can be measured. Due to delay, utterances are not placed at the same point in time at each end, which leads to different communication “realities” at the different sites. As a result, the turn-taking system becomes unbalance and communication problems become far more prevalent.

The third component which is considered in detail here, the Interaction Partners, plays another important role. Transmission delay is not directly perceivable as a

technical impairment and thus its reflection in the Conversational Quality rating is not straight forward. Participants do experience changes in the interaction which can be explained by the inherent tendency to attribute situational aspects, such as in this case technical impairments, to the disposition of people instead. In this context, users rate the current attentiveness or the personality of the Interaction Partners differently when delay is present in the line.

In summary, the sole Conversational Quality assessment is not enough to understand the consequences of transmission delay and it is recommended to follow a more comprehensive approach combining the engineering and psychological perspectives and corresponding analysis methods to gain insights into the Quality of Mediated Conversations.

To my parents.

Contents

Acronyms

IP	Internet Protocol
LTE	Long-Term Evolution
QoE	Quality of Experience
QoS	Quality of Service
QoMC	Quality of Mediated Conversations
CQ	Conversational Quality
MOS	Mean Opinion Score
RTT	Round-Trip Time
SCT	Short Conversation Test
iSCT	interactive Short Conversation Test
RNV	Random Number Verification
RNT	Random Number Verification Timed
3CT	three-party Conversation Test
3SCT	three-party Short Conversation Test
3RNT	three-party Random Number Verification Timed
CNG	Celebrity Name Guessing Task
ACR	Absolute Category Rating
NB	Narrowband
WB	Wideband
FB	Fullband

Part I
Theoretical Background and Research Questions

Chapter 1
General Perspective and Definitions

1.1 General Perspective on Mediated Communication

For the purpose of this work, a mediated communication situation is considered to encompass the persons participating in the conversation, the communication behaviour delivered via a medium and the context and frame in which the communication takes place. Figure 1.1 illustrates such a communication situation with its different factors. In the following, each factor is described in more detail and possible relationships among factors are highlighted.

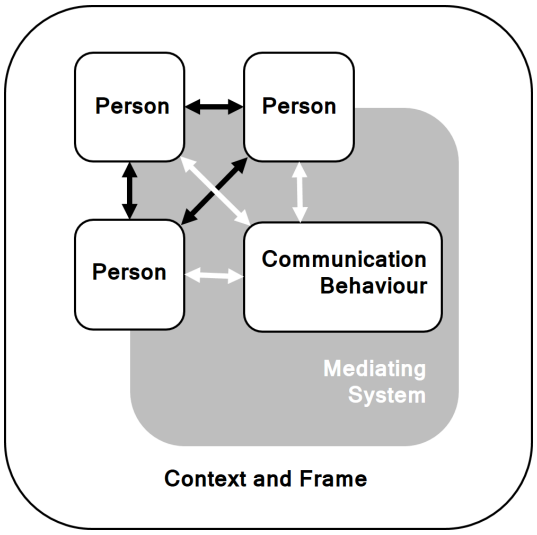


Fig. 1.1: Illustration of the general perspective on mediated communication, black arrows indicating the relationship between people, white arrows indicating the communication between people over the medium

The **medium** in such cases is given through a communication network and the appropriate devices used that may modify the communication behaviour delivered to the other end. Nowadays, most networks used for communication are based on the Internet Protocol (IP). In the scope of this work, the impact of pure overall transmission delay, which is a property of the communication medium, will be in focus. Transmission delays arise as a result of different processes in the transmission chain that take up time (Lakaniemi et al, 2001; Raake, 2006; Thomsen and Jani, 2000). At the sending side the recorded speech, audio or audio-visual data needs to be encoded and packed into container formats as used, for example, for Voice over Internet Protocol (VoIP) or Voice over Long-Term Evolution (VoLTE). The time needed for this process is typically predefined by the codec and the number of coded frames packed into one packet. The packets are then sent via the network which represents the most unpredictable part of the overall delay time. Different packets may be sent over different routes which is why some packets that were sent out earlier can arrive later. To ensure a coherent signal is played out to the receiving side, a so-called de-jitter-buffer is used. Here, the packets that have arrived are queued to re-collect them in the right order. After a certain predefined period of time (which can also be adaptive), packets that are still missing will be dropped and the data will be unpacked, decoded and played. Packet-loss in this case is typically referred to as packet drops. The processes of coding, network transmission and jitter-buffers and the according times of these processes determine the overall transmission delay. If no echo occurs in combination with delay this is typically called pure delay.

Apart from the impairment pure delay, two further aspects of the medium shall be investigated here. On the one hand side, the transmitted speech bandwidth may influence the communication situation. It is determined in the according codecs that are supported in the transmission chain. In the case of audio-video data transmission in contrast to audio-only data, the sending and coding process can influence the synchronisation of the audio and video content. None synchronised content is expected to affect the quality of the conversation over the medium as well.

Each **individual person** communicating over the medium can be characterised by his or her cognitions, emotions and behaviours (Petty and Wegener, 1998). Only the behaviours of a person are accessible to other persons, however, cognitions and emotions can be expressed by means of behaviours. In a particular situation, the expressed behaviour can be influenced by the persons current emotions associated with the situation (sometimes also called mood or status) and their own role in the situation. Also more long term aspects, such as prior experiences with similar situations or the persons' personality and temperament may determine the cognitions, emotions and behaviours of him or her. Cognitive and communication abilities of a person may further affect the communication situation, as well as, the currently available resources of attention for the conversation of each person.

On a **group** level, the relationship between communication partners is considered to be important for the interaction (Jones et al, 1999). Aspects that may determine the relationship could be for example mutual sympathy, trust and knowledge that

the interlocutors share (Gibbs, 1987; Keysar et al, 1998, 2000). These aspects can be related to prior experience with each other. The number of interlocutors and the different relationships among individual persons or subgroups of persons can alter the dynamics in conversations. Within a particular communication situation, the currently perceived status (e.g. current mood, attentiveness ect.) of the interaction partners often influence their own interaction behaviour.

In communication research, two ways in which **communication behaviour** can be expressed are typically distinguished: *verbal* versus *nonverbal*. Some researchers have criticised this distinction and argue that there is only communication (Jones and LeBaron, 2002; Kendon, 1972; Streeck and Knapp, 1992). It is true that all communication, including the communication of other species, represent "[...] the species' adaptation to the exigencies of a particular ecological niche in which communication facilitates survival " (Krauss, 2001). Different ways of expression may support this need of survival on different communication levels. For this reason, the above distinction is considered to be useful to gain a detailed view on the entire spectrum of communication (Nöth, 2000).

Verbal communication or language is often described as a system of *symbols*. Krauss (2001) defines symbols as being related to its message by social convention. A particular sound pattern, the symbol (i.e. the sound pattern for "flower'"), in a particular language (in the example English) can stand for an object while a different sound pattern (i.e. "Blume") can stand for the same object in another language (i. e. German). Symbols are usually used in verbal communication. Symbols need to be distinguished from *signs*. Signs are causally related to the message they deliver. For example, if a persons face is turning white this may indicate that the person is not feeling well. Signs are often part of nonverbal communication (Burgoon et al, 1996). However, the definitions for the terms symbol and sign are quite different depending on the author and in some cases they have even been switched (Nöth, 2000). For this work, the theoretical format of the expressed communication behaviour as a symbol or a sign, is not of importance.

Here, the different types of communication behaviours and how they are exchanged will be of relevance. Besides verbal and nonverbal communication, a second perspective for distinguishing communication behaviours has been proposed: *vocal* versus *nonvocal* (Laver and Hutcheson, 1972; Poyatos, 1976). When combining all possible communication types, four categories can be observed (adapted from Nöth, 2000; Poyatos, 1976), that will be described further below:

- *verbal and vocal*: spoken language
- *verbal and nonvocal*: written and sign language
- *nonverbal and vocal*: vocalics, also known as paralinguistics
- *nonverbal and nonvocal*: kinetics, haptics and physical appearance

According to this categorisation, ***verbal***, can be described as "communication by means of words" while ***nonverbal*** covers "all communication except for words". Definitions for nonverbal communication can be quite different in terms of what

kind of assumptions they make. The above definition can be considered as the broadest one. When defining nonverbal communication the kind of behaviour that should be included is often highly debated. A much more restricted definition was postulated by Burgoon et al (1996), saying: "We consider nonverbal communication to be those attributes or actions of humans, other than the use of words themselves, which have socially shared meaning, are intentionally sent or interpreted as intentional, are consciously sent or consciously received, and have the potential for feedback from the receiver". In this definition, not all behaviours evident need to be considered as nonverbal behaviours but only those which all parties can understand (socially shared). This can be particularly challenging in intercultural interaction (Knapp and Hall, 2010). Even though a particular nonverbal behaviour, i.e. head shaking, may be shown intentionally it does not necessarily have the same meaning for all participants of a conversations (i.e. in India soft head shaking symbolises attention for a conversation while it represents disagreement in most other countries). Furthermore, according to Burgoon et al's definition, only the behaviours that were either sent or received intentionally and consciously need to be included. Goffman (1963), on the other hand, suggests that "we can stop speaking but we cannot stop to communicate".

In the scope of this work, nonverbal behaviour will be assessed from an observation perspective. From this point of view, it is difficult to determine whether a particular behaviour was intentionally or consciously sent or received. For this reason, all nonverbal behaviours observed with a particular measurement technique will be considered and thus the broadest definition will be applied for practical reasons.

Coming back to verbal and nonverbal communication, Burgoon et al (1996) suggested six ways how the two communication types can work together: Both can either be used to express the same (redundancy); nonverbal behaviour can substitute; complement; emphasise or contradict the verbally expressed. Finally, nonverbal behaviour can also regulate the communication, in particular the turn-taking. This functionality is important when the mediating system comes into play.

The distinction between vocal and nonvocal is based on "using the voice" (***vocal***), or "not using the voice" (***nonvocal***), for communication purposes. The so-called *vocalics*, which fall into the category of nonverbal and vocal communication, go back to Trager (1958). He described three dimensions of vocalics: the voice set, voice qualities and the vocalization. The first dimension is related to the physiological properties and physical surrounding of the speaker. Voice qualities include aspects like the pitch range, the resonance of the voice, rhythm control or the tempo. The third dimension includes peculiarities of the voice, for example, the intensity or the pitch height or whether someone is laughing or whimpering. The so-called vocalized pauses (i.e. sounds like "uh", coughs or sniffs) also fall into the category of vocalizations. Jones et al (1999) later included aspects such as the turn length, the speech rate or the interruption frequency as vocalics. Furthermore, the latency of responses can also be considered to fall into this category (Knapp and Hall, 2010).

The fourth category listed above, nonverbal and nonvocal communication, subsumes all communication behaviours that are not using words and not using the

voice for expression. *Kinetics* fall into this category and they are represented by gestures or facial expressions. Also *haptics* such as e.g. clapping on the shoulder or hugging, and *physical appearance* can be considered as nonverbal and nonvocal.

When comparing the different nonverbal behaviours, Burgoon et al (1996) point out that vocalics, which are vocal, and kinetics, which are nonvocal, are most heavily used whereas haptics and physical appearance (both nonvocal) seem to be more difficult to code or decode and are therefore not used as extensively. They may still be very important in some situations.

The actual usage, particularly of nonverbal vocal and nonvocal, communication behaviours is strongly dependent on communication rules particular to the situational setting (e.g. business versus private, see components context and frame below) or the cultural background that a person has acquired (Gellri and Kanning, 2007).

In the context of mediated conversations, the **mediating system and the communication behaviours** are related such that the transmission though a system with certain technical properties may modify the delivered communication behaviour, and thus also impact the turn-taking (more details will be given in Chapter 3). The transmission delay of the mediating system plays an important role in this context because it alters timing-related aspects of the nonverbal vocal communication and therefore may change the interpretation of the messages' meaning at the receiver side.

From a psychological point of view, the *richness* of a communication-mediating system (Darft and Lengel, 1986; Rice, 1992) facilitating *social presence* (Short et al, 1976) has been stated as an import factor in this context. The richness of a transmission medium is defined through the type and amount of communication behaviours that are delivered (Sassenberg, 2004). The related theory on social presence was created based on results comparing different types of communication media. They showed that mediating systems including a visual channel better allow for an exchange of social behaviours (intra- and interpersonal) for tasks requiring those kind of behaviours, such as conflict or negotiation tasks, than systems without any visual input, hence, people feel more comfortable with the visual channel available (Apperley and Masoodian, 1995; Isaacs and Tang, 1994; McGrath, 1984; Olson et al, 1995; Short et al, 1976; Whittaker, 1995). Of course, in a real life context, a high degree of social presence is not always desirable. It can enhance the effort for monitoring the persons own impression and thus being "observed" via a video connection can lower the performance (Brander and Mark, 2001; Goffman, 1961). A possible reason for this could be that it distracts the individual from the actual content that needs to be discussed.

In general, however, the literature seems to find rather no difference when comparing audio-only to audio-video systems in terms of task effectiveness (Apperley and Masoodian, 1995; Brander and Mark, 2001; Dennis and Kinney, 1998; Gale, 1991). The task effectiveness is in this context often termed task performance.

It could be possible that in case of an audio-only connection people use speech differently to deliver a certain meaning due to the lack of visual input. For exam-

ple, more vocalics may be deployed because other nonverbal nonvocal behaviours cannot be used. Pye and Williams' assumption that properties of the mediating system influence the way in which the communication system is applied is not consistently supported by the literature (Doherty-Sneddon et al, 1997; Jones et al, 1999; O'Conaill et al, 1993; Sellen, 1995).

Nevertheless, both types of mediating systems were usually found to be different to face-to-face communication and therefore it can be assumed that people use a different communication style for mediated communication. Face-to-face communication should thus not be considered as the baseline condition or considered similar to a perfect mediated communication (Dourish et al, 1996; Fish et al, 1993; Friebel et al, 2003; Sellen, 1995).

The **context** of a communication situation covers properties of the surrounding environment and the **frame** defines the initial reason for the realisation of the conversation. The frame can be characterised by aspects like the reason for or underlying motivation to communicate (business versus private, important versus casual) and associated rules or norms (Goffman, 1974). The context, on the other hand, is constituted by the perceptual conditions that a person experiences, such as lightening, background noise, ect.. The possible distraction that the surrounding environment may cause can result in a reduction of the available resources of attention for the communication itself. Both of these components, which can be subsumed to the circumstances of the communication, can lead to different assumptions in the mind of the interlocutor and can alter the way communication behaviours are used. In an important business conversation, for instance, taking place in an office people are unlikely to laugh aloud to show their consent but rather only smile and nod.

Based on the relationship of **communication behaviours to the context and the frame**, it is possible to determine whether the expressed communication behaviour was comprehended. Besides individual factors, such as willingness, readiness to listen and prior knowledge to be able to integrate the information, three factors are of importance for comprehension (adapted from Raake, 2006):

- comprehensibility: "addresses how well the speech [...] allows content to be related to it"
- intelligibility: "refers to how well the content of an utterance [...] can be identified on the basis of the form"
- communicability: "means that a speech message is such that it can serve to communicate, that is, can fully be understood by recipients, ideally as it was intended by the sender"

With this general perspective on a communication situation in mind the goal of this work can be addressed, that is, to describe and quantify the *Quality of Mediated Conversations*. For this reason, the following sections will first look at definitions of quality in general, then the *Quality of Experience* will be considered, followed by a detailed description of the concept of the Quality of Mediated Conversations.

In the following chapters, the proposed components of the Quality of Mediated Conversations will then be examined in more detail and findings from the literature will be reported.

1.2 Classical Definitions of Quality

For a general understanding of the term **quality**, definitions made by standardization organizations and by Jekosch will be looked at first.

The International Organization for Standardization (International Organization for Standardization, ISO, 2005) uses the following description.

ISO 9000
Quality "[...] is the ability of a set of inherent characteristics of a product, system or process to fulfil requirements of customers and other interested parties".

In this definition, the view of a possible stakeholder, such as a communication service provider, is taken. In contrast the German Institute for Standardization (Deutsches Institut für Nomierung: DIN) defines quality in a more general way (2008-05; 1989-03).

DIN 55350 Part 11 and 12
Quality is the "[...] composition of an entity with regards to its ability to fulfil determined and necessary requirements"

Where a ***composition*** can be understood as "[...] the totality of the features and the values of features [...] of an entity "[...] ."

Within this, the terms are defined as follows:
feature "[...] characteristic for recognizing or distinguishing between entities [...] "
value of feature "[...] the value attributed to the manifestation of a feature."
entity "[...] material or immaterial object under observation"

On this basis, Jekosch (2005) chose the following definitions in the scope of voice and speech quality assessment:

Quality is the "[r]esult of judgment of the perceived composition of an entity with respect to its desired composition."

A ***perceived composition*** is understood as the "[t]otality of features of an entity.", a ***desired composition*** as the"[t]otality of features of individual expectations and/or relevant demands and/or social requirements."

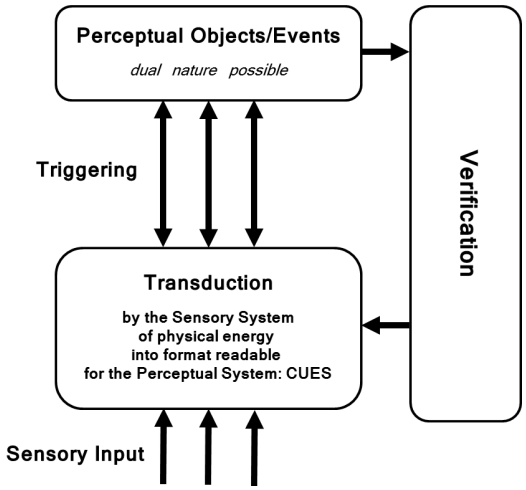


Fig. 1.2: The formation of perceptual object/s or event/s, adapted from (Mausfeld, 2010b)

and the term ***feature*** as the "[r]ecognisable and nameable characteristic of an entity".

Jekosch's definition includes the involvement of perceiving, experiencing and reflection of or on an event to be judged in terms of quality.

1.3 Formation of Quality Judgment

As the above definitions of quality imply, perception of an event under assessment by the perceiving person is an essential component of the quality formation. Based on sensory input, auditory, visual or otherwise, the sensory system provides *cues* to the *Perceptual System* (see Fig. 1.2). These cues are *transductions* of the physical energy into a data format that is easy to read for the Perceptual System (Mausfeld, 2010b). Through *triggering*, the concepts of the Perceptual System are elicited. These concepts can be regarded more precisely as functional structures and are usually called *Percepts*. It should be pointed out that the format of perceptual cues correspond to the format of the Percepts. Therefore, a linear picture involving processing step after processing step is regarded as incorrect (Mausfeld, 2002). Instead, the format of the Percepts determine in a top-down manner the cues provided by the *Sensory System* (bottom-up). As Arnauld and Nicole (1662) already clearly stated: "None of our ideas come from sense[s], sense[s] supplying only the occasion for the mind to form ideas of its own." This quote underlines the importance of the Percep-

tual System for our *perceptual reality* as it acknowledges that sensory input causes the triggering of Percepts.

A perceptual object or event, subsumed under the terminology Percept, is not necessarily stable. Through verification mechanisms (see Fig. 1.2), such as, moving the head to get more binaural input for locating an auditory object, the Percept can be stabilized (Hoffman, 2000). Beyond this, more than one Percept can be elicited. Mausfeld calls this situation the *dual nature* of perception (Mausfeld, 2003, 2010a, 2011). In the case of perceiving a picture or movie for example, two different Percepts are triggered. On the one hand, the picture is recognised as an artefact with certain properties such as the surface texture or the material, being flat (2-D). On the other hand, the scene in the picture is perceived with certain depth (3D) and objects, such as, humans showing emotional expressions. It could be said that the picture depicts its own reality. The world of the scene is as present as the artefact picture itself, following, both can be evaluated separately. In the case of a communication mediating system, a similar dual nature can be observed. The communication system is recognised as a technical artefact, and simultaneously people can use and evaluate the *virtual reality* (Mausfeld, 2013) created by it. For both, the picture and the communication system, the artefact is not completely independent from the virtual world. The colours used for painting determine the atmosphere created, similarly technical impairments of a communication system influence the communication process.

Upon both, the virtual reality itself and the technical details a human can reflect and attribute causes for possible abnormalities (Fig. 1.3). As a part of this, the human is able to recall quality features from memory (desired features), which he or she expects, and compare them to the event perceived. The result will be a judgment on the quality of the current event. It should be noted that the recall from memory is determined by the context or the current mood of the person (Schoenenberg and Raake, 2011), and cannot be regarded as an objective ground truth. Typically the judgment result is expressed on one or multiple scales.

1.4 Quality of Experience

In this work, quality related to the usage of telecommunication services will be examined - the **Quality of Experience (QoE)**. This terminology aims to address all facets of the "positive and negative experience" related to the usage of a technological system. In the majority of the literature, however, QoE is still expressed in terms of a sole technological quality, typically addressing a telecommunication or IP-based transmission system.

Therefore, the latest definition of QoE is specified for applications, services or systems and involves aspects related to the those using it. According to Raake and Egger (2014) and Möller et al (2013) and in line with Kahneman (2003) it is defined as:

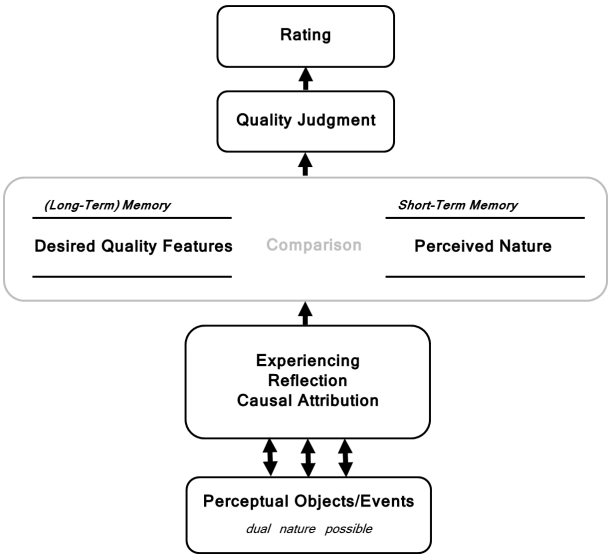


Fig. 1.3: The formation of a quality judgment, adapted from Jekosch (2005)

Quality of Experience is the degree of delight or annoyance of a person whose experiencing involves an application, service, or system. It results from the person's evaluation of the fulfilment of his or her expectations and needs with respect to the utility and / or enjoyment in the light of the person's context, personality and current state.

In the above definition, three points are described in more detail than in the definition by Jekosch (2005). Firstly, it states that the outcome of the comparison process is understood as the degree of delight or annoyance, implying a positive - negative nature. Secondly, it states that QoE always includes the use of or encounter with an application, service or system. As a third point, the desired composition is more precisely referring to utility and enjoyment (Kahneman, 2003).

The experience of the sole technological side of QoE is now encompassed by under the **Quality based on experiencing**. The formation of the quality judgment and conversion to fit the provided quality assessment scale is defined by Raake and Egger (2014) and in line with the quality definition by Jekosch (2005):

Quality based on experiencing results from the "judgment of the perceived composition of an entity with respect to its desired composition".

QoE furthermore has to be distinguished from "***[a]ssumed quality*** [which] corresponds to the quality and quality features that users, developers, manufacturers or service providers assume regarding a system, service or product that they intend to be using, or will be producing, without however grounding these assumptions on

an explicit assessment of quality based on experiencing." (Raake and Egger, 2014). For this kind of quality no actual encounter with an application, service, or system is needed, the pure anticipation of its use shapes it.

Moreover, QoE differs from **Quality of Service (QoS)**. While QoE represents the users point of view, QoS encompasses the viewpoint of the provider. In ITU-T Rec. E.800 (2008) ***Quality of Service*** is defined 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." (for more details see e.g. Möller, 2000; Raake, 2006; Raake and Egger, 2014).

1.5 Quality of Mediated Conversation

In principle, the concept of QoE could be applied to the situation of mediated communication and the latest definition could serve as its basis. However, research mistakenly associated to QoE is in most cases only concerned with what is now called Quality based on experiencing and thus only measuring the perception of the technical or transmission side of the communication situation.

This work aims to address the perception of the transmission and explicitly other component that contribute to the experience. For this reason, a new term - **Quality of Mediated Conversations (QoMC)** - has been chosen to specify when mediated communication is addressed. It will be assumed that a perfectly well mediated conversation serves as the reference to which an experienced mediated conversation is compared in order to gather a quality judgement in the sense as described above. If users are unfamiliar with using a system it can be expected that this reference is an abstract one. A face-to-face conversation is not considered to be appropriate and therefore not a used reference. It has been shown in the past that people develop a specific style of communication to fit the available modalities (Doherty-Sneddon et al, 1997; Dourish et al, 1996; van der Kleij et al, 2009; Fish et al, 1993; Pye and Williams, 1977). Despite that communication itself is based on face-to-face interaction, it is believed that people are able to construct a mediated communication reality that follows different behavioural strategies.

It is proposed that the ***Quality of Mediated Conversations*** is a function f determined by the following components which are expected to be, to a large extent dependent on delay and interdependent to each other. [1]:

$$QoMC = f(CQ, MI, P, E, C) \tag{1.1}$$

with

$$MI = f^*(ver, nver, voc, nvoc) \tag{1.2}$$

[1] The author is aware of the fact that most variable names are problematic from a mathematical point of view because it may appear to the reader as if several variables are multiplied. Nevertheless, these variable names were chosen to better associate them to the according concepts.

- The component ***Conversational Quality (CQ)*** is a measure for the perception of the connection or mediating system, as termed by ITU-T Rec P.805 (International Telecommunication Union, ITU-T, 2007b). It can also be referred to as a Quality based on experiencing.
- The component ***Mediated Interaction (MI)*** of a conversation is composed of the sub-components: *verbal* (*ver*), *nonverbal* (*nver*), *vocal* (*voc*) and *nonvocal* (*nvoc*). These are measures of the type of the communication behaviour and are combined in the function *f**.
- The component interaction ***Partners P*** subsumes the respective attributes assigned to interlocutors by the participant based on prior knowledge, experience in interaction with this person and on the situation itself.
- The component ***Experiencer* (E)** constitutes the individual person who is experiencing the communication situation. He or she is also assessing the QoMC and has certain personality and situation-related attributes that can influence the judgment and other components.
- The component ***Circumstances C*** subsumes all aspects related to the perceptual impact of the context and the perception of the communication frame representing the initial motivation for a particular conversation.

Based on the above equation 1.1, the CQ can serve as a predictor for the QoMC, as well as the component MI or any of the other components. The best prediction can be expected if all components are taken into account. For the scope of this work, the focus will be on investigating the components CQ, MI and P in case of transmission delay.

1.6 Focus of This Work

The main goal of this work will be to examine the impact of the independent factor the *level of pure (one-way) delay* which is a property of the transmission medium, through the three components Conversational Quality; Mediated Interaction and the perception of the Interaction Partners. [2]

Most other properties of the transmission chain will be kept constant throughout all empirical evaluations. Audio-only communication applications will be in focus, however, an initial analysis of audio-video communication will be undertaken. Audio-only and audio-video applications can be considered as comparatively synchronous (in contrast to, e.g., email), and rather rich in terms of manifoldness of channels (speech, visual) and context information accessible to the interactants (in contrast to, e.g., a fax; Rice, 1992).

The component Mediated Interaction will mostly rely on measures addressing nonverbal and vocal aspects because the delay impairment can have a direct influ-

[2] If no explicitly indicated as Round Trip delay Time (RTT), all delay values reported in the following are one-way delay times in milliseconds

ence on them, as explained above. Nonverbal and nonvocal communication will be investigated through an audio-video experiment. Verbal communication behaviours will not be in the main focus. Nevertheless, for a precise estimation of the QoMC an analysis of verbal information and how the content exchange is structured (as e.g. in Schrire, 2004) or in how far people enjoy the content (as e.g. in Tam et al, 2012), should be taken into account. This work will look at two-party and three-party interaction. Undoubtedly, groups with higher number of interlocutors should be examined in future work to enable greater generalisability of the results.

The Circumstances will be kept constant by using a laboratory setting and the frame of the communication will be set by the experimental tasks, even though this can only be considered as an "artificial" variation of the frame. Field evaluations should complement the research in the future. The influence of the Experiencer will be considered to be randomised in the best way possible. The author is, however, aware that the drawing of participants from the total population is not entirely independent due to the recruitment process of participants.

More details on the theoretical background for the individual components will be given in the following Chapters 2, 3 and 4. The particular operationalisation of the components will be described in the Chapter 7.

Chapter 2
Conversational Quality

In this chapter, approaches described in the literature for measuring the influence of pure delay in mediated communication on the Conversational Quality will be explored. The goal will be to identify the most beneficial approaches for extending them later in the Empirical Part II.

2.1 Assessment of Conversational Quality

In general two approaches need to be distinguished when talking about the assessment of quality. The utilitarian approach looks at the overall impression of quality. The respective measurement is typically referred to as *integral quality* or *overall quality* (Möller, 2000; Raake, 2006) and will be in the focus of this work. In contrast, analytic approaches assess the perceptual features that contribute to the overall quality (Möller, 2000; Raake, 2006; Wältermann, 2013).

Subjective quality assessment is necessary to identify relationships of certain technical degradations and the perceived overall quality or perceived quality features. Such investigations fall under the category of perception-based methods. Instrumental models, on the other hand, try to predict the perceived quality. They are used in the context of network planning and they base their prediction on certain assumptions, such as, additive impairment factors or on mathematical models of the perception.

Two types of subjective quality can be distinguished the **Conversational Quality (CQ)** and the **Listening-only Quality**. Some authors mention a third type the **Talking Quality** as the equivalent for Listening-only Quality (Möller et al, 2011). The type of quality is related to the type of assessment method: *conversation* versus *listening-only tests*. A typical **listening test** will use two to five short speech samples of about 2 to 3 seconds. Speech material from at least two different male and two different female voices will be recorded. The speech samples are typically processed with the system under test. Stimuli are then either presented in isolation

using single-stimulus methods such as the *Absolute Category Rating*, or by comparing pairs of multiple stimuli such as in the *Degradation or Comparison Category Rating* method. When using pairwise comparison, one stimuli usually constitutes an unprocessed reference stimuli. Then either participants can be asked how much better one of the stimuli was than the other *Comparison Category Rating*) or the reference stimuli needs to be adjusted to an impaired stimuli (*Isopreference Test*). Accordingly, detailed descriptions of listening test methods can be found in ITU-T Rec. P.800 (1996). Listening tests have the advantage that they can be set up easily and the conduction does not take up much time because several participants can be tested simultaneously using the same material. However, they lack a natural conversation situation and degradations that are related to the conversation cannot be examined.

Each trial in a **conversation test** lasts about 2-3 minutes for a two-party and 4-8 minutes for a three-party test (depending on the task and degradation) as described in ITU-T Rec. P.805 (2007b; audio-only), ITU-T Rec. P.920 (2000; audio-video) and ITU-T Rec P.1301 (2012c; the multi-party case). Different aspects of quality play a role in conversation tests, especially in terms of listening and/or viewing one or several other conversation partners, talking (possibly leading to echo, or some specific sidetone perception), and conversation (in terms of the interaction with the other). Conversation tests in contrast to listening-only tests encompass the assessment of the influence of degradations that impact the conversation, such as delay, as well as listening degradations such as packet loss or echo. Beyond this, conversation tests represent a more realistic usage situation, and results can thus be considered as more externally valid than those of listening-only tests. This is underlined by the assumption that people adapt their communication behaviour to certain communication conditions. As a result of this adaption, the degradation may be more or less severe in terms of quality. For listening tests there are usually only a few speakers recorded to compile the files for the quality testing, whereas in conversation tests each participant has a different voice and communication behaviour. Here, the outcome is based on a higher variability of speakers and hence is more generalisable.

Conversely, conversation tests usually imply a certain predefined task, whereby more attention is needed to accomplish the task (in contrast to the case when participants only listen) and less attention may be available for the reflection on the quality judgment. For this reason, tasks should carefully be selected. Especially in the case of video-based communication, tasks should also comprise components involving visual communication. This can be hindered if the task requires a lot of paper material that takes the focus away from the screen. Furthermore, the coordination of participant groups for conversation tests requires much more effort for the examiners than finding single persons for a listening-only test. However, conversation tests have the advantage that for each successfully finished session, ratings of at least two participants are collected (depending on the number of interlocutors per group).

Obviously, conversation tests are the only means for validly assessing the influence of delay on quality due to the high impact of transmission delay on the conversation structure. For this reason, the Conversational Quality will be in focus of this work, the sole listening quality will not be assessed. Conversational Quality can be

ausgezeichnet	gut	ordentlich	dürftig	schlecht
5	4	3	2	1

Fig. 2.1: The Absolute Category Rating Scale (ACR) according to ITU-T Rec P.800 (1996) with German labels: 5 = ausgezeichnet (Engl.: excellent), 4 = gut (Engl.: good), 3 = ordentlich (Eng.: fair), 2 = dürftig (Engl.: poor), 1 = schlecht (Eng.: bad)

assumed to comprise aspects of talking and listening quality (further discussion can be found e.g. Möller et al, 2011).

A variety of **conversation tasks** have been proposed for evaluating the CQ of audio-only and videotelephony systems (International Telecommunication Union, ITU-T, 2007b, 2000). The tasks define more precisely what participants need to do in a conversation test. A subset of those tasks are typically referred to as *conversation scenarios* because they try to generate close-to-natural conversations. As will be explained in more detail later, the task type and the related interaction speed is expected to have a major influence on the extent to which the delay impairment is reflected in the quality judgment. The task types used in the empirical part of this work will be discussed in more detail in Chapter 6.

ITU-T Rec P.800 and P.805 specify the five point **Absolute Category Rating (ACR)** scale to collect quality ratings from test participants during conversation or listening-only tests (Fig. 2.1). Based on the ratings from all participants of a test, the Mean Opinion Score (MOS) is calculated. Although the scale level is ordinal rather than interval, statistics such as the mean and standard deviation are typically calculated. Therefore the scale is treated as if it was an interval scale. Strictly speaking, this is not correct, but the statistics employed for analysing the ratings of most scales used in various research areas face the same problem.

The ACR scale has been criticised in the past for the issue of unequal intervals between categories (Aldridge et al, 1995; Watson, 2001). It should be noted that intervals between German scale labels were found to be almost equivalent. Similar problems can occur with continuous scales. For example, people may rather apply a logarithmic or even completely different spacing on continuous scales. Therefore, the problem of non-equally distributed categories is comparable with effects due to other scales. Numbers added to the verbal descriptions of the categories may help to equalise the distribution, as implied by the similar usage of ACR-type and continuous scales found in a study by Huynh-Thu et al (2011).

Continuous versions of the ACR-scale have been proposed, adding two non-labelled extreme points to the ends and smaller ticks between the categories (Bodden and Jekosch, 1996). Outer categories are sometimes avoided by raters since a subsequent event which is perceived as being of even higher quality may not be accounted for correctly if the prior event was already rated using the highest available categories. This newer version of the ACR scale tries to solve these problems.

Including these considerations, the biggest advantage of the standard ACR scale is that most research in the context of delay was performed using the standard version. Furthermore, currently the most widely known prediction model for quality in the context of network planning, the E-Model, maps to ACR *Mean Opinion Score (MOS)* values. Since it is the aim of this thesis to compare outcomes to prior research and to be able to update the current E-Model based on the collected data, the original version of the ACR-scale will be used here.

When **planning a network**, the CQ that can be achieved is one of the key criteria for planning choices. The ITU-T Rec. G.114 (2003a) recommends to not exceed a one-way delay of 400 ms when planning any type of speech communication application (see Fig. 2.2). If delays are kept below 150 ms "most applications, both speech and non-speech, will experience essentially transparent interactivity", as stated in ITU-T Rec. G.114. For the case of audio services, these values address the case of pure delay in the complete absence of any echo. Effects of talker echo are described in ITU-T G.131 (2003b). When delay is presented in combination with audible echo, it becomes directly perceivable for the interlocutors. As a consequence, the CQ ratings are considerably more critical for echo than for pure delay (Guguin et al, 2008; Helder, 1966; Riesz and Klemmer, 1963). The additional effect of delay on perceived quality is rather minor if it occurs together with more perceptionally obvious degradations such as packet loss or packet drops due to jitter (Berndtsson et al, 2012; Bräuer et al, 2008; Cermak, 2002; da Silva et al, 2008; Raake, 2006). For videotelephony and the combination with packet loss, bandwidth and delay, the factor packet loss was found to be most decisive for the resulting MOS (Cermak, 2005). These outcomes may, however, be determined by the chosen strength of the individual degradations.

When considering pure delay, early contributions to ITU-T in the 1990ies reported a higher impact on ratings of interruptibility than for delays ranging from 0 ms to 1250 ms one-way (ITU-T Delayed Contribution 21, 1990; and Delayed Contribution 22, 1990). Based on the results, the authors recommended to keep the upper limit of 400 ms from ITU-T Rec. G.114 as the maximum delay for network planning. It should be noted that these findings already suggested that the sole assessment of CQ only partly reveals the effect of pure delay on conversations.

Another contribution to ITU-T addressed the limits set in ITU-T Rec. G.114 and the impact of pure delay from another perspective (ITU-T Delayed Contribution 214, 2004). The conversational structure, measured in terms of double talk times and length of pauses was found to be affected for delays lower than 400 ms already, though, this impact was not dramatic in the range between 150 ms to approximately 300 ms. Nevertheless, the authors questioned whether the experience of the user is indeed not impaired for delays reaching 200-250 ms. They summarised the following: "Those familiar with the laboratory data from which earlier versions of G.114 were formulated know that this depiction of delay's effect on voice quality is a simplified one. Very different results were reported by different laboratories, and the resulting recommendation was a compromise." (ITU-T Delayed Contribution 214, 2004). As this statement suggests other measures besides the sole CQ assessment

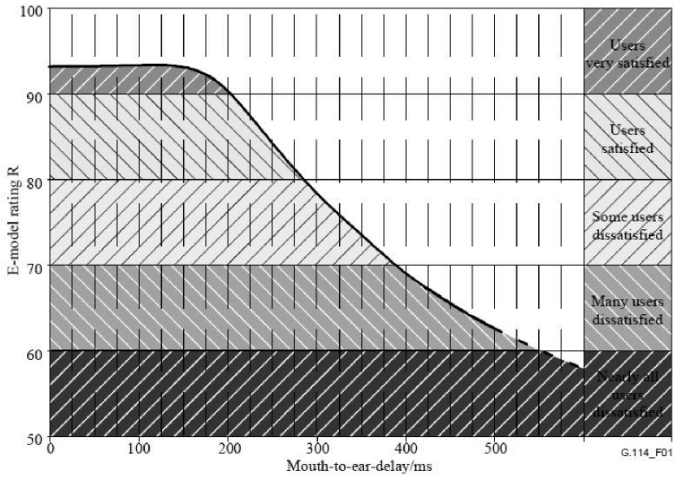


Fig. 2.2: Conversational Quality as a function of one-way transmission delay as predicted by the E-Model as available prior to the present work, adapted from ITU-T Rec. G.144 (2003a)

could help to better understand the entire impact of delay on the conversations. Such alternative measures will be developed further in the following chapters.

Regarding audio-video synchronisation, the ITU only gives recommendations for broadcast video services to date. It is recommended that sound should not precede the video by more than 90 ms and that video should not precede the sound by more than 185 ms (ITU-R Rec. BT.1359-1, 1998). These values are based on the known acceptability thresholds for audio-video asynchrony in broadcast video (ITU-R Rec. BT.1359-1, 1998). A recommendation by the ITU on the impact of audio-video synchronisation on the CQ in interactive services is currently lacking. Since interactive services require a higher degree of user involvement, less attentional capacity may be available and higher asynchrony values may be required to have users realise the effect of asynchrony and reflect it in terms of CQ.

In the following subsections, findings regarding the impact of transmission delay on rejection of calls and on CQ will be reported. An in-depth discussion of how the conversational structure changes when delay is present will be given in Chapter 3. The following subsections present literature findings from the (audio-) telephony, video-telephony and videoconferencing domains. Most studies reported for the audio domain are based on two-party interaction, multi-party studies (i.e. conferencing studies) are usually covering the audio-video case. Besides basic considerations, the aspect of audio-video synchronisation will be discussed.

2.2 Audio Transmission Delay

Back in the sixties, when scientists assessed the impact of delay in satellite transmission, the effect due to delay was not primarily assessed in terms of CQ. Back then, the **percentage of rejected calls** was assessed, which can be considered as a measure very similar to CQ. [1] For naturally occurring telephone conversations, users of the delayed telephone lines were instructed to dial a particular digit without hanging up, to show that they rejected the line as unsatisfactory for a normal use (Klemmer, 1967; Riesz and Klemmer, 1963).

In a first study by Riesz and Klemmer (1963), administrative staff of the Bell Laboratories were exposed to telephone circuits with 600 and 1200 ms Round-Trip Time (RTT) echo-free delay in their frequent internal calls (inserted delays altered each working day). [2] They were told that some calls would be routed over a simulated satellite circuit, but it was not indicated which calls. During the first four weeks, there was hardly any rejection recorded. For this reason, the 1200 ms RTT condition was altered with a condition with 2400 ms RTT delay in week five and six. In the following six weeks, the rejection rate for the 600 ms RTT condition increased to 25 % and for 1200 ms RTT it went up to 43 %.

Klemmer (1967) repeated this study with different staff again over a period of 12 weeks without any exposure to delays higher than 1200 ms. Even though a much larger number of delayed calls were made, the satisfaction remained rather high and hardly any calls were rejected for the entire period.

After this first phase, people were exposed to telephone lines with 1800 ms RTT delay every day over a period of two weeks, followed by two weeks with conditions of 2400 ms RTT delay every day. In spite of the conditioning, the rejection rate of the 600 ms RTT condition was not increased in the following 10 weeks (during trials conducted in a similar manner as for the earlier 12-week test) but for the 1200 ms RTT condition a rejection rate of 12 % was found.

Since these outcomes were not showing a sensitisation effect as prominent as in the prior study the authors concluded that the magnitude of a sensitisation effect is difficult to predict (Riesz and Klemmer, 1963). Nevertheless, these studies suggest that sensitisation can occur in the context of measuring the impact of pure delay. It can be questioned if such effects are still relevant today because people are generally more used to delays from using IP-telephony services or mobile phone connections.

Two known studies on the **CQ of telephone calls** have been published in the early 1990ies reporting inconsistent results. In Karis' study (1991), participant pairs were asked to find matching halves of postcards while delays up to 1200 ms RTT

[1] Nowadays, researchers call this kind of measure the *acceptability* of a connection. The acceptability is always given as a "yes - no" or "acceptable - not acceptable" judgment and it is not exclusively assessed for real life calls but also in laboratory experiments.

[2] The RTT describes a full circle of transmission starting from the sending point to the receiver side and going back to the sending point. In contrast, the one-way delay includes only the path from sender to receiver side and can be considered as one half of the RTT if delays are stable and symmetric.

were inserted into the used mobile telephone lines. The rated quality of the connection, as well as the rated listing effort and the task performance, measured in terms of incorrect and incomplete pairs, did not differ significantly between delay conditions. However, the number of interruptions, counted by two raters listening to the taped conversations, increased significantly. Karis (1991) further observed that fast-pace conversations are more likely to show communication problems when delay is present.

On the other hand, Kitawaki and Itoh (1991) found a clear dependency of perceived quality and transmission delay (up to 4 sec RTT). The drop in CQ was essentially determined by the interaction speed required to accomplish the task. They tested the following six tasks, listed in the order of decreasing expected interaction speed:

1. Take turns reading random numbers aloud as quickly as possible.
2. Take turns verifying random numbers as quickly as possible.
3. Words with missing letters are completed with letters supplied by the other talker.
4. Take turns verifying city names as quickly as possible.
5. Determine the shape of a figure described verbally.
6. Free conversation.

The outcomes of the study showed, the higher the initially expected speed of a task, task 1 being the fastest and task 6 being the slowest, the more critically CQ was rated by test participants.

Three reasons may explain the mismatching results of the work carried out by Karis (1991) and Kitawaki and Itoh (1991). The general sensitivity of participants for both must have been the same since both studies were conducted at approximately the same time, in the early 90ies. In Kitawaki and Itoh (1991) though, participants were trained for thirty minutes prior to the delay experiment which facilitated a high awareness of the difficulties related to long delays. Furthermore, Karis (1991) examined mobile connections where participants may have had lower expectations regarding the quality and thus a lower internal quality-related reference. As a third point, the task used by Karis can be considered as requiring only medium to low interaction speed, and maximum delays tested were not as high as in the study by Kitawaki and Itoh. This is further underlined by another outcome of the quality testing conducted by Kitawaki and Itoh (1991). When testing the detectability, delays were already detected at 90 ms for task 1 by trained participants but only at 1120 ms for task 6 in the case of untrained participants.

It can be concluded that the sensitisation for or awareness of delay and the initial conversation speed required for the task are two important factors regarding the impact of pure delay on CQ. It remains unclear at this point how the initial speed related with a conversation task can be determined otherwise than by intuition. This topic will be addressed in the empirical part of this work (Part II).

About ten years later and after a time of rather low delay with ISDN.type digital telephone services, applications transmitting speech and video signals over the internet were established and the question of the impact of pure delay was raised

again. This time it seemed to be even more important to address the issue of pure delay especially since since VOIP is increasingly replacing an other type of voice transport mechanism. The underlying process of packetisation, sending data via internet connections and queuing the packets to play them out in the right order, makes the occurrence of delay very likely. When designing a codec for this purpose, the 'available time frame' for coding and decoding as well as possible data compression and decompression is essential. When the processing time of the codec or end device is long or if there is the need for substantial buffering, the process of waiting for unreceived packets will result in a long delay or packet drops.

On the basis of previous findings, Hammer (2006) used different task types to assess the impact of transmission delay for CQ in IP-based telephone connections. It was his goal to use the most valid conversations for the assessment. The Short Conversation Tests (SCTs; ITU-T 2007b; Möller, 2000) are conversation scenarios designed for CQ assessment triggering close-to-natural conversations. They address topics such as booking a flight or ordering a pizza. Participants are provided with basic information, like tables and short statements to facilitate the conversation.

Hammer (2006) also used a more rapid version of the Short Conversation Tests (SCTs), the interactive Short Conversation Tests (iSCTs) which were developed by Raake (2006). They were meant to induce fast but natural telephone conversations. Participants were asked to exchange a set of numerical or lexical information in each iSCTs scenario. Additionally, a clarification phase was part of the conversation. It was initiated by one participant as a matter of one item of information missing for her or him. Even though the interaction speed was supposed to be increased, Hammer (2006) could not find a decrease of CQ for delays up to 1000 ms one-way. In a second experiment, using the Random Number Verification (RNV) task (2. task in Kitawaki and Itoh, 1991), an asymmetrical version of the iSCTs (asymmetrical because one person was given all information), the SCTs and free conversations, no increased effect due to the more interactive iSCTs on the CQ could be found either (delays up to 500 ms one-way). This is surprising because not even the RNV task, expected to require fast interaction, seemed to cause a great decrease in CQ judged by participants, which stands in contrast to the results of Kitawaki and Itoh (1991) and the recommended upper limit of 400 ms for delay by the ITU-T Rec. G.114 (2003a).

Alongside CQ ratings, judgments of perceived conversation flow ranging from 0-100 were collected from participants by Hammer (2006). For the the asymmetrical version of the iSCTs, the results indicate a significant decrease from approx. 88 points at 200 ms delay to 80 point at 500 ms delay. Regarding the naturalness of the calls, free conversations were found to be the closest to natural conversations followed by SCTs, the asymmetrical version of the iSCTs, and last the RNVs.

Subsequent studies confirmed these findings. Holub and Tomiska (2009), for example, tested only SCTs combined with delays ranging from 0 to 900 ms one-way. The MOS rating dropped only by appox. 0.2 MOS points from no delay to 500 ms of delay and further appox. 0.35 MOS points for 900 ms delay. Similarly, Egger et al (2010) found only a slight decrease in CQ for delays ranging from 100 ms to 1600 ms delay. For the SCTs employed, the drop of the MOS was appox. 0.55

MOS points, the iSCT MOS dropped by appox. 0.8 MOS points and the RNV MOS dropped by 1.55 points from the lowest to the highest delay condition. In both studies, the greatest decrease could be observed for delays greater than 800 ms, however, the decrease was presumably not statistically significant (no statistics reported). It can therefore be summarised that in only one study, where participants were aware of the inserted impairment due to training and were interacting fast, an impact due to transmission delay on the CQ could be found (Kitawaki and Itoh, 1991).

2.3 Audio-Video Transmission Delay

Bouch et al (2000) conducted a study to gain first insights into the question which **aspects** are most **important** for users of mediated communication. For each application case, two different usage scenarios were described to the participants. After reading the descriptions, participants were asked to name the most relevant determinant regarding quality. For the case of network-based audio, a business conversation was contrasted to listening to streamed music. The audio-video case was exemplified with the case of listening to a remote lecture while completing another task on the computer, and contrasted to taking part in an interactive tutorial. For both interactive examples, the business conversation and the interactive tutorial, the "speed" was clearly identified as the most important aspect. In contrast, the aspect "smoothness" was most important for music-steaming, and the "reliability" of the audio-video system was most important for listening to a remote lecture. The given answers for the music-streaming and the lecture listening case were generally more distributed over the possible categories than the more interactive scenarios.

Zuberbühl (2003) tried to determine **detection** and **acceptance** values for absolute delay in interactive audio- and videoconferencing applications. Acceptance is a concept closely related to QoE but has a yes-no response format (Möller, 2000; Raake, 2006). Zuberbühl (2003) summarised that either no sharp thresholds exist or that slopes and thresholds vary greatly between people. He also pointed out that the detection and acceptance of delay is strongly dependant on the task. Zuberbühl (2003) suggests that the degree of interactivity can serve as a classification variable for tasks and communication settings. This is in line with the findings by Kurita et al (1994). The detection limit for delays ranging from 0 to 600 ms were clearly different for task a) where participants narrated the numbers from 1 to 10 in turns as fast as possible compared to task b) a free conversation over one minute. It is interesting to point out that no significant difference regarding the detection limits could be observed for the usage of audio-only versus audio-video systems which is why the authors conclude that voice delay is the dominant factor for delay detection. Braun (2003), who was also looking at delay acceptability thresholds, found that people in three-party conversations detected lower delay times than people in two-party interactions. Additionally, the acceptance threshold was reached at lower delay values in three-party interaction when compared to the two-party case.

Real-time network gaming can be considered as a highly interactive way of computer mediated interaction which is why some insights on the impact of delay from this domain will be examined briefly. In gaming the audio-video interaction is very time sensitive because interaction speed is very often the most critical aspect of the game (Beigbeder et al, 2004; Pantel and Wolf, 2002; Quax et al, 2004; Schaefer et al, 2002). For this reason, it is commonly accepted that a delay of approximately 100 ms should not be exceeded (see e.g. Ishibashi and Tasaka, 2003; or Pantel and Wolf, 2002). Computer game interaction essentially differs from videotelephony and conferencing interaction because the players do not interact directly with their own voice or appearance but via their avatar. Following from this, only the behaviours accessible to the avatar can be expressed.

Yamagishi and Hayashi (2006), on the other hand, investigated **videotelephony** in their study assessing CQ in a multi-dimensional manner. They inserted delays up to 1500 ms one-way into two-party videotelephony conversations of participants (name guessing task from ITU-T Rec P.920, 2000). Besides the delay, they also altered the video coding bit rate, the video frame rate and packet loss rate. After each conversation participants responded on twenty-five bipolar adjective scales. This technique is known as *Semantic Differential* (Osgood et al, 1957). They extracted two factors "aesthetic feeling" and "feeling of activity" from the data. The impairment delay was mainly related to the factor 'feeling of activity'.

Tam et al (2012) also did not assess the CQ but the "perceived comfortableness", the "perceived naturalness" and the "likeability of the conversation topic" for free conversations with seven selected topics. The tested delays ranged from 67 ms up to 900 ms. They investigated audio-only and audio-video communication systems and found that all three measures were significantly decreased in both systems for delays of 600 ms or higher compared to delays below 600 ms. Regarding the perceived naturalness, the audio-only conversations were rated about 0.5 MOS points lower than the audio-video conversations for delays between 400 and 500 ms but similarly low for delays of 600 ms or higher. Furthermore, Tam et al (2012) asked participants for the *perceived pace* of the conversation whereby it was perceived to be significantly slower for delays of 600 ms or higher compared to delays below 600 ms. However, the general main effect was to their surprise not significant for this measure.

Summarising these initial considerations, to be able to interact with reasonable speed seems to be a critical aspect for gaining a high CQ in interactive audio-video applications.

A first attempt to investigate the **CQ in two-party video applications** was taken by Bräuer et al (2008) and Egger and Reichl (2009). Unfortunately the selected task, a LEGO ® building task as recommended in ITU-T Rec. P.920 (2000), was not very delay-sensitive (Bräuer et al, 2008). Pure (symmetrical) delays from 100 ms up to 1800 ms one-way were tested but led to a decrease of only one MOS point from the shortest to the longest delay condition. A study by Iai et al (1993) tested the effect of audio-visual transmission delays on the perceived conversational quality for "saying numbers from 1 to 10 alternately" and for free conversations. Unfortunately, peo-

ple were also asked whether they detected the delay and in which modality, which most likely changed the sensitivity towards the impairment, as explained earlier. Nevertheless, they found a quick decrease for the MOS ratings in the number verification task (down to about 1.3 MOS point for 180 ms RTT) when using audio-only or audio-video connections. For the free conversations the decrease of MOS scores was much smaller but also fell down to about 3 MOS points for 900 ms RTT delay (for both audio-only and audio-video). Regarding the modality in which participants had perceived the delay, they could select either the audio-only, video-only or both modalities. Up to about 90 ms RTT most people responded to perceive only audio delays. For higher values people suggested audio-video delays, pure video delays were hardly selected. This is in line with the findings of Kurita et al (1994) who asked people to select the modality that lead to the detection of an audio-video synchronous delay. For delays higher than 90 ms RTT, both channels were selected as being responsible for sensing the delay.

In a study by Hashimoto and Ishibashi (2006) participants played the game rock-paper-scissors over an audio-video connection. One person, the game-initiator, initiated the game by saying "Rock, paper, scissors, go!" and afterwards both people had to show one of the three options with their hands. Time is very important in this game because both players need to show their choice at the same time. Hashimoto and Ishibashi (2006) found that the MOS ratings by the game-initiating person were strongly decreasing down for appox. 2.3 MOS points for the highest delay condition of 140 ms additional one-way delay (approx. 30-40 ms system delay). The ratings of the other person were on the other hand completely unaffected by the delay condition.

The most important result on **audio-video synchronisation** was confirmed in two different studies, both using free conversations (Berndtsson et al, 2012; Hayashi et al, 2007): A synchronised audio-video transmission is more important than avoiding a small delay for one modality only. This outcome was reflected in the CQ judgments of participants, in Hayashi et al (2007)'s study. Berndtsson et al (2012) could not show this in the overall CQ ratings of participants, but it was evident in their rating of the synchronisation. In their setup they connected two participants and asked them to rate the overall CQ, a scale termed "interaction quality", the "perceived synchronisation", the "conversation difficulty" and the "acceptability of the call". When the video was delayed with 500 ms one-way and the audio signal by 200 ms (and thus the audio was 300 ms ahead of the video) the perceived synchronisation was rated on average 1.5 points lower on a five point scale than for a condition where the audio was more delayed with 400 ms but less ahead, only 100 ms, of the video which was delayed with 500 ms. The closer the audio delay was adapted to the video delay (of 500 ms) the higher also was the score on a scale assessing the acceptance of the connection. A study related to this by Kurita et al (1994) assessed the degree to which audio-video desynchronisation is acceptable for conversations. The limit for the acceptable audio-video synchronisation delay was similar to the limit of audio-only delay in audio-video asynchronous conditions.

Berndtsson et al (2012) investigated the **CQ** for groups of four to seven people interacting via an **audio-only or videoconferencing** system. They seated two participants in two separate small rooms and the remaining participants in the group were put in one large room. The rooms were connected with either audio and video or with audio-only delay of 200, 400 or 800 ms. Two types of tasks were used: free conversations and a quiz game. The scales overall CQ, the interaction quality and the acceptability were assessed after each call. When one-way delays of 400 ms were compared to delays of 200 ms they could not find a decrease in the CQ ratings for both audio-only and audio-video setups. In the 800 ms delay condition a small decrease of about 0.5 MOS points could be found for the audio-video setup. The results regarding interaction quality were similar to the CQ results. A slightly larger decrease was found for longer delays on this scale in the audio-video setup (approx. 0.7 MOS point from 200 ms to 800ms condition). When looking at acceptability scores, the 800 ms condition was less acceptable in the audio-video setup than in the audio-only case (approx. 1.25 points on a five point scale).

Schmitt et al (2014) investigated five-party desktop video conferencing calls where each participant was seated in a separate room. Participants conversed freely whilst delays of no, 500 ms, 1000 ms or 2000 ms on top of the initial system delay of approximately 75 ms were inserted. Three scales were assessed: the "overall CQ", the "extent of annoyance" by the delay and the "noticeability" of the delay. Unfortunately, the three questions were asked together which most likely lead to a enhanced sensitivity for the impairment and more critical quality ratings. For this reason it is not a big surprise that a significant effect for CQ and annoyance of delay could be found. Regarding a main effect for noticeability, it did not turn out to be significant, however some post-hoc pairwise comparison were according to the authors. This is surprising because post-hoc tests, when applied correctly, seldom turn out significantly if the main effect was not. Furthermore, participants were clustered into two groups related to their total speaking time and the CQ ratings of active and non-active participants were compared. It was observed that active users rated the CQ lower in the 500 ms compared to the no additional delay condition. For non-active users a decrease from 500 ms to 1000 ms was shown. It can be concluded that the participants of this study must have experienced the 500 ms condition differently. The question on annoyance of the condition followed a similar pattern.

As highlighted in earlier in the context of audio-only applications, these two studies indicate that the impact of delay is not always reflected in ratings on the CQ of the connections but instead sometimes affects the interaction process. The speaking behaviour of the participants in turn can influence users' perception of the CQ.

2.4 Predicting Conversational Quality Based on Conversational Measures

Kitawaki and Itoh (1991) did the first attempt to predict CQ ratings based on conversation structure measures of two-party audio-only conversations. *Objective Quality (Od)* attempted to predict the rated CQ results and is formed by a linear combination of the talkspurt durations T_p and T_{ps}, the reciprocal number of speaker changes R_n and two weights to "minimize the differences in subjective and objective assessment values for detectability of delay", W_1 and W_2 (Kitawaki and Itoh, 1991). A replication or development of their approach is unfortunately not possible because no details on the size or precise nature of the weights were given.

$$Od = T_p + T_{ps}xW_1 + (1/R_nxW_2) \tag{2.1}$$

Nevertheless, a study by Wang et al (2010) took their approach of testing scenarios of different interaction speeds. They created different tasks with varying utterance lengths for the to-be-exchanged information to achieve slow or fast interaction and applied them in two-party videotelephony quality tests. The results showed that the longer the utterance length, the less delay affected the perception of quality. The relationship of MOS and the utterance length was proposed to be a logarithmic function or a quadratic polynomial because they found that in case of longer utterances (mean utterance length $>$ 2 sec) participants were less critical regarding their MOS ratings. It should be mentioned that participants of this study were told to specifically evaluate the delay effect which most likely, similar as in Kitawaki and Itoh (1991), lead to very critical MOS ratings due to greater attention to the technical issue. In real life calls, however, the presence of delay is usually not in focus (more details see below or in Chapter 4). Furthermore, the scenarios employed were mainly relying on audio-based interaction, even though there might have been some turn alignment based on the visual input, it was not actually needed to accomplish the task. Thus, the results explain little about visual interaction. Beyond this, it remains open how the relationship of surface measures, delay and CQ appear in more natural conversations.

Wah and Sat (2009) also aimed to relate quality ratings to a conversational surface structure measure. Their *Conversational Efficiency (CE)* relates the sum of speaking and listening time to the total time of the call (for more details on the CE, see Chapter 3, equation 3.2). Unfortunately, quality ratings were only obtained in listening tests where participants listened to computer generated conversations that didn't allow for *double talk* (two parties talking at the same time). Even though the authors had their reasons for this approach, it is widely known that the evaluation of transmission delay can realistically only be carried out with conversational tests. Quality ratings differ when based on the listening or conversation tests (Guguin et al, 2008; International Telecommunication Union, ITU-T, 2007b; Karis, 1991). The specific problem with the CE metric proposed by Wah and Sat (2009) is that it remains unclear if double talk, as it occurs in human conversations, would be

counted as speaking or listening time. This in turn may change the predictive value of the measure.

A third attempt to predict quality based on conversational surface structure measures was done by Issing and Nikolaus (2012). In their work, RNV tasks were adapted in order to gain different levels of *Speaker Alternation Rate (SAR)*. The SAR has initially been described by Hammer et al (2004) to count the successful speaker changes in a conversation per minute interval (for more details on the SAR, see Chapter 3, Fig. 3.2). In Issing and Nikolaus (2012), different SAR levels were obtained by altering the rate of matching numbers in the number verification task. Furthermore, the instructions were changed in such a way that participants only needed to reply for non-matching numbers. This implied that the higher the number of mismatches, the more often participants needed to interact. A corresponding task with different levels of matches was also created for a text proofreading task. As a control task, the SCT flight booking was included. Tested delays ranged from 100 ms to 800 ms one-way. Lowest SAR values were found for the text verification tasks, the SCT scenario showed a medium SAR and all RNVs SARs were found to be the highest. As intended, the higher the mismatch rate in text or number verifications was the higher was the measured SAR. Based on the data, three SAR classes were built and linear models were applied to verify the impact on the CQ. A statistically significant impact of the SAR class and the factor delay could be found for CQ ratings, with the factor delay contributing slightly stronger to the prediction.

2.5 The E-Model

Instrumental models calculate an estimation of the CQ based on different approaches. In general, signal-based models such as PESQ (Perceptual Evaluation of Speech Quality; ITU-T Rec P.862, 2001) or POLQA (Perceptual Objective Listening Quality Assessment; ITU-T Rec. P.863, 2014) need to be distinguished from parameter-base models such as the E-Model (ITU-T Rec. G.107, 2011a). Classical signal-based models predict the quality by comparing a non-degraded to a degraded speech signal while parameter-based models use individually measurable parameters to give estimations. The **E-Model** can be considered as the most popular model in the context of telephone network planning (International Telecommunication Union, ITU-T, 2011a; Möller and Raake, 2002; Raake, 2006), providing estimates for two-party interaction. It assumes that degradations affecting certain parameters can be transformed to the Transmission Rating Scale called R-Scale. On the **R-Scale**, impairment factors which are built based on the defined parameters are assumed to be additive, see equation 2.2. Using an S-shaped relationship, R-values can be transformed into MOS-values (see ITU-T. Rec G.107, 2011a).

$$R = R_0 + Is - Id - Ie,eff + A \tag{2.2}$$

R Transmission Rating Scale
R_0: basic signal-to-noise ratio

Is: Simultaneous Impairment Factor
Id: Delayed Impairment Factor
Ie,eff: Effective Equipment Impairment Factor
A: Advantage Factor

In this context, the delay impairment factor, *Id*, is of particular interest. It covers all impairments due to a delayed signal. Sub-factors are added to calculate *Id*. The first sub-factor, *Idte*, calculates the impairment due to talker echo, the second one, *Idle*, does the same for listener echo and the third one, *Idd* accounts for the impairment due to pure delay without any echo. In this work, the latter factor *Idd* is of interest because only the impact of pure delay without considering any audible echo will be examined in this work.

$$Id = Idte + Idle + Idd \tag{2.3}$$

The parameter *Ta* represents the absolute one-way delay of the entire transmission path and is most important for calculating *Idd*. For $Ta \leq 100$ ms, *Idd* is set to zero because the model assumes no impact on the quality in this case. If the pure delay, *Ta*, is longer than 100 ms the following equation is defined:

For Ta <= 100 ms:

$$Idd = 0 \tag{2.4}$$

For Ta > 100 ms:

$$Idd = 25\left\{\left(1+X^6\right)^{\frac{1}{6}} - 3\left(1+\left[\frac{X}{3}\right]^6\right)^{\frac{1}{6}} + 2\right\} \tag{2.5}$$

$$with: X = \frac{log_{10}\left(\frac{Ta}{100}\right)}{log_{10}2} \tag{2.6}$$

The E-Model, originally suitable for Narrowband (NB), has also been extended to Wideband (WB) showing that an extension of the R-scale up to 129 is needed (Raake, 2006). In the case of NB, R-values lie in the range from 0 to 100.

2.6 Extension of the E-Model

Raake et al (2013) proposed to **extend the E-Model** to be able to capture aspects related **to the type of interaction** that is undertaken over the (echo free) delayed line. As will be described in detail later, the effect of transmission delay on the perceived CQ is strongly related to the type of interaction and the sensitivity of the participants to the impairment. For this reason, two new parameters were included: ***sT*** the delay sensitivity and ***mT*** the minimal perceivable delay. The parameter *mT* replaces the

former value of 100 ms to be able to have lower or higher values according to a particular interaction type (see equation 2.7). For example, for slower interactions mT may be higher than 100 ms because people cannot detect the small delay due to the slow interaction. The parameter sT captures the degree to which people attribute the impairment to technical system quality, which also includes how much they reflect on the perceived quality change in their quality judgments (see equations 2.8-2.9).

For Ta $<= mT$:

$$Idd = 0 \tag{2.7}$$

For Ta $> mT$:

$$Idd = 25 \left\{ \left(1 + X^{6 \cdot sT}\right)^{\frac{1}{6 \cdot sT}} - 3 \left(1 + \left[\frac{X}{3}\right]^{6 \cdot sT}\right)^{\frac{1}{6 \cdot sT}} + 2 \right\} \tag{2.8}$$

$$with: X = \frac{log_{10}\left(\frac{Ta}{mT}\right)}{log_{10}2} \tag{2.9}$$

Raake et al (2013) calculated three different models to fit their CQ NB data to the new equation. Their data was composed of the experiment A.NB.2a presented in this work, and the data of a study by Egger et al (2010). In the first model (#1) computed, both parameter were free, in the second (#2) model, mT was set to 150 ms for SCTs and iSCTs and to 100 ms for RNVs and RNTs. In the third model (#3), mT stayed at its default of 100 ms for all task types and only sT was fitted. The determined mT and sT values for experiment A.NB.2a are shown in Table 2.1. The according prediction of the data with fitted mTs and sTs afterwards lead to a much better prediction as the former version of the E-Model. Thus the authors recommended the

Table 2.1: Fitting Results from Raake et al (2013), for experiment A.NB.2a; values in small letter, e.g. $_{100}$, indicating that they were preset

Task Type	**Band-width**	**Group Size**	**Order**	**Model**	***mT* [ms]**	***sT***
SCT	NB	2	rand.	#1	175.5	0.78
				#2	$_{150}$	0.61
				#3	$_{100}$	0.42
RNV				#1	117.5	0.84
				#2&3	$_{100}$	0.69
RNT				#1	114.9	1.24
				#2&3	$_{100}$	0.98

Table 2.2: Delay-sensitivity classes for different use cases, adapted from "E-model update regarding delay and interactivity Rec. ITU-T G.107", 2013

Class of Delay-Sensitivity	*sT*	*mT* in ms	**Use Case**
default	1.00	100	Applicable to all types of telephone conversations Must be used for: Carrier-grade fixed or mobile telephony Enterprise-grade fixed or mobile telephony When targeted user group and delay requirements are unknown
low	0.55	125	Applicable only in cases where it is known that users have low sensitivity to delay, e.g. non time-sensitive conversation scenarios.
very low	0.40	150	Applicable only in cases where it is known that users have very low sensitivity to delay, e.g. in primarily non-interactive cases, such as to mainly listen into a conversation or to a lecture.

use of the new parameters as explained in Table 2.2 (taken from "E-model update regarding delay and interactivity Rec. ITU-T G.107", 2013).

To be able to use the new model, it will be beneficial to confirm *mT* and *sT* values for other NB data sets and to see whether the values change for WB data. Furthermore, it should be investigated whether they change for three-party conversations, even though the E-Model was originally designed for quality prediction of two-party interactions.

Raake et al (2013) also related *mT* and *sT* to a parameter addressing the structure of the recorded conversations and found a fitting mapping function. More insights on suitable conversation metrics to predict *mT* and *sT* will help to increase understanding of when delay problems manifest themselves in CQ-ratings. Further, with the link to conversation structure metrics, it is possible to monitor CQ in running calls.

An equivalent **model** to the E-Model **for video-telephony** is based on a parametric estimation of the video-quality, audio-quality (based on the E-Model) and an multimedia integration function of both (ITU-T Rec. G.1070, 2012b). The pure audio delay parameter *Ta* (or in that context *Ts*, *s* for speech) and the pure video delay parameter *Tv* are considered in the multimedia integration function only. The audio-visual impairment factor MM_T estimates the impact of audio-visual delay *AD* and synchronization *MS* on the estimated quality. The function is still under investigation (see e.g. Hayashi et al, 2007; Belmudez and Möller, 2013). More details on the current version can be found in the ITU-T Rec. G.1070 (2012b).

2.7 Conclusion

From the above literature review it can be concluded that two factors seem to determine the extent to which transmission delay affects the perceived CQ of a connection. First of all, it is important to know how sensitive to, or aware of the impairment users are. This includes, on the one hand, prior training, experience or knowledge of the technical conditions. On the other hand, it addresses the awareness of a technical problem in the usage situation itself. The major issue with delay-only is that it is not identified solely as a technical impairment. As Riesz and Klemmer (1963) stated: "Indeed during most of a conversation over a delayed circuit there is no degradation. When degradation does occur it can often be misinterpreted by the user as being due to the other speaker. Slow responses, excessive interruptions and complete failures to respond (because the question was lost in the circuit) are examples of such difficulties." Misinterpretations or misattributions regarding the delay source, as described in this quote, will be addressed in more detail in Chapter 4.

Secondly, the initial speed that an interaction via a communication system requires, and to which extent this speed can be maintained, seems to be of importance for the CQ judgment. Insights into the relationship of CQ ratings and conversation surface measures were described. However, approaches are incomplete. In Part II, this will be examined further. For a better understanding of possible relationships, the following chapter will go into detail about the literature on conversational measures.

Chapter 3
Mediated Interaction

This chapter explores the question how communication behaviour changes with the presence of pure delay in Mediated Interaction, and how, in particular, these changes can be quantified. The focus will be on nonverbal measures of vocal communication that can be assessed in an automated manner. This component is considered to be important because transmission delays are not always identified as a technical issue (Brady, 1971; Klemmer, 1967; Riesz and Klemmer, 1963; Schoenenberg et al, 2014b; Vartabedian, 1966) and may thus not be reflected in CQ judgments even though they may severely affect the interaction.

In the following, at first the terms *interaction* or *interactivity* will be defined. Subsequently, findings from the literature on how the interaction changes under pure delay will be reported.[1]

3.1 Defining Interactivity

There are numerous definitions of **interactivity**. Depending on the scholar of research, different aspects are important (for a detailed discussion see, for example, Kiousis, 2002). In Kiousis' (2002) classification, some researchers see interactivity as a property of technology, others define it in terms of its subjective nature. A third category of theoretical considerations defines interactivity based on the communication setting. Here, the latter view will be taken because the measurement of the communication processes is in focus.

As Stromer-Galley (2004) points out, two cases need to be distinguished when discussing research on interactivity which are based on different types of interaction partners. If the interaction partner is another human (human-to-human), the research will be oriented towards the process of interactivity. The communication is always two-way in this kind of research. On the other hand, the interaction partner can be a computer (human-to-machine). In this case, interactivity can be considered as a

[1] Parts of this chapter have been published before in Schoenenberg et al (2014a)

product and two-way communication is not required for it to occur, it can be only one-way (Gellri and Kanning, 2007). [2]

When defining interactivity, one aspect, termed *third order dependency*, is mentioned by different researchers and seems to be of highest importance. Third order dependency implies that interaction is not simply a reaction but a cyclic process relying on reciprocity (Crawford, 2003; Rafaeli, 1988; Rice, 1987). In a strict sense, individual turns and roles of interaction partners should be interchangeable. Besides reciprocity, the ability to induce feedback is mentioned as another important aspect of the interaction process (Rafaeli, 1988). Researchers do not completely agree on what should be considered as feedback. Nevertheless, most agree that one important function of feedback (among others) is to send or receive information indicating attention (Kendon, 1967) and comprehension (Kraut et al, 1982). Thus, it always refers to prior messages.

The concept of interactivity can thus also be discussed in the context of the more global concept of communication because interaction implies that the behaviour of interactants is dependant on each other (Nöth, 2000). This assumption is not necessarily true for communication, but it can be if the communication is interactive.

For the scope of this work, the following definition from Egger et al (2014) for ***interactivity*** is chosen:

"An interactive pattern is a sequence of actions, references and reactions where each reference or reaction has a certain, ex-ante intended and ex-post recognisable, interrelation with preceding event(s) in terms of timing and content."

The concept of interactivity does not need to be reduced to two-way communication, but can be applied in the same sense for multi-way communication (Kiousis, 2002).

To summarise, the view of interaction as a process best fits to the context of mediated audio-only or audio-video communication. Participants do not interact solely with a machine in these applications but over a machine or system with one or several other humans, which is why the interaction process and its relationship to the technological settings (the machine), especially the transmission delay time, will be examined. For describing the process of human-to-human interaction, it may be valuable to take a closer look on how conversations are organised in general. For a better understanding on the organisation of turns, an excursus to conversation analysis will be given in the next section.

3.2 Conversation Analysis

One of the most known articles on conversation analysis by Sacks et al (1974) describes the basic organisation of conversations in the context of small group be-

[2] For a discussion of similarities and differences between human-to-human and human-to-machine interaction, the reader is referred to Egger et al (2014) and Stromer-Galley (2004).

haviour. The most essential concept for the organisation of conversations is **turn-taking** which can be considered as a general rule system of how turns are allocated. An important presumption of this system is the *invariance to parties*. It implies that whatever variations are brought to a conversation through contributions, they will be accommodated without the need for change of the rule system.

Furthermore, the organisation system is characterised to be a *local management system* which is *interactionally managed* and *party administered* (Sacks et al, 1974). The importance of local management is emphasised by the authors. More precisely, turn size and turn order are managed locally which means they are managed from one turn to the next, and are not predefined but variable. The participants control the turn-taking (party administered). They do this in an interactive way (interactionally managed), meaning that conversation behaviour permits for projection of possible completion of an utterance; allows the other one to begin their turn at a transition place; or allows the commencement of ones own talk, placed at a point where it may show the other one to stop talking (or not).

The model is described as consisting of turn-construction and turn-allocation components. Various unit types (sentential, clausal, phrasal, lexical) can be selected by a speaker to construct a turn. For the allocation, two types need to be distinguished: the *current speaker selects the next speaker* or *self-selection*.

A set of rules determines the local allocation of the next turn in the most efficient way (adapted from Sacks et al (1974), page 704):

"(1) For any turn, at the initial transition-relevance place of an initial turn-construction unit
(a) If the turn-so-far is constructed as to involve the use of the current speaker selects the next speaker technique, then the party so selected has the right and obligation to take the next turn to speak; no others have such rights or obligations, and transfer occurs at that place.
(b) If the turn-so-far is so constructed as not to involve the use of a current speaker selects the next speaker technique, then self-selection for next speakership may, but need not, be instituted; first starter acquires rights to a turn, and transfer occurs at that place.
(c) If the turn-so-far is constructed as not to involve the current speaker selects the next speaker technique, then current speaker may, but need not continue, unless another self-selection.

(2) If at the initial transition-relevance place [...], neither 1a nor 1b has operated, and, following the provision of 1c, the current speaker has continued, then the rule set a-c re-applies at the next transition-relevance place, and recursively at each next transition-relevance place, until transfer is effected."

Sacks et al (1974) further propose that the model of **turn-taking organisation** is and needs to be capable of the following points (adapted from Sacks et al (1974), pages 700-701):

- speaker-changes

- one party talking
- more than one party talking at a time
- transitions from one to the next turn with no or only slight gaps or overlaps
- variable turn order, but not randomly
- variable turn size
- conversation length not predefined
- what is said is not predefined
- relative distribution of turns is not predefined
- variable number of parties
- continuous or discontinuous talking
- turn allocation techniques are used
- different turn construction units are used (one word or a whole sentence)
- repair mechanisms are used for dealing with turn taking errors and violations

The last point of this list, the organisation of **repair mechanisms** was further explained by Schegloff et al (1977). The authors describe the self- and other-correction as a subtype of the repair mechanisms. Strictly speaking a self- or other-correction replaces an error or mistake by what is thought to be correct.

Schegloff et al (1977), however, pointed out that the phenomenon is not contingent upon error which means after an error, a correction or repair does not necessarily need to follow. Additionally, there may be a correction or repair even though there was no error. Repair mechanisms should not be limited to simple word replacement and can be more multifaceted. Schegloff et al (1977) distinguish between self- and other-initiated repairs which are characterised by the person placing the repair. The self-initiated repair is done by the speaker who caused the trouble in a conversation and the other-initiated repair is done by any other party besides the one who caused the trouble. These two kinds of repairs usually differ in the placement relative to the incident of error and the techniques used to initiate the repair. Self-initiated repairs are mostly placed within the same turn as the error incident and use non-lexical means, such as cut-offs or sound stretches. On the contrary, other-initiated repairs are always placed in a different turn than the one which contains the error and use techniques such as partially repeating the error incident possibly followed by a question. Regarding the position for placing an initiation, three cases are described: in the same, first, turn (self); in the transition space between turns (self); at the next, second, turn (other) or in the third turn (self, in multiparty possibly also other). Due to the fact that other-initiated repairs always require a new turn, the authors postulate that self-initiation is preferred by the interactants. It should be mentioned that both types of repair initiation, self and other, are related and should not be seen as distinct categories. They are both possibilities in handling the same communication trouble. Examples for repairs are word replacement, a repair when referring to anther person or a repair when selecting the next speaker. Whether such repairs are successful or fail is a question of the outcome of a repair, which needs to be distinguished from the sole initiation organisation.

Jordan and Henderson (1995) emphasised the need to extend the pure speech-based analysis of an interaction to an analysis of the entire interactional exchange

system which also considers, for instance, *turns with bodies* and many other turn-taking behaviours. The works of different researchers (i.e. Argyle et al (1968); Aran and Gatica-Perez (2011); Battersby (2011); Gatica-Perez (2006); Jokinen (2011)) have addressed such **visual turn-taking** behaviour in face-to-face communication. Argyle et al (1968), for example, assessed the impact of visibility of the interaction partner in face-to-face conversations and found that with decreasing visibility speech pauses became longer and interruptions more frequent. Synchronising the turn-taking was best if only the eyes or the bodies could be seen. In line with Kendon (1967) they concluded that eye movement together with head nods are the most important factors for floor control.

Jordan and Henderson (1995) further described vision-based and speech-based interaction on a continuum. One end, termed "talk-driven interaction", is the interaction constituted by speech, and the other end, termed "instrumental interaction", is the interaction that requires the accomplishment of a physical task involving a lot of visual interaction. Examples for the later can be a surgery or a car repair where speech communication alone is not sufficient but where the alignment of action is required.

Communication over an audio-only communication system should be considered as talk-driven interaction due to the lack of any other than speech (vocal) input available. For the case of audio-video communication systems, the placement is not as clear. On the one hand, it depends on the task people are accomplishing over the system and, on the other hand, the richness of information delivered by the video system most likely varies the importance of visual interaction.

Coming back to the scope of this work, transmission delays are likely to disturb the described turn-taking organisation (if speech based or non-speech based). Since delay causes different conversational realities at the different ends in terms of timing, turn-taking organisations are likely to fail and repair attempts will most likely be needed. They in turn are likely to fail, too, due to time shifts between the different realities. This will have various consequences on structural aspects of the conversation. It can be expected that conversations take longer, more words are needed to finish, longer time of speech overlap will appear. In the following, first results from a micro-analysis of conversations under transmission delay by Ruhleder and Jordan (1999, 2001) will be reported. The micro-analysis technique is based on a detailed linguistic exploration of the turn-taking organisation. Not the interaction structure in terms of speech on-off pattern is considered but the speech content in respect to its functionality for turn-taking organisation. Afterwards, findings from the literature supporting the proposed impact of delay on the interaction structure will be considered in more detail.

3.3 Micro-Analysis of Conversations under Transmission Delay

Ruhleder and Jordan (1999, 2001) analysed conversations over a videoconferencing system based on transcriptions of recorded video tapes suffering a delay of approximately 1000 ms. Recordings were done at both ends of the system. From the micro-analysis, they detected several phenomena caused by the transmission delay. For example, in cases where the previous speaker does not select the next speaker at the end his or her turn, and the connection is delayed, continuation of the previous speaker and self-selection of a next, different, speaker can collide. Following this, the respective other person is often interrupted at his or her side even though this was not intended and is probably not perceived as an interruption by the causing person. This phenomenon is commonly called *unintended interruptions* and will be discussed in detail again below. Ruhleder and Jordan (1999; 2001) also observed that people generally tend to rephrase their utterance more often probably because they believe, so the authors, that the other will react negatively. This belief may arise from long silence caused by delay, before reactions that were expected earlier. Long silences in turn imply that the person thinks longer before responding, which is often related to a negative attitude towards the prior utterance. In addition to this, words get swapped and feedback often appears to be misapplied, both phenomena can be ascribed to delay. Repair attempts also tend to fail because one side does not recognise them due to the incorrect timing of their arrival.

When looking at delayed communication on the micro-level, it can be summarised that the entire turn-taking organisation becomes unstable.

3.4 Interaction Structure under Transmission Delay

For analysing the interaction structure, some of the approaches reported below require a **conversational state model**. The commonly used state model for two-party interaction is based on Markov Chains. Four states are distinguished which are assigned to the speech on-off patterns derived from recorded conversations:

$S_{2,0}$: silence
$I_{2,A}$: individual/single talk of person A
$I_{2,B}$: individual/single talk of person B
$M_{2,AB}$: multi/double talk of person A and B

These states and their possible transitions are illustrated in Fig. 3.1. Transitions are defined for every sample of a recording, which makes a transition from no interlocutor talking to both interlocutors talking at a time (or reverse) very unlikely, this case therefore is indicated by a grey arrow.

A number of different approaches have been described in the literature to examine the structure of conversations affected by transmission delay (caused, e.g., by

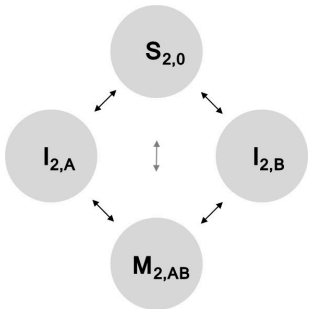


Fig. 3.1: Possible states (sets and subsets) for a two-party conversation situation; $S_{2,0}$: silence, $I_{2,A}$: individual/single talk of person A, $I_{2,B}$: individual/single talk of person B, $M_{2,AB}$: multi/double talk of person A and B; State transitions indicated by arrows. Improbable transitions indicating grey arrow.

satellite or packet-switched transmission). Table 3.1[3] summarizes all conversational parameters considered relevant, and are discussed in more detail in this thesis. Most of the described studies are based on assessments of audio-only communication systems, only a few studies address audio-video connections.

One of the earliest and most simple parameters used to describe the impact of transmission delay of the call is the **call duration (DUR)**. Vartabedian (1966) examined the DUR of three-party calls. In the delayed condition, the connection of one end to the conferencing bridge was impaired by 300 ms delay, the connection of a second end received 600 ms of delay and the third connection had no delay. In contrast to a condition where none of the connections were delayed, participants took significantly longer (28 %) in the delayed condition to complete a figure matching task. Riesz and Klemmer (1963) investigated delay with and without echo and echo suppressors in naturally occurring two-party telephone conversations of laboratory administrative staff. After each call the staff were asked to rate their satisfaction of the call conditions. The results showed a much higher probability for long calls to be perceived as unsatisfactory ($\geq$ 2.5 min, 51 %) when compared to short calls ($\leq$ 0.5 min, 8%). This gives a strong indication that the DUR is not a clear indicator for the quality of a call. On the contrary it appears that in the particular business context of their study, the calls were somewhat unsuitable when they lasted longer than expected primarily due to the induced delay. This means calls were perceived more unsatisfactory due to the delay condition and not due to the longer DUR, which was only a consequence of the delay condition. In a different context the better the connection that is provided, the longer people talk (Skype, 2010). In this case, the probability to perceive short calls as unsatisfactory can be expected to be higher. Still, the change of the DUR under disturbed versus undisturbed conditions remains an interesting aspect. Hammer et al (2004), for instance, found that the DUR satu-

[3] This Table was partly published before in Schoenenberg et al (2014a)

rates for one-way delays starting at 360 ms when using close-to natural interactive Short Conversation Test scenarios (iSCTs). This outcome indicates that people find a way to communicate equally fast even though the transmission conditions complicate the conversation. Unfortunately, the DUR parameter does not tell us how people cope with long transmission delays and can therefore not serve as a good indicator of delay induced conversation problems.

Krauss and Bricker (1967) were one of the first to give insight into how peoples' behaviour changes with increasing transmission delay. They assessed the **number of words needed (NoW)** to accomplish a shape-ordering task, and found a significant increase of words used by the sending person for 1800 ms RTT delay, in comparison to no or 600 ms RTT delay conditions. Furthermore, participants of their study

Table 3.1: Conversation parameters described in the literature

Duration of a Call	DUR	Riesz and Klemmer (1963), Vartabedian (1966), Hammer et al (2004)
Number of Words	NoW	Krauss and Bricker (1967); Krauss et al (1977)
Number of Back-Channels	NoBC	O'Conaill et al (1993)
State Probabilities	P	Brady (1965, 1968, 1971) O'Conaill et al (1993), Braun (2003), Geelhoed et al (2009)
State Sojourn Times	SOJ	Hammer et al (2004)
Speaker Alternation Rate	SAR	Hammer et al (2004) Egger et al (2010)
Conversational Efficiency	CE	a) Kitawaki and Itoh (1991), b) Sat et al (2007), Wah and Sat (2009)
Involuntary/Unintended Interruption (Rate)	UI(R)	Richards (1962), Ruhleder and Jordan (2001), Egger et al (2010), Tam et al (2012)
Intended Interruptive Interruptions (Rate)	III(R)	Egger et al (2012)
Transition Tendencies	TEND	Brady (1971)
Conversational Synchrony	CS	Sat et al (2007), Sat and Wah (2009)
Conversational Interactivity	CI	Sat et al (2007)

reported significantly more “difficulties in communicating due to the circuit”, they rated their partners to be less attentive, and also detected the delay more often in the 1800 ms condition. In a similar investigation by Krauss et al (1977), participants used more words for accomplishing the task, when approximately 1000 ms of access delay was inserted in comparison to no delay. Access delay is a particular kind of delay. It can be described as the time a person has to wait before being able to access the audio channel after the turn ending of the current speaker. When adding a non-delayed video channel to the conversations, the effect of increased number of words was substantially mitigated (Krauss et al, 1977). The authors explain this by the possibility to send and receive visual back-channels in the audio-video setting. The visual feedback seemed to have functionally replaced the speech-based feedback. Consequently, the speaking person felt understood and did not use more words for the explanation. As it was shown by Kraut et al (1982), the more feedback that was given to persons telling a story the more comprehensible their narratives were. In their study, both conditions were without any delay but it is likely that there is also less feedback in delayed conversations, and thus the comprehensibility may be lower.

O’Conaill et al (1993) looked into the **number of back-channels (NoBC)** and found that participants used less back-channels in a half-duplex connection of low video quality with delays of 410 ms to 780 ms in comparison to a full-duplex audio broadcast of a high quality video connection. Furthermore, they found significantly less turns with more words per turn, resulting in a greater turn length. The lack in feedback could have allowed speakers to continue talking and elaborate further.

The actual surface structure of telephone communication was first considered by Brady (1965, 1968, 1971). He developed a state model to assess conversations, and tested its properties under different transmission conditions. The conversation analysis revealed longer **sojourn times (SOJ)** in the states of *double talk*, *mutual silence* and *pause in isolation* for 600 ms and 1800 ms RTT delays than in the no delay condition. With ***mutual silence***, he referred to all conversational states where no speech can be detected in the recording. It included ***alternating silences***, which are silences occurring before the speaker changes, and speaker ***breaks***, which are silences occurring before the same person starts talking again (Hoeldtke and Raake, 2011). Brady called the latter *pause in isolation*, but here it will be referred to as speaker break. At this point it can be highlighted that not only the alternating silences were found to get longer, but also the breaks. Regarding audio-video connections, Braun (2003) found an interesting U-shaped relationship of delay and the **state probability (P)** for *double or triple talk* time. In particular, he differentiated three stages: the first, for delays up to 700 ms RTT, where the percentage of double talk increases slightly with increasing delays, the second, up to 1700 ms RTT, where the percentage of double talk drops again, and the third stage for higher delay values, where it rises again. He explains these results as follows: In the first stage, double talk is mainly caused by back-channels and by the seamless (with slight overlap) beginning of the next speaker, both facilitating a fluent conversation. This explanation is in line with Goodwin and Heritage (1990), who explain that overlaps are a natural part of

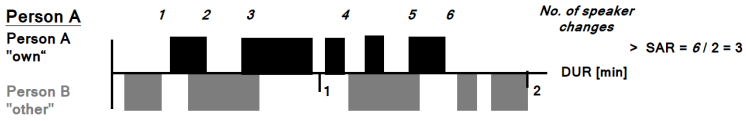


Fig. 3.2: The Speaker Alternation Rate (SAR) calculated as the number of speaker changes per minute

conversations and are initially non-destructive [4]. In the second stage, the delay diminishes the occurrence of back-channels and seamless turn hand-overs, however a mostly undisturbed conversation is still possible. For delays over 1700 ms RTT (or approx. 850 ms one-way), the amount of double talk increases again because the number of interruptions increases. In this stage, more effort needs to be taken for coordination of the conversation. Similar conclusions could be drawn for triple talk, however, in Braun (2003) there was not enough data to give statistically relevant answers. It should be kept in mind that the found relationship could be related to the used threshold method.

Geelhoed et al (2009) analysed the amount of double talk for conversations done over a Halo telepresence system comparing the initial system delay to an added 2000 ms RTT delay. For three out of four observations an increase in double talk was found, however, the fourth showed a decrease. As participants of this study were aware of the inserted delays due to the design of the questionnaire, different groups may have adapted their speech behaviour differently.

The reason for speech overlaps (double talk) percentages to increase or decrease was discussed by O'Conaill et al (1993). They found a lower percentage of overlaps resulting from projections/completions per turn and a higher percentage of overlaps resulting from simultaneous starts when delay was present in the audio-video communication system.

Hammer et al (2004) took up the idea of Brady (1968; 1971) to look at conversational state walks. In their work, Hammer et al (2004) focused on a particular subset of state walks that lead to speaker changes. Two cases can be distinguished here, ***alternating silences*** which are silences between speaker changes and ***interruptions***, which are the interruptions where single talk of the interrupter follows after the double talk. The **Speaker Alternation Rate (SAR)** includes both possibilities and relates them to time (see Fig. 3.2). However, as Hammer himself discerns, the SAR decreases with increasing delay, but as the increase of mutual silence saturates at 360 ms delay, so does the SAR in his study.

Relating occurrences of certain events or task outcomes to time is beneficial if the aim is to examine the efficiency of conversations. Kitawaki and Itoh (1991), for example, defined **Conversational Efficiency (CE)** as the percentage of completion

[4] The statement on the non-destructiveness of speech overlaps was addressing face-to-face interaction

(x %) of a task during a given time window in the presence of transmission delay as compared to percentage (y %) of the same type of task without the impairment, see equation 3.1. They reported a greater drop of CE for tasks like verifying random numbers or completing words with missing letters in conditions of 500 ms RTT delays compared to conditions of no delay. Sat et al (2007) and Wah and Sat (2009) defined CE based on conversational states and related the sum of the overall speaking time and the overall listening time (listening to someone else speaking) to the total time of a conversation, see equation 3.2. They also reported a drop of CE with increasing delay.

$$CE_{V.1} = x\ \%_{delayed} / y\ \%_{non-delay} \tag{3.1}$$

$$CE_{V.2} = \textit{(Speaking Time + Listening Time) / DUR} \tag{3.2}$$

In conclusion, CE-measures generally drop under transmission delay, which means less information can be delivered in the same amount of time. This negative relationship is evident since delay lengthens the time during which the same amount of information can be delivered. Despite the value to assess the amount of decreased efficiency, these measures do not tell us what consequences delay exactly has on the communication structure.

Besides longer alternating silences, other reasons have been reported which explain why conversations take longer when transmission delay is induced. Brady (1971), for example, observed **confused situations** significantly more often for circuits of 600 ms and 1200 ms RTT delay than for circuits with no delay. He spoke of a confused situation if, after an involuntary (or unintended) interruption, one of the interlocutors stopped his normal speech flow to readjust to the conversation. In line with this, in the study by Vartabedian (1966) participants complained more often about "talking together" in conversations with delay, and Tam et al (2012) found that the perceived number of interruptions increased significantly with delay (one-way delays up to 900 ms). The phenomenon of **unintended interruptions** is also typical of delayed circuits and was first mentioned by Richards (1962). It describes the situation where a response of the respective other person does not arrive within the silence of the current person, but within his speech and induces a speaker change, even though it was not perceived and not meant as an interruption by the causing person at the far-end (see Fig. 3.3). Egger et al (2010) decided to look at this phenomenon in more detail and investigated the **Unintended Interruption Rate (UIR)**. The UIR calculates the number of unintended interruptions (or more precisely *interruptive unintended interruptions*) per minute that take place within a conversation with a particular delay (see Fig. 3.3). As expected, the UIR increases with increasing transmission delay. A complementary measure to the UIR is the rate of **interruptive intended interruptions (IIIR)** which counts the number of interruptions that were delivered intentionally to the other end (Egger et al, 2012). Observations confirmed the expected decrease with increasing delay (Egger et al, 2012). Consequently a third measure the **non-interruptive intended interruptions**

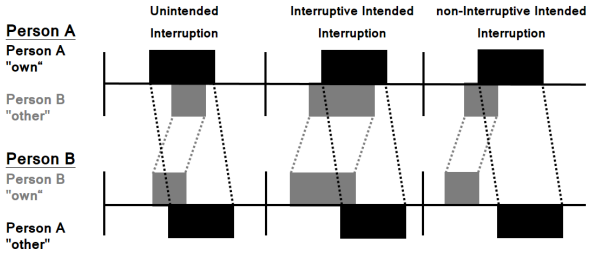


Fig. 3.3: Different interruption types, by Person A

(nIIIR) could be introduced, which counts those intended interruptions of a person that are not delivered as interruptive to the far-end person due to transmission delay (see Fig. 3.3). Investigating unintended interruptions and non-interruptive intended interruption rates appears to be a meaningful way to describe the divergence of the two perceived realities. Nevertheless, in line with the critique for the SAR and other CE-measures, in this work absolute occurrence values are considered to be more appropriate because they do not dependant on the overall duration.

Getting back to the roots of conversation analysis for telephone conversations, Brady went far beyond simple state probabilities and sojourn times with his on-off pattern analysis (Brady, 1968, 1971). One aim of the on-off pattern analysis was to calculate **tendencies (TEND)** for certain state transitions. This gave a similar but more precise view on the situation when delay is present. With delay the probability to stay in a silent state before starting to talk again, and the probability to stay in a double talk state after being interrupted were increased. These outcomes on the TEND were the same for both sides.

Let us imagine person A stands for the person of the current viewpoint at the near-end talking to a person B at the far-end. Brady (1971) discovered an interesting discrepancy on how person B was perceived by person A, in contrast to how B actually behaved. In particular, he found a greater probability for person B to be perceived as interrupting and taking longer to respond when in fact B did not attempt to interrupt more often and responded similarly quickly. These findings were the first to show that with delay the realities of interlocutors diverge.

Conversational Synchrony (CS) is a measure following a similar approach to the afore mentioned and was described by Sat et al (2007) and Wah and Sat (2009). It is calculated as the ratio of the maximum and the minimum silence perceived by a particular person (i.e. Person A) for a given time interval, see equation 3.3.

$$CS_A = max(S)_A / min(S)_A \tag{3.3}$$

The motivation is as follows: When delay is present in a connection people generally need to wait longer to receive a response. After waiting for a while the wait-

ing person will either begin to speak again (in this case, silence is perceived as an own break) or, if he continuous to waited he may receive a response by the interlocutor (person B, according to the example above), an alternating silence. In both cases, the silence in between two talk spurts is likely to get longer (maximum silence). If, on the other hand, this person (person A, according to the example above) is responding he will do so immediately (minimum silence). Consequently, CS is thought to reflect greater asynchrony or divergence of the two perceived conversational courses for long delays. Sat et al (2007) proposed another similar parameter, the **Conversational Interactivity (CI)**. It is calculated as the ratio of the alternating silence duration (durAS) that a person A waited after person B finished to speak (durAS.BA) and the alternating silence duration of the next turn that was taken by the other person B after person A finished to speak (durAS.AB), both seen from the persons' A perspective (_A). In this measure the divergence is also reflected because the responses are always longer than the turns that are taken by person him- or herself due to and proportional with delay.

$$CI_A = durAS.BA\_A / durAS.AB\_A \tag{3.4}$$

Even though these measures seem interesting in the first place they cannot be calculated for natural conversations because they do not take into account turns being handed over by means of interruptions. The situation is indeed even worse. As we have seen from the review so far, the higher the delay gets the greater the number of turns that are ended by interruptions. The authors simply neglect this by using simulated conversations not allowing for double talk.

3.5 Conclusion

Concluding, it can be said that all measures revealing a divergence of the different perceived *conversational realities* are beneficial to better understand the destructive impact delay may have on conversations. It is suggested to include all possible state walks when analysing the divergence of the two interacting parties' realities. This approach will be developed further in Chapter 7.

Furthermore, the assessment of the amount of double talk seems to be interesting considering the changing source for double talk, as described by different studies. For this reason, it will be examined if a similar u-shaped relationship, as in Braun (2003), can be found for audio-only connections.

Completely absent from the literature that addresses mediated-conversations under transmission delay is the assessment of visual interaction (for the case of audio-video connections) through, for instance, analysis of body posture or face movements. This aspect will also be explored in Chapter 7 and in the following empirical Part II.

Chapter 4
Interaction Partners

This chapter will be discuss how much a communication system can influence the perception of the attributes of the interaction Partners. As there is limited research in the domain of audio- and videotelephony and -conferencing in this regard, this chapter will investigate further areas, and also include other communication systems. Based on respective suggestions, conclusions will be drawn about the possible effects in audio- and videotelephony and -conferencing.[1]

4.1 Problems of Computer-Mediated Communication

An article by Cramton (2001) summarises five common problems for the usage of computer mediated communication systems in *virtual group work* that may serve as an explanation as to why certain impressions of the interlocutors at the far-end arise. She based her analysis on a study where people were accomplishing a project without face-to-face interaction and only communicating via electronic media such as email, chat, internet-based voting tools and telephone.

Three of Crampton's five extracted problems are relevant to telecommunication and conferencing services. The first relevant problem was a **lack of communication** and **retention of contextual information**. Group members in Cramton's study, for example often forgot and failed to talk about differences in constraints such as deadlines or evaluation criteria which caused misunderstandings. Since telecommunication partners are distributed geographically and calls usually focus on the proposed topics, specific circumstances of each location are more likely to be neglected.

The second type of problem Cramton observed was related to differences in the **salience of information**. Nonverbal communication such as facial expression and body language can change the meaning of exchanged information. However, nonverbal behaviours are often only partially delivered when communicating over a telecommunication or conferencing service. This is particularly the case for audio-

[1] Parts of this chapter have been published before in Schoenenberg et al (2014b)

only services. Some nonverbal information can be delivered via the tone of voice (nonverbal and vocal), but then a new risk for misunderstandings is formed by technical impairments that can distort these cues.

Cramton also describes the problem of **misinterpreting the meaning of silence** as a major pitfall of computer mediated communication. Misinterpreting the meaning of silence is considered to be a key problem for examining the effect of transmission delay. A moment of silence in a face-to-face conversation can have several causes, most of which reside in the talking behavior or circumstances of the conversation partner. In the particular case of conversations, communication trouble arises due to unexpected long silences (caused by transmission delay) which disrupts the turn-taking system. The trouble's source usually can't be identified by the interactants (Ruhleder and Jordan, 1999, 2001). Even though technical issues could explain the troubles, for the conversation partner who attempts to identify the source of the other's silence, it is more likely that it is due to the other person expressing hesitations, doubts, uncertainty or similar. Humans are inherently socially oriented and tend to attribute dispositional rather than situational, a phenomenon widely known in the field of social psychology under the term **fundamental attribution error** (Ross, 1977) or **correspondence bias** (Gilbert and Malone, 1995). In a mediated communication situation this implies that a person will more likely explain, for example, an exhausting conversation with the personality or current mood of the interlocutors (dispositional) than to consider current technical issues of the communication system (situational).

Let us imagine a telephone conversation at a point where a person A is waiting for a response concluding that the other person B intends to deliver some message by not responding directly. Person A may start to worry what is wrong, may rephrase what he or she just said, or may simply move on with a new topic to stop the awkward silence. At the same time, person B does not experience any trouble or a very long silence and replies straight away. Person A may, however, have continued talking already, which appears to person B as if person A is not very competent or is uninterested (Ruhleder and Jordan, 2001). Topic crossovers or speech collisions are likely consequences (Ruhleder and Jordan, 2001).

For the case of videotelephony a similar example can be found. Kendon (1967) explained two main functions of vision in interaction: tracking the other's attention (eliciting feedback); and the control of turn-taking. He pointed out that people usually look up at the end of a sentence to see if the other is listening, understanding and fine with what was just said, in this case the other will express verbal or visual feedback. Looking up can also hand over the turn. With respect to delay, the looking behaviour is delayed which gives the speaking person the impression that the other is not listening because he or she is not giving feedback in an appropriate time frame, or that he or she does not want to hold the floor. To the listening person, the speaker may seem to be impatient and not leaving the floor. One should be mindful, however, that eye contact is not possible in most videotelephony applications which is why the functions of looking are somewhat disturbed anyway.

From these small examples it can be concluded that the technical delay will not necessarily affect the perception of the connection because the interaction structure

changes (the turn-taking system gets out of balance) which may likely be interpreted as an intended communication behaviour of the person(s) at the far end.

4.2 Findings from Paralinguistics

Findings from the linguistic domain support the assumption that time-related aspects of speech can alter the perceptions of the speaker. Miller et al (1976), for instance, examined conversations from field research and found that fast speakers were perceived as more persuasive. However, in their study other paralinguistic aspects that may have influenced the results were not controlled for. Apple et al (1979) supported this through their findings that slower speech was perceived as less persuasive and more passive. Nevertheless, the perception was also dependant on the topic. People who responded very quickly to a question on money were perceived as less persuasive than people who responded at normal speed. Lay and Burron (1968) investigated the perception of hesitation in speech. They compared how desirable hesitant speech stimuli were in comparison to non-hesitant speech stimuli and found that the hesitant stimuli were less desirable. They also checked whether the manipulation of hesitant speech could be used to artificially generate non-hesitant speech. For the non-hesitant speech stimuli, "ahs", repetitions and longer silences were excluded from recordings of hesitant speech. These manipulated recordings were indeed perceived as less hesitant and more fluent. A perceived hesitant reaction of the other person can further signal to the requesting person that his or her statement was not desired (Pomerantz, 1984). As a consequence, the requesting person may weaken their statement which facilitates a vague communication situation. Wollermann and Lasarcyk (2007) carried out a more detailed analysis of paralinguistic factors that may cause the perception of uncertainty. They asked participants to rate the perceived certainty for answer-response stimuli with different levels of response time (delay), with or without fillers (the sound "hmmm" was used) and with rising or falling F0 contour at the end of an utterance. A rise of intonation itself already led to a higher degree of uncertainty. Increasing the response time in addition did not lead to a significant increase of uncertainty. However, when combining longer response times with fillers and rising F0, a very strong effect on the perception of uncertainty was observed. The effect of delay-only was unfortunately not tested, but as explained earlier, fillers or repetitions are likely to be used for bridging long awkward silence in conversations under transmission delay.

4.3 Conclusion

To date, there does not seem to be any explicit study identifying a dependence of pure delay and perceived attributes (i.e. the current state or the personality) of the interaction partners in a mediated conversation setting. Different results strongly

support the assumption that delayed responses are likely to lead to a different perception of interlocutors' attributes because the reason for silence are difficult to identify, particularly when no or only reduced visual communication behaviours are accessible. In the empirical part of this work (Part II) it will be examined whether such misattributions of long silences caused by the technical impairment delay can indeed be observed.

Chapter 5
Hypotheses

Based on the discussion of the previous chapters, the following hypotheses are formulated to be tested in the empirical investigations:

Hypothesis Ia
Different levels of pure one-way transmission delay lead to a decrease in Conversational Quality ratings.

Hypothesis Ib
The decrease of Conversational Quality ratings is greater if the interlocutors are willing to interact quickly (RNT task type[1]) than if there is no need to accomplish the task in fast interaction speed (SCT task type).

Hypothesis Ic
If people experience the impairment delay in an interaction that was supposed to be fast (RNT task type[1]) prior to slower interaction conversations (SCT task type) they will be more critical about their Conversation Quality ratings for the slow interaction conversations than compared to those with no prior sensitisation.

Hypothesis Id
People lose sensitivity for the ratings of Conversational Quality the bigger the group and thus rate less critical.[2]

Hypothesis IIa
Different levels of pure one-way transmission delay lead to a change of the interaction structure of the conversations. This change can be interpreted as a less comfortable interaction.

---

[1] RNT: Random Number verification task under Time pressure, see Chapter 6 for details

[2] The reason for this is less available cognitive resources. A three-party call is usually more complex and demanding than a two-party call (Skowronek et al, 2011).

Hypothesis IIb
Different levels of pure one-way transmission delay lead to the perception of decreased conversational fluency.

Hypothesis III
Different levels of pure one-way transmission delay lead to a change in the perception of the interaction partners' attributes (current state and personality), especially if the perceiving person and the interaction partner are unfamiliar to each other. For interaction partners who are familiar with each other this does not need to hold.

Part II
Empirical Study

Chapter 6
Methodological Aspects of Present Study

6.1 Overview

To answer the research questions, a total of eight experiments, seven addressing audio-only and one the audio-video mode have been conducted. They will be outlined in the following chapters. Table 6.1 gives an overview of all experiments including the most important details of the experimental designs. Besides the mode, "A." audio-only or "V." audio-video, experiments differed in the number of interlocutors per group. In accordance with this, different conversation scenarios were used, according to the number of participants per group. As technical system characteristics, the delay-value and codec was varied across experiments. In all empirical investigations the Mean Opinion Score (MOS) was assessed. In some experiments, additional scales were considered (for details on the assessed scales see Chapter 7.2).

Experiments comprised of either two-party or three-party groups as indicated by ".2" or ".3" respectively at the end of each experiment name (see Table 6.1). In this fashion, two main Narrowband (NB) experiments A.NB.2b and A.NB.3, two Wideband (WB) experiments A.WB.2 and A.WB.3 and two Fullband (FB) experiments A.FB.2 and A.FB.3 were conducted. The order in which experiments were conducted is indicated in the column "Order Number" in Table 6.1. The very first experiment A.NB.2a was replicated in A.NB.2b, with some details changed in the design. The two Fullband (FB) experiments were conducted prior to the WB experiments. In the FB experiments, the research focus was on assessing potential misattribution of the impairment transmission delay to attributes of the conversation partners. Hence, the design of those two experiments differed notably, especially regarding the delay range and the scenario types used. Finally, the two WB experiments were conducted adopting new assessment scales. The audio-video experiment, V.WB.2, was a first approach to extending the knowledge acquired so far from the audio domain to the audio-video domain. Following this, more details on the designs and distinct aspects of each experiment will be explained.

Table 6.1: Overview of the empirical investigations; Exp.: Experiment, No.: Number; **Scenario Types**: SCT&3CT: (Short) Conversation Tests, RNV: Random Number Verification, RNT&3RNT: Random Number Verification Timed, CNG: Celebrity Name Guessing; **Bandwidth**: NB: Narrowband (300-3400 Hz), WB: Wideband (50-7000 Hz), FB: Fullband (20-20000 Hz); **Scales**: MOS: Mean Opinion Score, ATT: perceived inattentiveness, FLOW: perceived fluency, PERSO: perceived personality; Details are given in later parts of Chapter 6.

Mode	Group Size	Exp. Name	Exp. No.	Scenario Types	Delay Range [ms]	Codec	Scales
audio	2	A.NB.2a	1	2SCT, 2RNV, 2RNT	100-1625	NB, G.711	MOS
		A.NB.2b	2	2SCT, 2RNT	100-1600	NB, G.711	MOS
		A.WB.2	7	2SCT, 2RNT	75/135-1600	WB, G.722	MOS, ATT, FLOW
		A.FB.2	3	2SCT	1.5-1200	FB, no coding	MOS, PERSO
	3	A.NB.3	6	3SCT, 3RNT	100-1600	NB, G.711	MOS
		A.WB.3	8	3SCT, 3RNT	135-1600	WB, G.722	MOS, ATT, FLOW
		A.FB.3	4	3CT	1.5-1200 +asymmetry	FB, no coding	MOS, ATT
audio-video	2	V.WB.2	5	CNG	100-1200 +asynchrony	G.722 & H.264	MOS

6.2 Participants

The test participants of the eight experiments were of similar age and cultural background. Details on the age distribution of the **samples** can be found in Table 6.2. Participants were paid for their participation. For all experiments, knowledge of German at a native-speaker level was required to ensure fluent conversations without hesitations based on translation problems. Consequently, it can be assumed that participants have been living in Germany for a substantial part of their life and have acquired conversational norms and rules of the German language and culture.

Table 6.2: Information on Samples of the Study

Study	Age		N	# Groups	Group Gender			Familiarity
	Mean	SD			female	male	mixed	
A.NB.2a	30.7	8.3	48	24	8	8	8	50:50
A.NB.2b	29.8	7.1	44	22	12	10	-	unfamiliar
A.WB.2	27.8	6.8	40	20	10	10	-	50:50
A.FB.2	27.3	6.1	44	132	18	-	26	unfamiliar
A.NB.3	28.2	10.9	24	8	4	4	-	50:50
A.WB.3	28.6	10.3	42	14	6	8	-	mixed
A.FB.3	34.5	11.41	21	7	1	2	4	unfamiliar
V.WB.2	28.1	8.7	44	22	8	7	7	familiar

The number of participants and the corresponding number of groups per experiment are summarised in Table 6.2. The gender allocation was kept constant wherever possible, ensuring an equal number of female-only, male-only or mixed gender groups. In four experiments there were no mixed gender groups to avoid inter-gender effects on the conversations. In Schoenenberg and Raake (2011) it was shown that the gender of the interaction partner may influence quality ratings in two-party groups. In experiment A.FB.2 and A.FB.3 it was not possible to control the group gender. As a consequence of the particular procedure of experiment A.FB.2, with each conversation having one out of three female fixed communication partners, there were no male-only groups. More details on experiment A.FB.2 will be described in section 6.6.

For experiment A.FB.3 participants with a high level of experience in multi-party teleconferencing were recruited. On average they had conducted 18.02 (SD = 14.23) conference calls in the preceding year in a business or private context. Unfortunately, as revealed by informal interviews after the experiments, their experience was generally bad, including a big variety of distortions, and their reference regarding conferencing quality was accordingly low. Since only delays were inserted in the connections during the experiment, hardly any decrease of quality was noticed by the participants (for details see Chapter 8). Consequently, the criterion for participants to be frequent users of multi-party conferencing systems was considered to be inappropriate for the remaining three-party experiments A.NB.3 and A.WB.3 because the tested conditions are considerably better than what people typically experience in real-life conference calls. A thoroughly-conducted warm-up phase ensured that participants were familiar with the technology in those three-party experiments.

In Table 6.2 details on the **familiarity** of participants of each group can be found. As a general approach, half of the groups were comprised of people who knew each other well, being good friends, relatives or in a relationship. The other half of the groups were people who did not know each other prior to the experiment. This distinction was made because the familiarity of interlocutors is considered to be an important factor when judging the degree of normal reaction speed of the conversation partner.

Communication partners were not familiar with each other in experiment A.NB.2b because here the influence of familiarity was not in focus. By choosing groups of unfamiliar interlocutors a common ground having no experience in interacting with each other was created for all of the respective participants.

In experiment A.FB.2 and A.FB.3, the goal was to assess the impact of transmission delay on the perception of the interaction partners' attributes, such as the perceived personality or the perceived current state of the other, when he or she was unknown to the partner. Therefore, groups were built by participants who did not know each other.

Taking the influence of interlocutors' familiarity on step further, the three-party groups in A.WB.3 was comprised of two persons who were very familiar with each other and a third person who was completely unfamiliar to the other two. This way, the impact of an asymmetrical familiarity constellation on ratings of quality, personality or current state attributes and on the conversational structure could be examined.

In the audio-video experiment, V.WB.2, familiarity was considered to be a key factor for proper usage of the video channel as a communication medium. Informal pre-tests showed that people who did not know each other hardly used the video to communicate. The higher demands to the management of the persons own impression (Brander and Mark, 2001) may have been the reason for avoiding the video channel. For this reason all groups were built of familiar participants in this experiment.

6.3 Tasks

To choose an appropriate task for participants when examining the perceived quality of a telephone or conferencing call is not straightforward. The degradation transmission delay investigated here does not become apparent in a pure listening setting which is why tasks involving conversations are needed. Moreover, a high degree of control is needed regarding the technical setup, for instance, delay times must be controllable on a millisecond level, which is why the following empirical work will be based on laboratory setting. This in turn implies a certain difficulty to enable natural conversations, as a result of the laboratory environment. This difficulty has been addressed with the choice of elaborated tasks, typically referred to as *conversation scenarios*.

6.3.1 Audio Tasks

The Short Conversation Tests (SCTs) were developed for quality testing of **two-party telephone calls** (ITU-T Rec. P.805, 2007b; Möller, 2000). They aim to be able to test telephone connections in a controlled laboratory environment while facilitating natural conversations at the same time. People are guided by rough structure documents, mainly showing symbols, providing indications for participants to talk about everyday topics without much effort. The following scenarios were used: travel agent (warm-up), information on flights, theatre box office, pizza service and doctor's appointment. Each SCT call lasts for about 2-3 min. Due to the predefined structure, conversations are always similar and comparable but still leave the freedom to express things individually (no given sentences).

In the context of transmission delay, Kitawaki and Itoh (1991) have shown that the nature of the experimental task plays an important role for the reflection of delay in the quality judgments. They found that the shorter the exchanged intervals for information were the more the perceived quality was affected by delay. For this reason, the RNV task, a task type that was later recommended by the ITU-T as a delay sensitive test task (ITU-T Rec. P.805, 2007b; Egger et al, 2010), was used in this context.

In the first experiment conducted in the series A.NB.2a, the SCTs and RNVs were considered as the appropriate task types. However, on the basis of the results reported by Egger et al (2010), the RNV task seemed to have less of an effect in terms of quality than reported by Kitawaki and Itoh (1991). Different aspects may have caused these conflicting results. In the earlier study, people were trained prior to the experiment which is assumed to be why they paid more attention to the degradation transmission delay. Furthermore, the 20 year laps between the two studies, may have caused the sensitivity of participants to change. Nowadays, people are used to latencies, for example, from using mobile phones or IP-telephony services, and as a result they may be less critical of them.

When looking at the conduction of the RNVs more closely, it appeared that they can also be accomplished at a rather low speed. As long as the participants are not motivated or under pressure to finish the task fast the quality perception does not necessarily need to be affected, when the speed is not as fast as it could be. To arouse the motivation for fast interaction, a timed version of the RNV task, the Random Number Verification Timed (RNTs), was invented. The RNTs are similar to the RNVs, with the additional instruction to participants that a prize would be given to the best team in the RNTs. The prize, a voucher for every team member worth as much as one hour paid to each participant, was given to the group with the overall fastest and most correct finalisation of the RNTs.

Similar to the SCTs, scenarios for **three-party teleconferences** are available (ITU-T P.Sup26 2012a; Raake and Schlegel, 2010; Raake et al, 2010). They are referred to as three-party Conversation Tests (3CTs), and are built on business topics, such as agreeing on the venue for a conference or finding a replacing person due to a colleagues illness. Having a business context rather than a private one is consid-

ered appropriate for three-party conversations because audio-only conference calls are usually carried out in business life.

In experiment A.FB.3, which was the first three-party experiment conducted, participants accomplished the 3CTs in their original version (Raake and Schlegel, 2010). Depending on the inserted delay, scenarios took up to 9 min and required a lot of concentration by the participants. Furthermore, the results of this experiment suggested that the 3CTs in their original version were hardly sensitive for quality measurement in the scope of delay. As a consequence, a shortened version with a reduced number of discussion points and reduced detailed information was developed and tested by Skowronek et al (2013). In this way, the complexity of the task was reduced for participants which had the advantage to leave more cognitive capacity for evaluating the quality of the connection. The shortened 3CT scenarios were revised another time by Skowronek et al. (2014). The final version is referred to as three-party Short Conversation Test (3SCT) and used in this context. Conversations based on these tasks last only 3-5 min, which makes it possible to assess multiple different task types in the same experimental session.

In addition, an extended version of the two-party RNTs, the three-party Random Number Verification Timed (3RNT; Schoenenberg and Schmieder, 2014), was included for the three-party experiments A.NB.3 and A.WB.3. The size of the number blocks stayed the same for 3RNTs but lines were coloured in red, blue and green. The colour indicated to the participant whether he or she had to read the numbers (red), respond first if he or she had the same number (green) or respond second (blue). For every line the colouring changed for each participant. If a person responded on the accordance of the number with "no" it had to be noted by striking the number out once. If two participants responded with "no" the number needed to be struck out twice. For "yes" answers, nothing needed to be done. All interlocutors were asked to note the answers at any time during the call. For the 3RNTs a prize was offered to the fastest and most correct team, similar to the two-party case.

All verbal and written introduction to the experiment and task description were given in German.

As the results of experiment A.NB.2a showed that the **order of the task types** may be important for the ratings. Once the motivation to interact fast was enhanced, participants did not differentiate between RNV and RNT tasks. This was why the RNV task has been dropped for all experiments following A.NB.2a.

The results of experiment A.NB.2a furthermore suggested that a randomised order of the tasks may lead to a biased and higher sensitivity for delay in the SCT scenarios. For this reason, experiment A.NB.2b investigated a block design and explored the order effect, having either only SCTs or RNTs in the first block. A small effect of the block order could be found, which was why it was decided to stay with a block design for the subsequent experiments always having the SCT block prior to RNTs. This effect was not confirmed in the according three-party experiment A.NB.3. Nevertheless, it was decided to stay with a block design for A.WB.2 and A.WB.3 and to always have the SCT block prior to RNTs block to avoid possible effects.

6.3.2 Video Task

An appropriate task for the audio-video experiment, V.WB.2, was needed to facilitate interactive conversations that use the audio and particularly also the video channel to communicate. Different tasks have been described in the literature. Berndtsson et al (2012) found a slight trend that quality ratings were more critical for free conversations than for a quiz game. This result could be due to more attentional resources being available (Kahneman, 1973) for evaluating the quality in the case of free conversations. Both of these task types are not specifically designed to facilitate visual communication, and as a result, participants may not use the video channel during the calls.

The Block-Building task, as recommended in the ITU-T Rec. P.920 2000, was found to not be sensitive for testing interactive audio-visual quality either (Bräuer et al, 2008).

Belmudez and Möller (2013) tried to find a conversational scenario format where the audio and video channel are used equally to test aspects of audio-visual integration in terms of quality. In a pre-study people had reported to use the audio and video channel about as often when accomplishing the Audiovisual Short Conversation Tests. This scenario type is an extended version of the SCTs, with an additional task part of holding objects into the camera for showing the other person some aspect related to the object. For experiment V.WB.2, however, it was not the goal that participants use each channel equally often but that they use both channels to actually communicate. In particular, it was the aim that interlocutors interact visually by means of smiles and other facial expressions, gestures, head nods, body posture and similar. For this reason, the main requirement for the audio-visual task was to facilitate visual communication.

In ITU-T Rec. P.920, a Name Guessing task is described which was later adopted by Hayashi et al (2007). In the this version of the question-answer game, one participant gets a description word of a brand or a well-known person and the corresponding name. The other participant has to guess the name through asking predefined questions.

For experiment V.WB.2 a different name guessing task, also known under the name “Who am I?”, was used. The Celebrity Name Guessing Task (CNG) has a strong social component because the respective other person is associated with the celebrity person that he or she has to guess. Furthermore, the person is eager to find out who he or she is to win the game. These aspects bring fun and a natural interaction including visual interaction cues. Due to the question-answer structure the alignment of turn-taking is necessary.

To facilitate the visual interaction even more, we invited only pairs of participants who were very familiar with each other. In this way, we aimed to keep the self-monitoring, in regard to what impression the other gets of the self in the social interaction (Brander and Mark, 2001; Goffman, 1961) low, to ensure that people had enough capacity for the quality evaluation.

In detail the CNG task works in the following way: Each participant receives a note with initials of a celebrity person that he or she has to guess. This note is

clipped to the shirt of the person so that they and their interlocutor are reminded of it. The respective other knows the full name of the unknown celebrity person due to a hidden note with the full name. One participant begins by asking questions that can only be answered with “yes” or “no”, for example, “Am I female?” or “Am I a movie star?”. As long as the response of the other one is “yes” the participant is allowed to keep asking. If the answer is “no” the roles of asking and answering person are reversed. In this way, the roles are switched several times until one person correctly guesses the name. If one participant wins a game he or she gets a point. After the entire session (nine games) the participant who won most games wins the entire session and with this an additional prize (voucher).

Regarding the audio-based conversation structure, this task type is expected to be similar to the RNTs due to its answer-response pattern. In terms of visual interaction the CNG task can be considered to facilitate natural visual interaction. As in the audio-only experiments, all contact, introduction to the experiment, task description and conduction was performed in German.

6.4 Impairment Levels

In general the **audio transmission delay** levels tested ranged from the smallest possible delay of the system to the maximum of 1600 ms one-way. The following

Table 6.3: Overview of the investigated delay levels in the different experiments (one-way delay time in milliseconds [ms])

Mode	**Group Size**	**Experiment Name**	**Delay Levels [ms]**
audio	2	A.NB.2a	100,225,425,825,1625
		A.NB.2b	100,200,400,800,1600
		A.WB.2	75,135,400,800,1200,1600
		A.FB.2	0,400,1200
	3	A.NB.3	100,400,800,1200,1600
		A.WB.3	135,400,800,1200,1600
		A.FB.3	0,400,1200 + asymmetry
audio-video	2	V.WB.2	0(,200),400,800 (+170) + a/v asynchrony

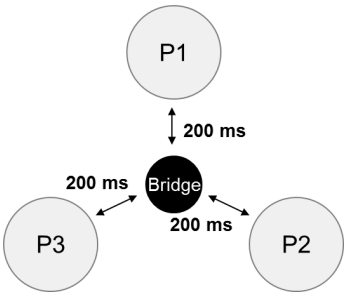


Fig. 6.1: Symmetrical Condition; P1, P2 and P3: 400 ms

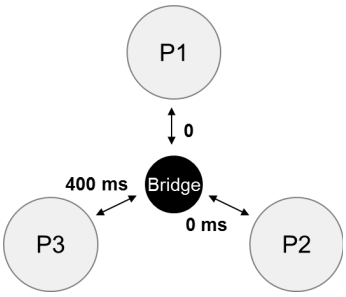


Fig. 6.2: Asymmetrical Condition; P1 and P2: 0ms | 400 ms, P3 400X ms

set of logarithmically distributed levels of transmission delay in milliseconds [ms] served as the basis for selecting the levels of test in the individual experiments: 100, 200, 400, 800, (1200,) and 1600. All delay levels for all experiments are summarised in Table 6.3.

Each delay level was combined with each task type once in a randomised order. This means, all experiments had a complete design. Regarding the delay level distribution, some small exceptions were made for some experiments as indicated in Table 6.3. In experiment A.WB.2 the conditions 100 ms and 200 ms were replaced by the levels of 75 ms and 135 ms. The condition of 135 ms was the minimal possible delay in experiment A.WB.3 replacing the two conditions 100 ms and 200 ms. Each task lasted slightly longer in the three-party experiments than in the two-party experiments. The condition of 1200 ms was considered an important inclusion. To keep the session duration below one hour the 200 ms conditions in the experiments A.NB.3 and A.WB.3 were omitted. In experiment A.NB.2a there was another deviation from the basic distribution of conditions. All conditions higher than 100 ms had additional 25 ms of delay, for example, the condition of 800 ms was actually 825 ms.

The design of the two FB experiments allowed for only three delay levels: no delay (1.5 ms), 400 ms and 1200 ms. The initial system delay, 1.5 ms, was very low due to direct connections and no coding, and will be neglected in the following.

Asymmetrical and symmetrical conditions were furthermore tested in the three-party experiment A.FB.3. The setup for the case of 400 ms and three participants is exemplary illustrated in Figure 6.1 showing the symmetrical and Figure 6.2 showing the asymmetrical case. As can be seen in these figures, in the symmetrical conditions each participant experienced the same transmission delay to each of the two other interlocutors. In the asymmetrical conditions, two cases need to be distinguished when looking from the viewpoint of one participant in three-party groups. First the indirect asymmetrical case, where the experience for a participant was similar to

Table 6.4: Delay conditions of experiment A.FB.3. T_{send}: delay to conferencing bridge, $T_{received}$: delay from conferencing bridge. |: The delay of the connections to one or the other interlocutor differed. X: The delay was similar to both interlocutors, however, the other two spoke with each other without delay.

Symmetry	T_{send}	$T_{received}$	Condition Name
symmetrical	0	0	0
	200	200	400
	600	600	1200
asymmetrical	0	400	0|400
	400	0	400X
	0	1200	0|200
	1200	0	1200X

the symmetrical condition, but their conversation partners were able to communicate without any delay between each other (condition named: 400X and 1200X). In contrast to this, for two participants there was a directly perceivable asymmetric perspective, the connection to one interlocutor was delayed but there was no delay to the third person (condition named: 0|400 and 0|1200). Since every participant was meant to run through every condition, the direct asymmetrical case was present twice for every participant, once for every interlocutor and the indirect asymmetrical case was carries out once.

In experiment V.WB.2, **asynchronous audio-video delays** were investigated besides synchronous conditions. The technical conditions for the two-party videotelephony experiment were chosen based on knowledge of detectability (45 ms to -125 ms) and acceptability (90 ms to -185 ms) thresholds for audio-video asynchrony of ITU-R Rec.BT.1359-1 (1998; positive values for thresholds indicate that the sound is ahead of the video). However, these values are rather conservative, some findings, for instance by Dixon and Spitz (1980) or Miner and Caudell (1998) suggest less strict boundaries for the detection of asynchrony.

In case of asynchronous delay, it is well-known that a delayed video channel is less acceptable than a delayed audio signal, for instance, from the field of broadcast television (International Telecommunication Union, ITU-R, 1998). For this reason, video-only delay was examined at a level of 200 ms and 400 ms whereas the audio-only delay was 400 ms and 800 ms. The asynchronous conditions were selected so that all condition were clearly worse than the upper acceptability threshold. They were chosen to be obvious because the focus of attention was not on asynchrony detection but on the overall CQ. Only the impact of the asynchrony on the perceived quality was assessed. Moreover, participants had to accomplish their conversation tasks, costing a higher amount of attentional resources.

Table 6.5 summarizes all conditions that were tested. The initial system delay of 170 ms one-way for both the audio and the video channel is not considered in these values and should be added.

Table 6.5: Delay conditions of the video experiment V.WB.2

Synchrony	**Delay [ms]**		**Condition**
	audio	video	**Name**
synchronous	0	0	0
	400	400	400
	800	800	800
asynchronous	0	200	200*v*
	0	400	400*v*
	400	0	400*a*
	800	0	800*a*

6.5 Apparatus

All setups were implemented as *one-per-site*, according to ITU-T Rec. P.1301 (2012c), this means that there was only one person at each end. This may have implications for the results because no face-to-face interaction was possible for any of the participants, but it assures similar communication conditions for all participants for each call.

Three setups can be distinguished. For the first **telephone setup** used for most experiments, Snom 870 IP-telephones served as terminal devices. The software Asterisk connected the two or three phones and enabled a typical telephone calling procedure. In the two-party experiments, one participant called the other and this role changed for every call. In the three-party experiments, every participant dialled into the conference, as it is usually the case in real-life settings. The role of the caller was nevertheless allocated to ensure a fluent beginning of the call. The participant being the caller was changed between the three members of each experimental group. For two- and three-party experiments, dial plans were created in Asterisk. The corresponding codec was set in the Asterisk configurations for each condition (for the codecs used see Table 6.1; for G.711, A-law was chosen in all case). All calls were recorded via the Asterisk software. The module NetEm (Network Emulator) of the Linux Debian packet scheduler allowed defined delays to be added to every packet going out.

The loudness was set to achieve an Overall Loudness Rating OLR according to ITU-T Rec. P.79 (2007a) of approximately 10dB. The Overall Loudness Rating is typically calculated as the addition of the Send Loudness Rating SLR (2 dB) and Receive Loudness Rating RLR (8 dB).

In experiment A.NB.2a, being the very first experiment conducted, the setup differed. Even though Snom phones were used, all aspects handled by Asterisk in later experiments were accomplished by a HeadAcoustics switch configured specifically for this experiment. Loudness attenuation levels were measured to be slightly higher with a Send Loudness Rating SLR of 7 dB and a Receive Loudness Rating RLR of 8.8 dB. The recoding was done by placing pressure zone microphones (Beyerdynamic, MPC 65) on the table in front of each participant in each room.

In the first experiment using Asterisk, A.NB.2b, it was not fully ensured that an acoustic echo would be avoided for particularly loud speech. For this reason a speech babble noise source with a sound pressure level of 68.2 dB was placed inside each room to mask potential echo. In later experiments, potential echo was fully eliminated by further adjusting the echo canceler.

The second headset setup, the **headset setup**, built for experiment A.FB.2 and A.FB.3, was characterized by a direct connection of the two or three interlocutors without any coding (FB: fullband bandwidth) via the software Pure Data (PD) (Puckette, 2007). Within PD the delays were added to the corresponding lines. Closed Beyerdynamics DT 290 headsets were used to communicate. The input of the microphones was captured and recorded directly. Regarding loudness, an attenuation of 25 dB SPL was measured in experiment A.FB.2 and 23 dB SPL in experiment A.FB.3, both times determined by a test signal played at the far-end of a dummy heads' ears, recorded from a dummy heads' mouth.

For the **audio-video setup** of experiment V.WB.2, audio and video information had to be delivered and each channels' delay had to be controlled. Therefore, the soft-phone client Linephone version 3.5.2 served as the interface and call client. The system behind the client was implemented in LINUX Free BST, and Asterisk was used to connect the two sides and record the audio of all calls. The videos of each room were recorded in a synchronised manner using the software ManyCam version 3.0, and additional web-cams were placed as close as possible to the built-in web-cams of the laptops. Due to technical problems with the recording system, lower and higher recording resolutions were used. For both resolutions, the video frame rate was set to 25 frames per second. The recording software combines the two images of the two cameras used during recording. The lower resolution produced an image size of 160x120 pixels for one person and therefore a resolution of 320x120 pixels for the combined pictures for both participants. The higher resolution produced a picture of 320x240 pixels for one person and 640x240 pixels for both persons. Approximately three quarters of the sessions were recorded with the lower resolution and one quarter with the higher resolution settings.

The Linux Free BST firewall was configured so the audio and video packets always arrived in particular ports. This configuration made it possible to add delay to

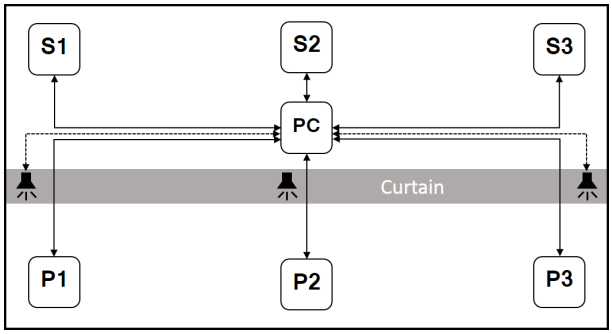


Fig. 6.3: Configuration of experiment A.FB.2. S1, S2, S3 were the positions of the fixed speakers. Positions P1, P2 and P3 belong to the participants. In order to avoid visual interaction between interlocutors and participants, there was a curtain in between. At every position there was a headset and a headset amplifier. In the center the PC (including the connection simulation tool), an external soundcard, and an amplifier were positioned.

the audio or video channel individually by slightly modifying the firewall properties. In each room, participants were equipped with a laptop, mouse and closed headsets (Beyerdynamics DT 290) as devices.

For all experiments (expect in A.FB.2), participants were seated in separate acoustically treated **laboratory rooms** (according to ITU-T Rec. P.800,1996) at Deutsche Telekom Laboratories, TU Berlin. Prior to the use of every setup, it was verified that potential echo was not perceivable for any delay condition.

The experiment A.FB.2 took place in a large anechoic chamber (area = 120 m^2 lower frequency limit = 63 Hz) at the Technical University of Berlin, see Figure 6.3 for setup. Previous work by Skowronek et al (2011) had shown that the acoustics of the room fitted the needs of a multi-party conference call with up to six interlocutors at the same time. A speech-babble noise at an approximate noise level of 58.2 dB SPL (measured at each seat) was played during all calls (not in the time between calls, to keep the stress level of participants as low as possible) to completely avoid direct speech transmissions between participants and to mask any speech cross-talk. Three loudspeakers playing the speech-babble noise were placed in the middle of the room. Beyond this, the closed headsets prevented cross-talk. The room was divided into two halves by a big curtain to avoid visual contact of the three fixed speakers and the participants (for details on the reason for the fixed speakers see sections 6.6 and 7.2).

6.6 Procedure

The procedure of all experiments was generally the same. After arrival, participants were introduced to each other, in case they did not know each other. They were then informed about the aim of the experiment, namely of acquiring insights to the quality perception of telephony or multimedia conferencing. Participants were not informed about what kind of impairments were to occur to avoid drawing the participants attention, which could have created an unnatural over-sensitisation.

The conversation task types were introduced to the participants, including the competitive part. After clarifying all questions, participants signed a disclosure to agree on being recoded.

Participants were then seated in separate rooms (according to ITU-T Rec. P.800, 1996) for the entire session. Each task type was conducted once using the system in the smallest possible delay condition. Participants were aware that the conditions of the warm-up phase were the best that they would experience - they were consequently anchored regarding the best quality. An anchoring to the worst condition was not possible because this was thought to have affected the sensitivity to the impairment and it may have implied that a high delay equals a low quality.

After the warm-up, the formal experiment began. Using the hands-free mode was not allowed throughout the experiments. In the case of a two-party experiment, the calling person was reminded of their role with a particular sign on the document. The callers role was altered for every call. In the case of a three-party experiment, all participants dialled into the conference. One person was assigned to be the caller and was therefore supposed to begin with the task. The test supervisor was able to talk to the participants and indicated, in the breaks between tasks, when participants could continue with the next call.

The RNTs were clearly tagged so that participants were always aware of whether the current task counted for the competition. Furthermore it was emphasised that shortcuts were not allowed, for instance, reading out five numbers in a row. This rule forced people to wait for the confirmation of a number before reading out the next number.

In the audio-video experiment, V.WB.2, each person saw the head and shoulder video of his or her friend on the screen, and a small version of their own video on the left corner of the screen.

After each call, the corresponding questions selected for a particular experiment needed to be answered using the respective scale(s). Overall, one session took approximately one hour up to 75 min, except for experiment A.FB.3 which took 1.5h per session.

As previously mentioned, in some aspects the procedure of experiment A.FB.2 differed from the other experiments. The introduction to participants and the conversation scenarios were similar. However, the experiment took place in a large anechoic chamber. For every session, three participants were invited and seated in one half of the room, as far away as possible, and facing away from each other. They were informed that they would speak to three different speakers seated in the other half of the room not visible due to a curtain placed along the middle line of

the room. Each participant was talking to each stationary interlocutor, referred to as speakers, once throughout a session. For one participant, one conversation was undistorted, which implies no delay, one was delayed with 400 ms delay and one with 1200 ms delay (randomised order).

After each call, participants filled out the questionnaire with a question about the quality of the connection and the perceived personality of the speaker. In total each participant performed three calls. The speakers were three female research assistants who did not know the current condition. To ensure that the presuppositions were similar for all conversations, it was necessary to inform the speakers in advance about the possible impairment delay. The speakers were asked to act in a similar manner for all conversations and as naturally as possible.

In the described design it was most important that participants and fixed speakers were unfamiliar to each other prior to the call to ensure that the ratings on the speakers' personality were mostly dependant on the inserted delay and not on prior experience in interacting with the speaker. The ratings were also expected to depend on the reflection of the initial personality of the speaker in her interaction behaviour and her voice. For this reason, three speakers were selected. Their gender was however kept constant to avoid explicit gender induced differences.

Chapter 7
Dependent Variables

7.1 Pre-Questionnaire

Prior to the introduction of the experiment participants were asked to fill out a short questionnaire assessing their age, cultural and language background and the familiarity to their co-participants (see Appendix III). These details can be found in in section 6.2. [1]

7.2 Ratings

7.2.1 Conversation Quality

In all experiments, after each conversation, participants were asked to rate the **Conversational Quality (CQ)** of the connection on a five point ACR scale according to ITU-T Rec. P.800 1996, to calculate the Mean Opinion Score MOS (see Appendix III).

7.2.2 Perceived Attributes of the Interaction Partners

After having conducted A.NB.2a-b and A.FB.3, it seemed as if the MOS ratings did not reflect the effects that participants experienced when delay was present, es-

[1] Furthermore, the current mood was assessed on a 7-point ARC scale ("How are you currently feeling?"; German: "Wie geht es dir gerade?") ranging from 3 to -3 with the end points "ideal" and "very bad" (German: "ideal" and "sehr schlecht"). The mood question was again asked after the experiment to examine whether mood changes may be related to the given Conversational Quality judgments. This aspect will, however, not be reported further here because it only constitutes a side effect. Some first insights can be found in Schoenenberg and Raake (2011).

pecially in rather natural conversations. Even though changes in the conversational structure could be measured, it was the aim to find measures which reflected the impact of delay–only in a better way.

For this reason in experiment A.FB.2 and A.FB.3, it was examined if people perceived a change in personality attributes of the person at the far-end (Table 6.1, PERSO). In A.FB.2, we decided to investigate how three fixed speakers were perceived in terms of personality traits dependant on the delay added to the connection. Therefore, after each conversation with one of three speakers, participants rated the quality and their impression of the speaker's personality on scales taken from a German translation of the International Personality Item Pool Representation of the NEO PI-R (Goldberg, 1992; Goldberg et al, 2006). Each participant talked to and rated each speaker only once (see section 6.6). In particular, the six scales of the **perceived personality (PERSO)** trait extraversion (E1 friendliness, E2 gregariousness, E3 assertiveness, E4 activity level, E5 excitement seeking, E6 cheerfulness) and the six scales of the trait consciousness (C1 self-efficacy, C2 orderliness, C3 dutifulness, C4 achievement-striving, C5 self-discipline, C6 cautiousness) seemed most interesting in this context. Per scale, ten items were assessed (five positive, five negative items) which led to a total of 120 items per questionnaire, plus the quality question which was asked as the very first question (see Appendix III). In the following, scores will always be reported as the sum of all items per scale.

Personality traits are generally considered to relatively stable over the lifetime of a person. This is why it does not make sense to ask for perceived personality traits in dependence of delay if people talk to each other more than once, or if they know each other. To be able to test assigned attributes for people who have spoken to each other before and thus know each other a little bit, a scale related to the current state of the interlocutors was tested in A.FB.3. Here, people were asked to judge on the **perceived inattentiveness (ATT)** of the other two interlocutors within the conversation (see Appendix III). Each other person was rated separately regarding this aspect on a continuous scale with the end-points "very attentive" and "very inattentive" (in German: "sehr aufmerksam" and "sehr unaufmerksam"). However, the results of A.FB.3 suggested that a continuous scale is not the most sensitive choice. [2]

Based on the gathered knowledge, the questionnaire was further developed for experiment A.WB.2 and A.WB.3. In addition to rating quality, participants now rated how attentively the other/s contributed to the call "How attentively did the following person contribute to the conversation" (in German:"Wie aufmerksam hat sich der folgende Gesprächspartner an diesem Gespräch beteiligt?") on a seven-point absolute category rating scale with the end points "very attentive" and "very inattentive" (in German: "sehr aufmerksam" and "sehr unaufmerksam"). It should be noted that in both three-party experiments (A.WB.3 and A.FB.3), the rating order of the voices were kept constant to avoid confusions (see Appendix III).

[2] Besides the attentiveness, the likeability of the interlocutors was also assessed on a continuous scale with the end-points "very pleasant" and "very unpleasant" (in German: "sehr angenehm" and "sehr unangenehm". According results will not be considered here but can be found in Weiss and Schoenenberg (2014).

7.2.3 Perceived Fluency

In experiment A.WB.2 and A.WB.3 participants were also required to judge a statement addressing the **fluency (FLOW)** of the call : “This was a fluent conversation.” (in German: “Dieses Gespräch war flüssig.”) with the end points “applies fully” to “does not apply at all” (in German: “trifft voll zu” and “trifft überhaupt nicht zu”) on a seven point Likert-Scale.

Participants were required to place a total of three or four ticks, depending on how many interlocutors were in the conversation (one quality judgment, 1-2 interlocutors inattentiveness, fluency of conversation). The scales for rating the inattentiveness and fluency differed in their appearance (see Appendix III). For one scale, grading markers were represented by small boxes, and for the other circles were used. Furthermore the span of the markers differed on the paper sheet. This design was implemented to prohibit people simply copying their rating from one scale to the next.

7.3 Mediated Interaction Measures

7.3.1 Nonverbal and Vocal Conversation Metrics

For analysing the nonverbal vocal conversation structure, the recorded speech was first transformed into on-off patterns. A voice activity detection algorithm was applied for this purpose using a window size of 100 ms. As recommended by Brady (1968), pauses from the same speaker that were shorter than 200 ms were filled. The single-channel on-off pattern served as the basis for extracting metrics on an individual level. For group level metrics, the single-channel on-off pattern were combined to reconstruct the different *conversational realities* of the different sites. The on-off pattern of the person, whose conversational reality was to be reconstructed, was combined with the other persons on-off pattern, which was priorly delayed according to the conversation’s technical condition. In line with the commonly used state model, conversational states were assigned to each sample of each reality.

Since not only two-party conversations were examined, the two-party state model (Fig. 7.1a) needed to be extended to the situation of more than two interlocutors. For the case of three interlocutors, Figure 7.1 b depicts how the complexity increases. The state of *individual or single talk* of the third person ($I_{3,C}$: individual/single talk of person C), two more states of *multi or double talk* ($M_{3,AC}$: multi/double talk of person A and C, $M_{3,BC}$: multi/double talk of person B and C), and the state of *multi or triple talk* ($M_{3,ABC}$: multi/triple talk of person A, B and C) were added in addition to the two-party states. Transitions that are very unlikely, those with an increase of two or more speakers in the next moment, are kept in the background (continuing

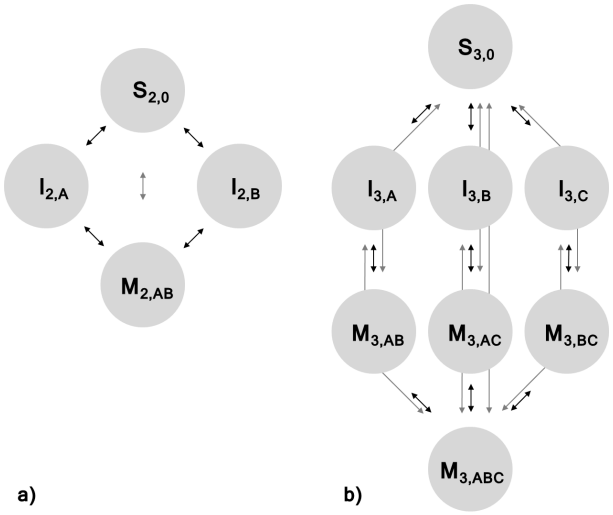


Fig. 7.1: Possible states (sets and subsets) for a two-party or multi-party conversation situation; $S_{N,0}$: silence in a group of N participants, $I_{N,A}$: individual/single talk of person A in a group of N participants, $M_{N,AB}$: multi/double talk of person A and B in a group of N participants, $M_{N,ABC}$: multi/double talk of person A,B and C in a group of N participants; state transitions are indicated by arrows. Improbable transitions have been omitted.

behind the states) and grey in colour. In a similar manner, the model can be extended to four or more interactants.[3]

The possible states can therefore be (re)defined as follows:

- S: Silence; No speech of any person is present.
- I: Individual or single talk; Only one person's speech can be detected at a particular point in time.
- M: Multi talk; The speech of two or more people can be detected at the same point in time.

7.3.1.1 Multi-Talk

From the literature review provided in Chapter 3, it can be concluded that the probability for multiple people talking at the same point in time seems to reflect stages in which transmission delay has different implications (Braun, 2003). In the case of low delays, double or triple talk seems to be a natural part of the turn-taking organisation used. For instance, for completions at medium delays, the number of

[3] To avoid confusions between the abbreviation "S" for silence and the according one for single talk, the latter was indicated by an "I" for individual talk

backchannels and seamless turn hand-overs are reduced to facilitate a mostly undisturbed conversation. For high delays, double and triple talk can no longer be avoided due to factors such as unintended simultaneous starts.

To investigate if such a U-shaped relationship between the number of state occurrences and delay can be held when investigating conversations, the probability of double talk and triple talk shall be examined. The probability of triple talk is usually rather small since it does not occur very often, however it should not be neglected. Therefore, the probability for double talk and triple talk will be added to the **probability for multi talk ($\mathbf{P}_{multi}$)**, with N reflecting the number of participants per group.

for N = 2

$$P_{multi} = P(M_{2,AB}) \tag{7.1}$$

for N = 3

$$P_{multi} = P(M_{3,AB}) + P(M_{3,AC}) + P(M_{3,BC}) + P(M_{3,ABC}) \tag{7.2}$$

7.3.1.2 Divergence

It was considered to be required to have measures that reflect the divergence of the two conversational realities caused by transmission delay. It was of particular interest to compare a person's actual behaviour to how their behaviour was perceived by the other person at the far-end. To this aim, sequences of conversational states will be examined in detail which are referred to as *state walks*. It will be the main goal here to look at divergences for all possible state walks, to not only acquire a more comprehensive understanding but to also include the case that one kind of state walk is perceived as a completely different kind of state walk at the other end. For instance, an alternating silence may be perceived as a successful interruption.

A state walk is a sequence of changes in the conversational state (shown in Fig. 7.1) as experienced by one participant, beginning in a period of individual/single talk and ending in the next period of individual/single talk. If divergences regarding the occurrences of certain state walks of one compared to the other one's reality can be detected, it confirms the assumption that the conversational courses get distorted by delay in a way that communication gets more difficult.

Table 7.1: Considered State Walks, for the viewpoint of person A

Name	**Abbr.**	**State Walk**	**Speaker Turn**
Break	BR_A	$I_{N,A} \rightarrow S_{N,0} \rightarrow I_{N,A}$	no
Alternating Silence	AS_A	$I_{N,A} \rightarrow S_{N,0} \rightarrow I_{N,B/C}$	yes
Non-successful Interruption	NI_A	$I_{N,A} \rightarrow M_{N,AB/C}$ ($\leftrightarrow M_{N,ABC}$) $\rightarrow I_{N,A}$	no
Successful Interruption	SI_A	$I_{N,A} \rightarrow M_{N,AB/C}$ ($\leftrightarrow M_{N,ABC}$) $\rightarrow I_{N,B/C}$	yes

Table 7.2: Confusion matrix for the reality of person A; entries in fields with ○ counting for *match*, ● counting for *mismatch*

A \ B	BR.AA_B	AS.AB_B	NI.AA_B	SI.AB_B	BR.BB_B	AS.BA_B	NI.BB_B	SI.BA_B
BR.AA_A	○	●	●	●	●	●	●	●
AS.AB_A	●	○	●	●	●	●	●	●
NI.AA_A	●	●	○	●	●	●	●	●
SI.AB_A	●	●	●	○	●	●	●	●
BR.BB_A	●	●	●	●	○	●	●	●
AS.BA_A	●	●	●	●	●	○	●	●
NI.BB_A	●	●	●	●	●	●	○	●
SI.BA_A	●	●	●	●	●	●	●	○

Based on the resulting state structure (Fig. 7.1), the following eight state walks (some of them being speaker turns) were extracted and analysed (Table 7.1). Note that a non-successful interruption (NI) describes situations where the person initially began to speak in single talk ($I_{N,A}$, $I_{N,B}$ or $I_{N,C}$) and then speaks again in single talk after a multi talk period ($M_{N,AB}$, $M_{N,AC}$, $M_{N,BC}$, $M_{N,ABC}$). This walk represents the complement to the successful interruption (SI) where the speaker changes but without a speaker turn. In case of transmission delay, an event perceived as an SI at the one end can be perceived as an NI at the other end, or even as an alternating silence (AS), and vice versa. In Figure 7.2, possible divergence is illustrated. The conversation shown there begins with a speech utterance by person B. In his view he pauses and starts to talk again after a break. The abbreviation for this transition walk (BR.BB_B) is structured as followed: First we find the abbreviation for the kind of walk that was perceived - in this case a break (BR). After a dot, an abbreviation for the person who spoke at the beginning, in this case person B, and the one who spoke at the end of the walk are placed, here, also person B (.BB). Following an underscore, we find the indicator for the person whose view we are looking at (_B). In contrast, from the perspective of person A, an alternating silence is perceived (AS.BA_A). After person B spoke and turned silent, person A responded (.BA). In the course of the conversation, a few other common possible examples of diverging realities are given (Fig. 7.2). In sum, it can be said that the higher the delay is, the more likely is an overall divergence of the two realities.

Here, the term **divergence** describes to which extent the courses of conversation are perceived differently by the interlocutors. For calculating the divergence, each walk occurring in each persons "own" reality was linked to the corresponding walk observable in the "other" persons reality. As a result, a confusion matrix was created for every person to every "other" person in each conversation. Only such walk where the person him or herself and one respective "other" were involved were considered for one matrix. This implies that in a three-party group two matrices per person are needed, for instance, one matrix out of person A's point of view compared to person B's point of view and one from A's perspective compared to person C's perspective (similar there were matrices calculated for B to A, B to C, C to A and C to B). The structure of each matrix is shown in Table 7.2. Walks found in the persons "own"

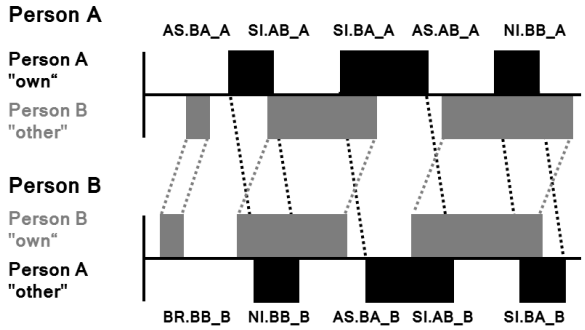


Fig. 7.2: Example of conversation with high divergence of the two realities due to a delayed circuit; BR: Break, AS: Alternating Silence, NI: Non-Successful Interruption, SI: Successful Interruption

reality (in Table 7.2 person As reality; for person Bs point of view, person Bs reality) are entered in the rows, the related walks in the "other" person's reality (in Table 7.2 person B's reality; for person B's point of view, person A's reality) are entered in the columns. The diagonal contains the number of walks in terms of the person's "own" reality that were perceived as the same kind of walk in the "other" person's reality. The sum of these will be referred to as the number of matches. All other cases, where a walk in the "own" reality was perceived as a different walk in the "other" one's reality, will be called mismatches. The divergence is calculated as the ratio of mismatches and matches. For three-party cases, the sum of both mismatch values and the sum of both match values obtained served to compute this ratio. For person A's point of view the divergence can be calculates as follows:

for N = 2

$$divergence = mismatch_{AtoB}/match_{AtoB} \tag{7.3}$$

for N = 3

$$divergence = (mismatch_{AtoB} + mismatch_{AtoC})/(match_{AtoB} + match_{AtoC}) \tag{7.4}$$

Additionally, the assumption is made that if the sum of mismatches is zero the divergence is zero, as well. We expect that a high degree of divergence makes a fluent conversation impossible. To further analyse this assumption, the mean divergence was taken as an indicator for the degree of distortion caused by transmission delay.

7.3.1.3 Speaker Alternation Rate Corrected

As already described the *Speaker Alternation Rate (SAR)* calculates the number of speaker turns per minute interval (Hammer et al, 2004). Due to delay, longer silence

periods e.g. when waiting for a delayed response of the other is likely to result in an overall longer duration of a conversation (there are also other reasons such as confusions in the conversation that add to the duration). For this reason, the SAR is typically very much dependant on the pure increase of the response time.

To see whether the density of speaker turns actually decreases or whether longer waiting times just lead to the drop of SAR with increasing delay, a corrected version of the SAR, **Speaker Alternation Rate corrected (SARc)**, will be examined. For this correction, the overall duration of the conversation will be reduced by the product of the number of alternating silences that a person started and the respective other person ended (for the view of Person A: # AS.AB_A) multiplied by the double delay time (Ta), is equal to the minimum response time. It is clear that this correction is not completely correct because for all walk types the delay also leads to behavioural changes. However, particularly in the cases where a person waited until the other person responded as a result of the delay added to a lengthening of the overall duration in minutes (DUR) and is assumed to proportionally lengthen the interval considered for speaker alternations.

Hence, to exemplify the calculation for Person A, SARc can be expressed as:

$$SARc = (\#AS.AB\_A + \#AS.BA\_A + \#SI.AB\_A + \#SI.BA\_A)/(DUR - (\#AS.AB\_A * 2 * Ta)) \tag{7.5}$$

7.3.1.4 Utterance Rhythm

Apart from the divergence evaluation, we were aware that fast interaction leads to more critical quality ratings than those of medium to slow pace (Kitawaki and Itoh, 1991). While a high degree of divergence is likely for quick interaction, without a clear conversation structure, participants need to slow down to keep the conversation feasible. For this reason, it was chosen to assess the utterance pace of individuals. The utterance pace of an individual is also dependant on the behaviour of the interaction partner, and on the content of the conversation.

The *Utterance RhYthm (URY)* was introduced as a parameter which is based on the individual talkspurts of the interlocutors A or B, and represents the mean time from the beginning of an "own" utterance until the beginning of the next "own" utterance of the same person, including the utterance and the silence duration (see Fig. 7.3 for illustration). URY thus differs crucially from the SAR. For the SAR, the number of speaker changes per minute is calculated, whereas for URY, the mean time for one utterance and the respective pause until the next one begins is computed for one person, while no speaker change is required in between. Hence, SAR addresses a group level and URY the individual level.

It should be noted that all possible state walks may be reflected in this measure, as well as the state lengths, and together they determine the value of URY. The "own" utterance can be a time where the person is holding the turn solitary (individual/single talk), where they are speaking while someone else is speak-

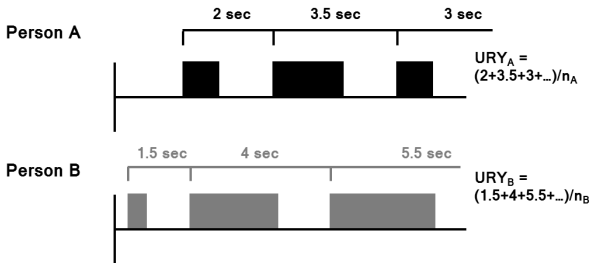


Fig. 7.3: Example for the measurement of the new parameter *Utterance RhYthm (URY)*

ing (multi/double talk), or a mixture of both (e.g. first individual/single and then multi/double talk). The end of an utterance can be caused by their own wish (AS) or due to an interruption (SI). A similar analysis is true for the "own" silence. This measure does not, like others, focus only on a subset of state walks, or exclude certain states but rather examines the speaking behaviour of an individual person as a whole.

Choosing a measure including both the utterance and the silence duration is considered as particularly meaningful as it covers the possibility that people adapt their utterance length and the silence length concurrently. From a theoretical point of view, the measurement of utterance rhythm seemed promising since rhythmic processes represent a key concept of the human organism (Glass, 2001) and especially in human language (Cutler, 1994), from an early age on (Jaffe et al, 2001; Trevarthen, 1993). It should be noted that the favoured interaction tempo appears to be highly dependent on the individuum, for example, whether he or she is an impatient or a calm person, as well as their current state, for example, whether they are in a rush or not. For short interactions with speech dialog systems, it could be demonstrated that if the response speed of the dialog system was adapted to the response time of the user, he or she preferred this over a slower or faster response speed (Ward and Nakagawa, 2004). If people tried to adapt their response time to each other, this failed in a delayed circuit. Dialogues of the experiments described here are rather short, which is why detecting an adaption over time is difficult. Nevertheless, to examine utterance rhythm adaption over time will be an interesting approach for future work. The approach of describing human behaviour in terms of a rhythmic process is common in the domain of sensorimotor synchronisation (SMS) (Repp, 2005) which is concerned with the coordination of motor behaviour involving references to former action, processes or events (Pressing, 1999), i.e. tapping a rhythm with a finger. In a similar sense, speech can be thought of as a motor action referring to a prior action, that is, a person's own speech or of the other person's speech.

In the context of delay and examining its influence on conversations, it seems to be of highest interest to investigate how much people adjust their mean utterance pace with the communication conditions. Additionally, it is known that different conversation types demand different types of interaction. For this reason, each

URY value measured for one person in a particular conversation was decreased by the URY value achieved in the same type of conversation by the same person without added transmission delay, calculating what will be referred to as **Utterance RhYthm change (ΔURY)** in the following. This results in a reduction of the dependency of the conversation feature on the initial task speed and the initial individual utterance speed.

7.3.2 Nonverbal and Nonvocal Conversation Metrics

As aforementioned, to date minimal work investigating automated assessment of visual activity based on recoding of videotelephony or videoconferencing has been reported. Only in one study by Kurita et al (1994) the degree of motion was deliberately varied by adding gestures to the investigated responses. For most tasks, no difference in CQ ratings were identified between conditions where additional movement was or was not present. Only in a toss-up task with video-only delay (asynchronous delay condition) ratings were more critical if the response-type involved gestures. These results indicate a correlation between the amount of motion and the experienced quality.

From the domain of social signal processing, the extraction of the visual activity is a well-known approach for analysing participant behaviour during meetings or conferences (Aran and Gatica-Perez, 2011). Here, the motion vector magnitude estimates the amount of motion shown by the people recorded. This feature has been used in the context of dominance, role and leadership prediction mainly in the analysis of face-to-face interaction in the literature so far.

In the present work, the motion vectors were extracted per pixel of each frame from the recorded one-person videos using a dense optical flow algorithm based on primal-dual convex optimization, as proposed by Werlberger et al (2010). The norm of the motion vector per pixels was then summed per frame and then per video. The video-motion values were, in the last step, related to the number of frames of the corresponding video. In the end, each video of each person led to one overall motion value for the feature here termed **Overall Motion (OM)**.

Considering that the image size of the recordings was small for low resolutions as explained in Section 6.5, the amount of detected motion can be expected to also be small. Therefore, even if the proposed dense optical flow algorithm generally provides accurate motion estimation, it can be challenging for the algorithm to estimate the very small amounts of motion in videos of low resolution. To avoid adding issues of differences in motion estimation performance between the “lower” and “higher” recording cases into the statistical analysis, the subsequent analysis is performed separately per image size.

7.3.3 Verbal and Vocal Conversation Metrics

Based on insights from the ISO 24617-2 standard 2012 called "Language resources management - Semantic annotation framework (SemAF) Part 2: Dialogue acts" which is i.e. based on Bunt et al (2010) and the works on microanalysis of delayed mediated communication (Ruhleder and Jordan, 1999, 2001), four dimension-specific dialogue-act functions were annotated with the software ELAN (Max Planck Institute for Psycholinguistics, The Language Archive, 2008).

Due to restricted resources and the novelty of the approach, not knowing which of the initially selected measures will be helpful, annotations were not done by several annotators separately, and therefore no κ values, indicating the correspondence of different raters, were calculated. Two linguistic students were conducting the annotations, regularly discussing and mutually assuring the strategy under the supervision of a Doctor of Linguistics and the author. Furthermore, annotations were only done for the symmetrical delay conditions of the fullband three-party experiment (A.FB.3) and the synchronous delay conditions of the audio-video experiment (V.WB.2) owing to the high effort (time and costs).

The following measures were annotated with the corresponding annotation strategies and the occurrence of the corresponding categories were counted:

Contact-Management

- *contact-check*; e.g. "Hello are you still there?"
- *contact-indication*; e.g. "I can here you."

A contact-indication does not require a prior contact-check.

Communication-Management

- *own-communication-management*, including retraction and self-correction
- *partner-communication-management* including completion and corrections

Own-communication-management is expressed by the speaker holding the floor while partner-communication-management is expressed by a person not holding the floor.

Corrections of the partner were only counted as partner-communication-management if it was clear to all parties that the previous expression was not correct and if the correction did not have a question format. Otherwise it was counted as an auto-negative feedback.

Feedback

- *auto-feedback*; The person currently does not have the floor shortly expresses their own processing of a given utterance from another person. Respective subcategories are:
  - *auto-positive*; e.g. "Mhm." "Right."
  - *auto-negative*; including clarifications, e.g. "Could you repeat this please."

  - *auto-repetition*; e.g. "... for this reason we take the green door." "The green one, ok."
  - *auto-summary*; including self- and partner-repetition and self- and partner-summary; e.g. "We need to do something for the birds, mice and fishes in that area." "Ok, protecting the animals."

- *allo-feedback*; When the person currently holding the floor elicits information on the processing status of the utterance priorly expressed from the person currently not holding the floor. With the subcategories:

  - *allo-positive*; If the person currently holding the floor thinks that the other one understood his or her prior expression. E. g. "That's clear, right."
  - *allo-negative*; If the person currently holding the floor thinks that the other one did not understood his or her prior expression. E.g. "Did you understand what I mean?"

Regarding the definition of the measure feedback, it was difficult to select the conditions in which an utterance should be counted. Converse to some definitions a rather strict strategy was followed, only including expressions as feedback if they were short (entire sentences were not counted) and did not involve a turn change. For example, responses to questions were not counted because usually they lead to a turn change. They were also not counted if the response was very short (i.e. "yes.").

When taking up the turn was followed instantly after the feedback by the same person it was not counted as feedback, but rather considered as a strategy for getting the floor. Short laughing or cheering to show that the prior expression was understood was counted as positive auto-feedback. Laughing with the other person together was not counted.

Stops

- *self-induced-stop*; When the person currently holding the floor suddenly stops their expression without completing it later for no obvious reason.
- *partner-induced-stop*; When the person currently holding the floor suddenly stops their expression without completing it later because someone interrupted them.
- *failed turn claims*; If a person tries to obtain the floor but the person holding the floor does not allow it.

Stops were counted in general only if expressions were terminated still being incomplete and were not taken up again.

If the partner finalised the expression it was only counted as a partner-communication-management because in most cases it was not clear if the stopping person would have taken the floor again or actually stopped.

In the case of a self-induced-stop the usually unobvious reason for stopping might be due to the person thinking they were already understood, realising that it was not their turn or they did not know how to continue.

Chapter 8
Results

8.1 General Approach

In the following, the eight experiments conducted are grouped into five data sets, per bandwidth (NB, WB and FB) and mode (audio-only versus audio-video). Experiment A.NB.2a is analysed separately because it was the very first experiment where the different task types were accomplished in a randomised order. As the results will show, this leads to slightly different results especially regarding the ratings of Conversational Quality. Experiment A.NB.2b and A.NB.3 are merged to one data set, abbreviated with A.NB. Similarly, A.WB.2 and A.WB.3 build one set, abbreviated with A.WB, and A.FB.2 and A.FB.3 build another one, abbreviated with A.FB. Experiment V.WB.2 is also considered separately due to its audio-video mode. The data grouping allows to statistically analyse the factor communication group size (.2 versus .3). As a further advantage, the factor *task-block order* can be examined for the A.NB data.

Due to changes in the experimental designs, the assessed scales and the tested delay levels changed. From the NB to FB experiments and from FB to WB experiments it can be expected that differences are rather high. For this reason, the entire data was not merged to one big data set. As a consequence of this approach, the influence of the factors bandwidth and mode can only be examined by visual inspection of the graphs but not on a statistical level.

For all analyses the factors *delay* condition and *task type* (if more than one task type was tested) are considered. These two factors were varied within each experiment, whereas the factor *group size* was varied between experiments. In some cases other factors were also investigated, these will be mentioned in the corresponding sections below. In case there is only one dependant variable, such as for the ratings of Conversational Quality, the data is analysed by means of an ANalysis Of VAriance (ANOVA) and subsequent post-hoc test with Bonferroni adjustment of alpha. If more than one dependant variable is considered and if it can be expected that the

variables are related, a Multivariate ANalysis Of VAriance (MANOVA) was used to analyse the data. [1]

8.2 Conversational Quality Ratings

Five full-factorial ANOVAs and subsequent post-hoc tests were conducted to be able to answer the hypotheses (corresponding ANOVA Tables can be found in Appendix I, Table 11.1-5). For all data sets the factor *delay* is examined. In the data sets A.NB.2a, A.NB and A.WB the *task type* and in A.NB, A.WB and A.FB the *group size* constituted additional factors. All factors were treated as fixed factor. As the dependant variable, the Mean Opinion Score (MOS) is taken as the indicator for the Conversational Quality.

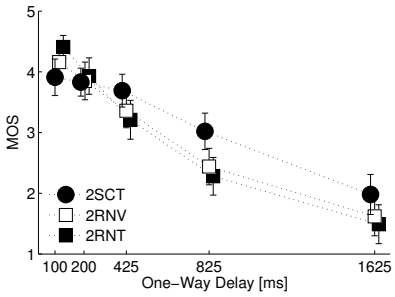
Delay

The factor *delay* is found to be a significant factor regarding the MOS ratings in all data sets except the A.FB data (Fig. 8.2 (b)). It also represents the factor with the biggest effect size among all main factors and interactions of factors, $\eta^2_{partial}$ ranges from 0.196 to 0.475 in the significant cases (for details see Appendix I, Table 11.1-5). Post-hoc tests show significant differences among nearly all conditions. Usually the low *delay* conditions are not significantly different compared to each other.

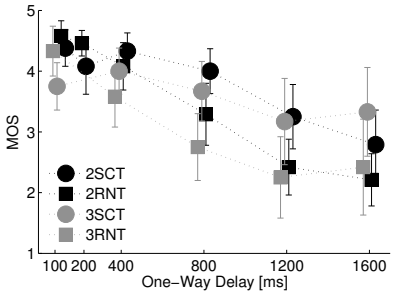
Significant Post-hoc test for the factor *delay* regarding the MOS ratings are obtained for the following conditions:

- A.NB.2a: all conditions differ significantly from each other, except for 100 to 225
- A.NB: all conditions differ significantly from each other: except for 100 to 200 and 200 to 400
- A.WB: all conditions differ significantly from each other: except for 75 to 135 and 400, 135 to 400
- A.FB: no effect for *delay*
- V.WB.2: condition 0 differs significantly from all other conditions, except 400 and 400*a*; condition 400*a* differs from 0 and 800*a*; and condition 800*a* significantly differs from all other conditions

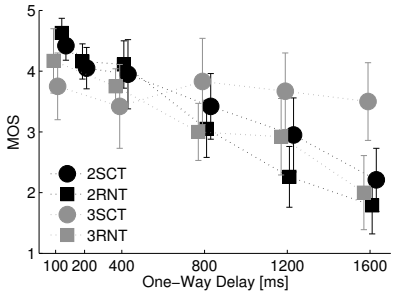
[1] For parametric statistical test methods, such as the ANOVA or MANOVA, it is assumed that an independence of observations, normal distribution of the residuals and homogeneity of variances is given on a theoretical level. Authors disagree about whether to apply or not to apply non-parametric test methods if the assumptions cannot be confirmed with the empirical data. It can be summarised that parametric test methods are robust against violated assumptions if cell sizes are equal and that effect sizes are difficult to interpret for non-parametric tests procedures (Bühner and Ziegler, 2009). For this reason, parametric models were applied in any case in this study.



(a) A.NB.2a



(b) A.NB; *task-block order* first SCTs then RNTs



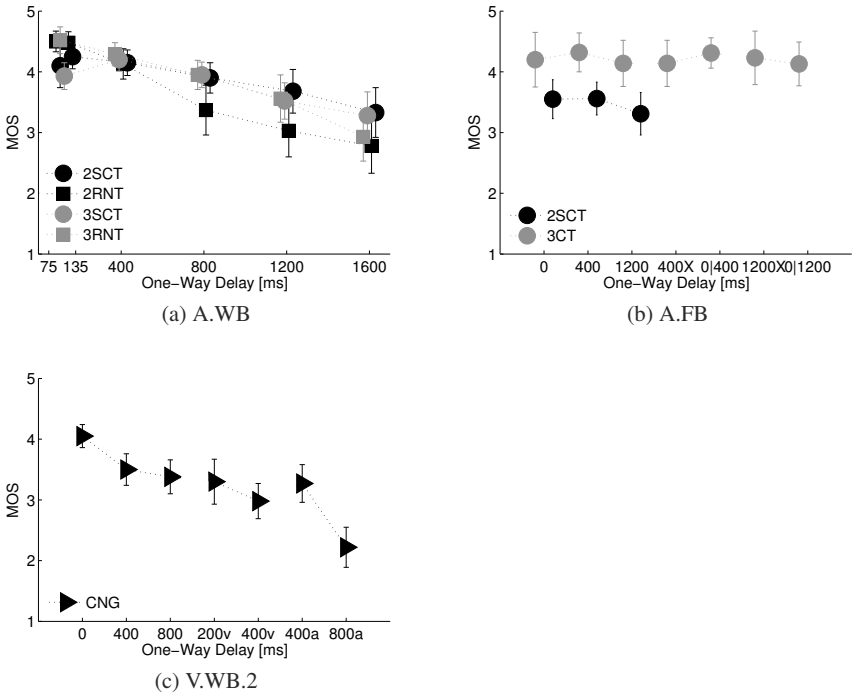
(c) A.NB; *task-block order* first RNTs then SCTs

Fig. 8.1: Mean Opinion Scores (MOS) and 95 % CIs for different task types in different delay conditions; 2SCT: Short Conversation Tests (two-party), 2RNV: Random Number Verification (two-party), 2RNT: Random Number Verification Timed (two-party)

There are no significant differences for symmetrical versus asymmetrical conditions in the A.FB.3 experiment. In experiment V.WB.2, asynchronous audio-video delay leads to lower MOS ratings. Particularly the condition 800*a* (800 ms audio-only delay) is rated lowest significantly, different from all other. Surprisingly, doubling the synchronous delay from 400 ms to 800 ms does not change the MOS ratings at all here.

Task Type

The factor *task type* turns out to be significant in the A.NB.2a and in the A.NB data but not in the A.WB data (there was no *task type* factor for the A.FB and the V.WB.2 data). In both data sets, A.NB.2a and A.NB, the SCT task differs significantly from



(a) A.WB

(b) A.FB

(c) V.WB.2

Fig. 8.2: Mean Opinion Scores (MOS) and 95 % CIs for different task types in different delay conditions; 2SCT: Short Conversation Tests (two-party), 2RNT: Random Number Verification Timed (two-party), 3CT: Conversation Tests (three-party, long version), 3SCT: Short Conversation Tests (three-party, short version), 3RNT: Random Number Verification Timed (three-party), CNG: Celebrity Name Guessing Task (two-party)

the RNT task (confirmed by post-hoc tests). The RNV task does not differ from the other two task types in a statistically significant way. For this reason, it was decided to not use the RNV task in any further experiments.

In all cases where the factor *task type* is tested (A.NB.2a, A.NB, A.WB) the interaction of *delay* and *task type* is found to be significant. This interaction shows the highest $\eta^2_{partial}$ values among all interactions although the values are not very high, ranging from 0.033 to 0.060. As can be seen in Figure 8.1 and 8.2, ratings of RNTs decrease more with increasing delay. This trend is smaller for the A.WB data than for the A.NB data. For the three-party experiment A.WB.3 there is hardly any difference between 3SCT and 3RNT MOS ratings (Fig. 8.2 (a)).

As explained earlier, for reasons of efficiency the 3CTs used in the A.FB.3 experiment were shortened for the A.WB.3 experiment. The shortened version, the 3SCTs, were thus investigated in A.WB.3. When comparing the two outcomes, a

small decrease of the rated quality can be found for the 3SCTs (Fig. 8.2 (a)) whereas no such decrease can be seen for the 3CT (Fig. 8.2 (b)). This difference could also be due to the higher bandwidth in the A.FB data or other aspects of the design that will be discussed later.

Task-Block Order

The factor *task-block order* is only addressed in the experiments A.NB.2b and A.NB.3 (combined to the A.NB data set). Surprisingly it is not found to be a significant main factor. However, the interaction of *group size* and *task-block order* is significant. In Figure 8.1 (b) and (c) the two *task-block order*s are compared. It can be seen that the graphs for 2RNT and 3RNT fall similarly, within a certain variance, with increasing delay for both *task-block order*s. When looking at the 2SCTs, a change in dependence on the *task-block order* can be detected. The slope for 2SCTs has a steeper shape in the RNT-SCT *task-block order*. In particular, in the RNT-SCT *task-block order* (Fig. 8.1 (c)) the SCT-MOS ratings for 800 ms of delay and higher values are rated about 0.5 MOS points lower than in the SCT-RNT *task-block order* (Fig. 8.1 (b)). The 3SCTs-MOS ratings, on the other hand, are similarly high in both *task-block order*s and hardly decrease with delay.

When further comparing the 2SCT slope of Figure 8.1 (a) (experiment A.NB.2a) to the 2SCT slope of Figure 8.1 (b) it becomes clear that the randomised order of tasks in A.NB.2a leads to more critical SCT-MOS ratings. The mean MOS ratings at 1600 ms delay is about one MOS point lower in A.NB.2a than in the A.NB.2b SCT-RNT order or in the A.WB.2 experiment (Fig. 8.2 (a)).

Group Size

The *group size* is found to be significant only in the A.FB data but not in the A.NB or A.WB data. Nevertheless, a significant interaction of factors *group size* and *delay* is observed in the A.NB data set. Three-party groups rate the quality generally lower than two-party groups for delays below or equal to 400 ms and on a similar level (in some cases higher or lower) for higher delay values.

8.3 Mediated Interaction Measures

8.3.1 Nonverbal and Vocal Conversation Metrics

For every data set one full-factorial MANOVA, was computed having the metrics probability for multi-talk (P_{multi}), Speaker Alternation Rate corrected (SARc), Utterance RhYthm change (ΔURY) and divergence as indicators for the nonverbal

vocal conversation structure and the *task type* (if varied), the *delay* and *group size* (if varied) as factors. The detailed MANOVA Tables can be found in Appendix I (Tables 11.6-9).

For all data sets the factors *task type* (if varied) and *delay* are found to be significant. Here, in contrast to the conversational quality ratings, the factor *task type* shows higher $\eta^2_{partial}$ values (range: 0.129 - 0.794) than the factor *delay* (range: 0.126 - 0.271). The factor *group size* is also significant in all data sets tested ($\eta^2_{partial}$: 0.320 for A.NB, 0.739 for A.WB and 0.698 for A.FB). The corresponding $\eta^2_{partial}$ values are lower than for the factor *task type* and higher than for factor *delay*.

In all cases where the *task type* is varied the interaction of *task type* and *delay* is significant, as well. This is in line with the Conversational Quality ratings. The interaction of *delay* and *group size* is also significant for all relevant data sets (A.NB, A.WB., A.FB). It is interesting to note, that the interaction of *task type* and *group size* is significant for the two cases where this interaction is possible (A.NB and A.WB) and there shows the highest $\eta^2_{partial}$ values among the interaction effects.

Even though all independent variables are examined together in one MANOVA for each data set, in the following, the outcomes for the independent variables will be described per variable and not per data set. This gives the possibility to compare the outcomes of one variable in one data set to the outcome of the same variable in a different data set.

The inbetween-factors analysis shows that, if a main effect or interaction effect is significant the single effects regarding the probability for multi talk, P_{multi}, and the Speaker Alternation Rate corrected, SARc, are in all cases significant, as well. For the divergence metric this is also the case except for the factor *group size* and the interaction of *delay* and *group size* in the A.FB data set. The Utterance RhYthm change, ΔURY, turns out to be a variable not varying significantly for the A.NB, A.WB and A.FB data sets and the factor *group size*, the interactions of *delay* and *group* size as well as the interaction of *task* type, *delay* and *group size*. In the audio-video data of experiment V.WB.2 the factor *delay* and the dependant variable ΔURY does not show a significant difference among conditions either.

When looking at the graphs, a more detailed picture can be gained. In Figure 8.3, the course of the **probability for multi-talk (P_{multi})** is depicted for the tested delay conditions in the different data sets. It can be seen that P_{multi} drops slightly for the 2RNT task with increasing delay which can be interpreted as a more disciplined turn-handover behaviour leading to less multi talk. The P_{multi} values for 3RNTs stay on a low level across all delay conditions in A.WB. In the A.NB data a small increase of the mean P_{multi} values for 3RNTs can be found. Since variances increase as well this seems to be indicating that participants had more difficulties to maintain the order of reading numbers and hence more multi-talk resulted. Most likely a few groups were causing a lot of double talk while others behaved as instructed. For the 2SCT task a general increase in P_{multi} up to 400 ms of delay detectable can be seen from the graph. For higher delay conditions P_{multi} stays constant, in A.NB.2a the values even drop again which may be related to the also comparably

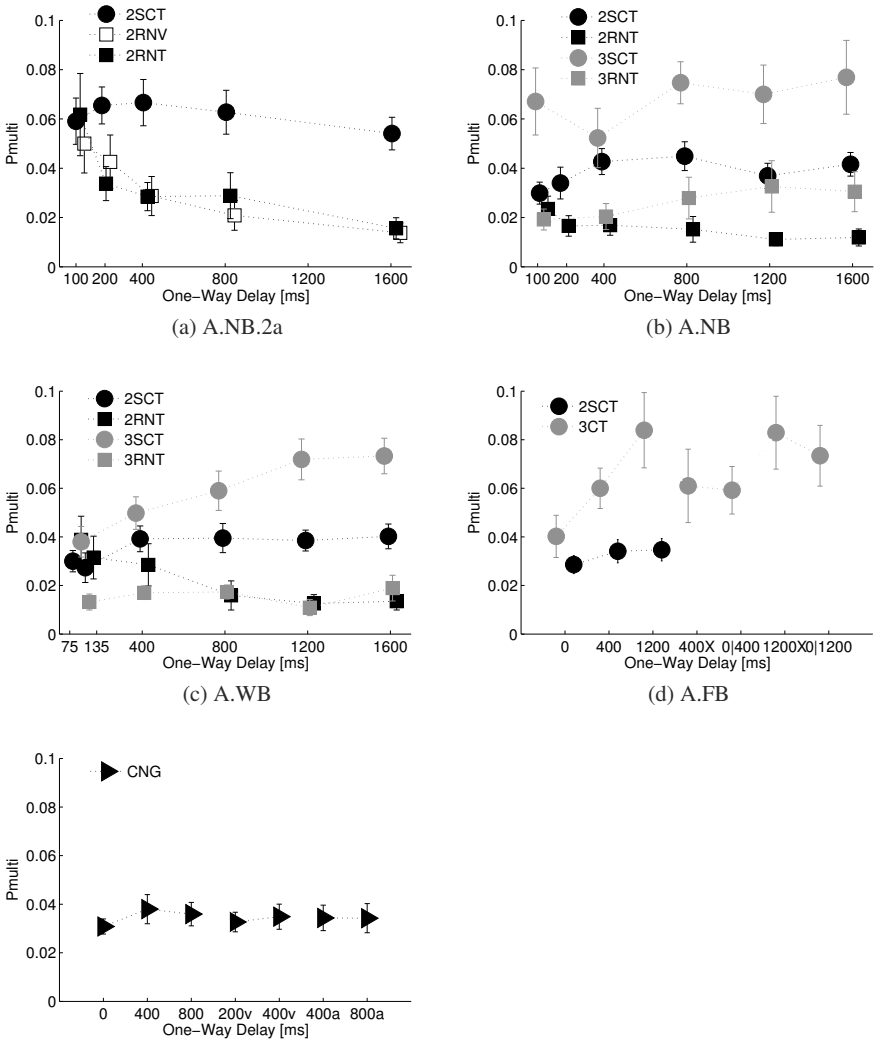


Fig. 8.3: Probability for multi-talk (P_{multi}) and 95 % CIs for different task types in different delay conditions; 2SCT: Short Conversation Tests (two-party), 2RNV: Random Number Verification (two-party), 2RNT: Random Number Verification Timed (two-party), 3CT: Conversation Tests (three-party, long version), 3SCT: Short Conversation Tests (three-party, short version), 3RNT: Random Number Verification Timed (three-party), CNG: Celebrity Name Guessing Task (two-party)

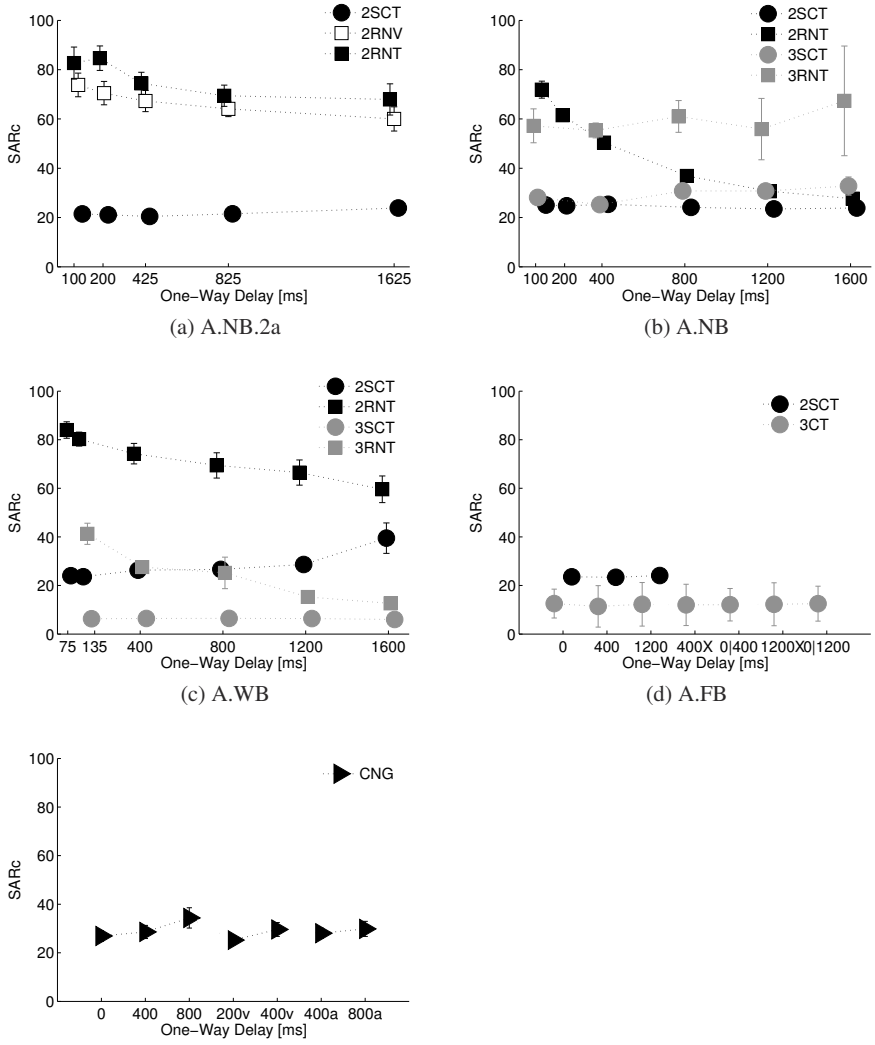


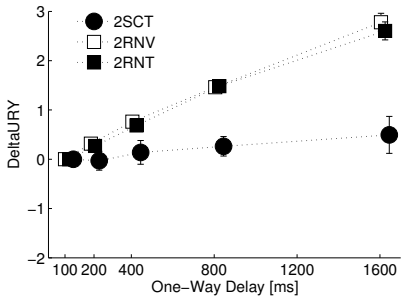
Fig. 8.4: Speaker Alternation Rate corrected (SARc) and 95 % CIs for different task types in different delay conditions; 2SCT: Short Conversation Tests (two-party), 2RNV: Random Number Verification (two-party), 2RNT: Random Number Verification Timed (two-party), 3CT: Conversation Tests (three-party, long version), 3SCT: Short Conversation Tests (three-party, short version), 3RNT: Random Number Verification Timed (three-party), CNG: Celebrity Name Guessing Task (two-party)

low MOS ratings for the SCTs in this experiment. Participants may have detected the technical impairment and thus controlled their talking behaviour. In the three-party SCTs (3SCTs and 3CTs) a general upwards trend can be registered which is more clear in the A.WB and A.FB data. For asymmetrical delay conditions it can be noted that P_{multi} changes in a similar way as in the symmetrical conditions even if the considered person is only speaking with delay to one person (conditions 0|400 and 0|1200). In the audio-video experiment the P_{multi} values change only little with the tested conditions. Nevertheless, P_{multi} is smallest in the reference condition (0) of no additional delay.

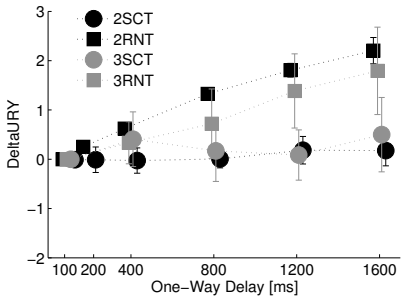
The picture for the **Speaker Alternation Rate corrected (SARc)** looks completely different. A task effect is clearly visible in the graphs. SARc values for 2SCTs and 3SCTs are much lower than the values for 2RNTs or 3RNTs indicating a slower interaction speed. The CNG task is on a similar level as the SCTs. This implies that, number verification tasks have a higher speaker change speed. The slopes for RNTs show a falling tendency with increasing delay which indicates that alternating silences and breaks must have become longer. In case of alternating silences it is even possible that they were longer than it would have been necessary because of delay. Longer overall durations of the calls cause lower SARc values if the number of utterances is constant (which can be expected for the RNT tasks due to the task design). This can be observed in particular for the 2RNT tasks of experiment A.NB.2b (Fig 8.4 (b)). Here, the SARc of the lowest delay condition is already lower than in experiment A.NB.2a (Fig 8.4 (a)) or A.WB.2 (Fig 8.4 (c)) and the decrease of SARc RNT values at 1600 ms delay is higher. This can only be explained by an initially lower interaction speed for RNTs in experiment A.NB.2b and even more slowing down in case of delay. Probably, the instructions to not use short cuts and to always wait for the response of the other before starting to talk again were more enforced in this experiment. 3RNTs seem to be slower already without delay, especially in experiment A.WB.3. In experiment A.NB.3 the SARc RNT values do not decrease but variances get higher indicating that some groups were keeping up their speed while others did not.

The variances of all SARc SCT values are generally very small and are for this reason not visible in most graphs. The SCT SARc values also stay rather constant on a low level across the different delay conditions which supports the assumption that the interaction speed in SCT task stays the same across delay conditions. This in turn implies that the individual utterance speed is likely to stay on a rather constant level as well (see below). Apparently participants of the A.NB.3 experiment were changing the speaker comparably fast because 3SCTs SARc values are on a similar level as 2SCTs and corresponding SARc 3RNTs values are also considerably higher than in experiment A.WB.3.

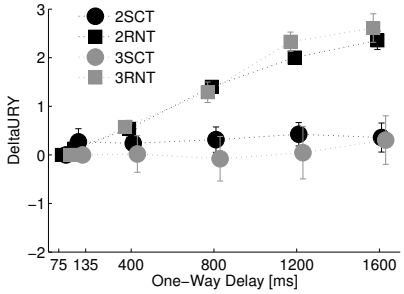
In Figure 8.5 we see the **Utterance Rhythm change (ΔURY)** plotted for the different data sets. It becomes clear why the factor *group size* is not significant regarding ΔURY. There is no systematic variation between the 2SCTs and 3SCTs or the 2RNVs and 3RNVs visible. ΔURY values are on a similar level per task.



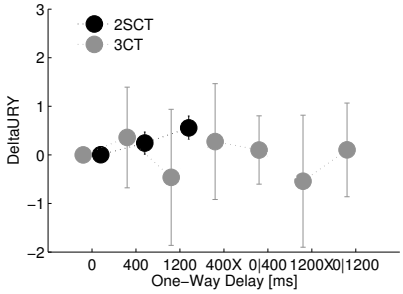
(a) A.NB.2a



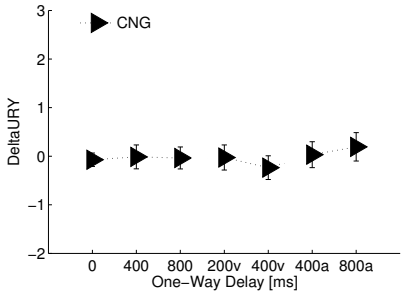
(b) A.NB



(c) A.WB

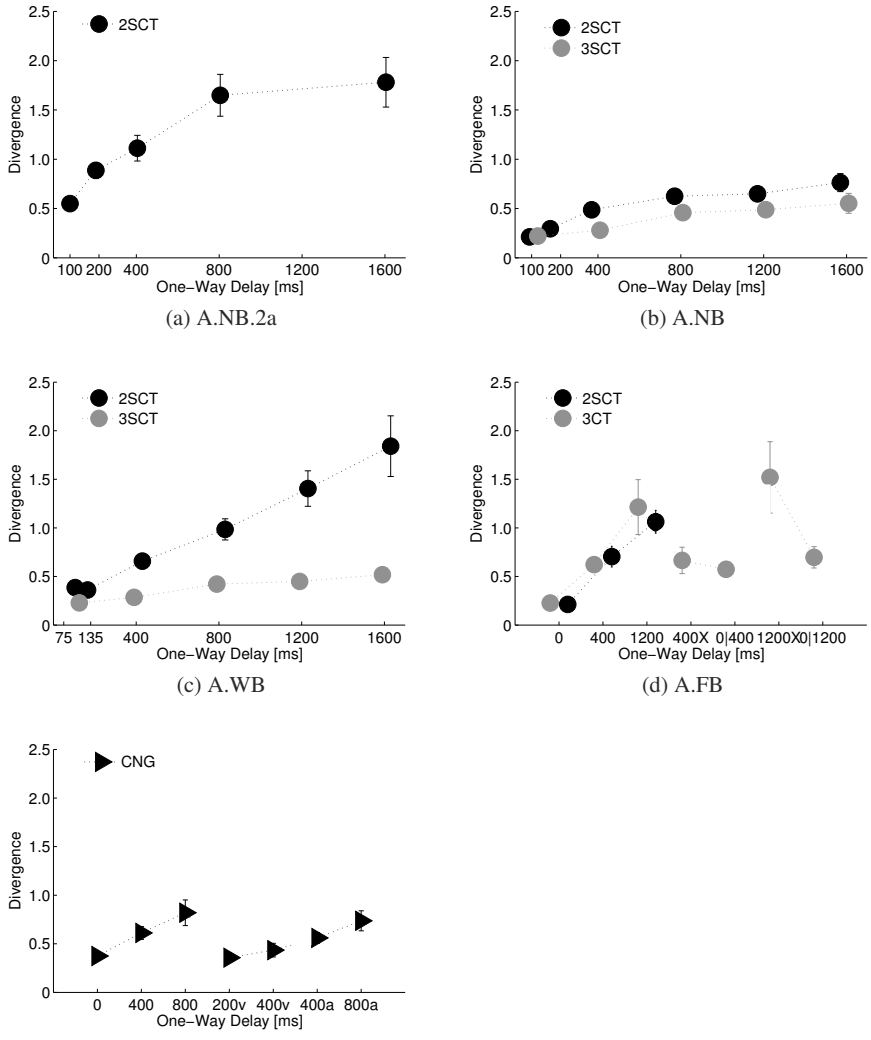


(d) A.FB



(e) V.WB.2

Fig. 8.5: Utterance RhYthm change (ΔURY) and 95 % CIs for different task types in different delay conditions; 2SCT: Short Conversation Tests (two-party), 2RNV: Random Number Verification (two-party), 2RNT: Random Number Verification Timed (two-party), 3CT: Conversation Tests (three-party, long version), 3SCT: Short Conversation Tests (three-party, short version), 3RNT: Random Number Verification Timed (three-party), CNG: Celebrity Name Guessing Task (two-party)



(a) A.NB.2a

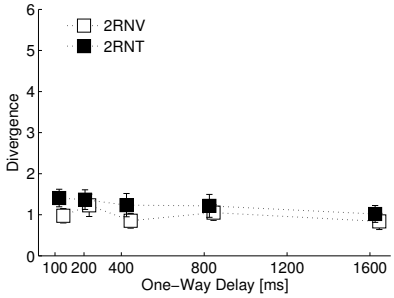
(b) A.NB

(c) A.WB

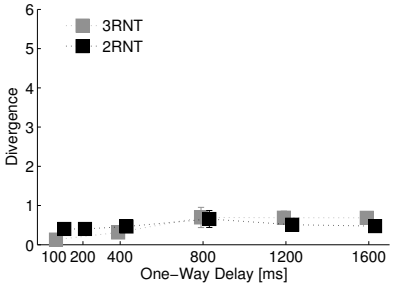
(d) A.FB

(e) V.WB.2

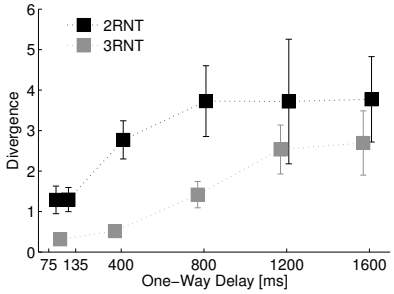
Fig. 8.6: Divergence values and 95 % CIs for SCT task types in different delay conditions; 2SCT: Short Conversation Tests (two-party), 3CT: Conversation Tests (three-party, long version), 3SCT: Short Conversation Tests (three-party, short version), CNG: Celebrity Name Guessing Task (two-party)



(a) A.NB.2a



(b) A.NB



(c) A.WB

Fig. 8.7: Divergence values and 95 % CIs for RNT task types in different delay conditions; 2RNV: Random Number Verification (two-party), 2RNT: Random Number Verification Timed (two-party), 3RNT: Random Number Verification Timed (three-party)

It is interesting to see how different the two task types progress with increasing delay. While the SCT ΔURY values rather stay on one level (around zero), such as the CNG ΔURY values, the RNTs (and RNVs) ΔURY values increase substantially. The 3CT ΔURY values show very high variances indicating that groups may have had very different interaction speeds. Even negative values can be seen which means in these cases participants utterance rhythm was faster in delayed than in non-delayed cases which may indicate impatience. In sum, it can be said that the utterance rhythm gets slower for the RNT/RNT tasks but not for the SCT tasks with increasing delay.

The graphs for the variable **divergence** were split up per *task type* for reasons of clarity. In Figure 8.6 the SCT divergence values are shown. We can see that the divergence increases with delay which means more state walks are mismatching than matching when comparing the courses of state walks at the individual sites. This increase is bigger in the 2SCTs than in the 3SCTs. Surprisingly, the increase in the A.NB 2SCT data is much lower than in all other data sets. For the 3CTs (A.FB

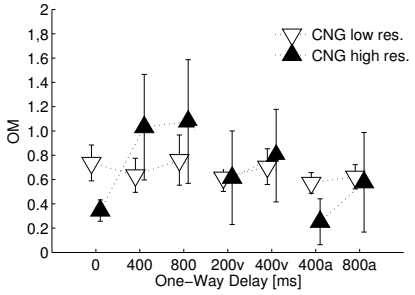


Fig. 8.8: Mean Overall Motion (OM) and 95 % CIs across different conditions for video recordings in high (320x240) and low (160x240) resolution of the two-person video, CNG: Celebrity Name Guessing Task

data) a similarly high increase as for the 2SCT can be found. The extend of increase could be related to how disciplined participants were waiting for the responses of their interlocutors. It could be the case that in the 2SCTs of the A.NB.a and A.WB.2 experiment people were interacting less disciplined than in the 2SCTs of A.NB.2b due to their degree of familiarity. In A.NB.2b all participant pairs were unfamiliar to each other which might have been a reason for a more polite interaction style waiting for the other to finish while half of the groups of A.NB.2a and A.WB.2 were closely familiar interaction partners.

The course of the CNG divergence increase corresponds to the course of the 2SCT divergence in the A.WB data set. The picture for the RNT task divergence values looks quite different (Fig. 8.7). While there seems to be hardly any change in divergence for 2RNTs and 3RNTs in the A.NB.2a and A.NB data sets, the RNT divergence clearly increases and even to higher values than the SCT divergence values in the A.WB data set.

8.3.2 Nonverbal and Nonvocal Conversation Metrics

The mean **Overall Motion (OM)** turns out to differ statistically significant for the tested delay conditions in the higher resolution but not in the lower resolution case (see Appendix I for ANOVA Table 11.11). This becomes clear when looking at the corresponding graphs (Fig. 8.8). In the high resolution case, the OM is significantly higher (confirmed with post-hoc test) for 400 or 800 ms of synchronous delay in comparison to the reference condition (0). The two synchronous delay conditions are also significantly higher than the 400*a* condition (audio-only delay) which shows similarly low motion values as the reference condition. The means of the two video-only delay conditions and the 800*a* condition are located in between the reference and the synchronous cases, but due to high variances no significant difference can be found. In the low resolution case, OM values for all delay conditions are on a

medium level, including the OM of the reference condition. This indicates that the algorithm most likely did not work properly for this resolution.

8.3.3 Verbal and Vocal Conversation Metrics

The verbal conversation metrics were analysed for the two experiments A.FB.3 and V.WB.2 (because they were only annotated for those[2]). A MANOVA with the main factor *delay* was conducted for each experiments data. In the corresponding MANOVA Tables we can see that all verbal conversation metrics differ significantly between the different *delay* conditions in the data of experiment A.FB.3 (see Appendix I; A.FB.3: Table 11.12, V.WB.2: Table 11.13).

In the experiment V.WB.2, however, only three variables are found to differ statistically significant. They are depicted in Figure 8.8 (e). We can see in Figure 8.8 (e) that the mean number of auto-negative Feedbacks (auto-neg) increases only for 800 ms of delay (significant compared to the reference condition in post-hoc tests). The mean number of partner-induced-stops (partner-stops) increases at 400 ms delay (significant to reference condition in post-hoc tests) and decreases again to the level of the reference condition but with a higher variance (no significant difference to any of the two other conditions). A reverse slope can be found for the contact-indications, here, the mean number drops at 400 ms delay (significant to reference condition in post-hoc tests) and rises up slightly again for 800 ms delay (no significant difference to any of the two other conditions).

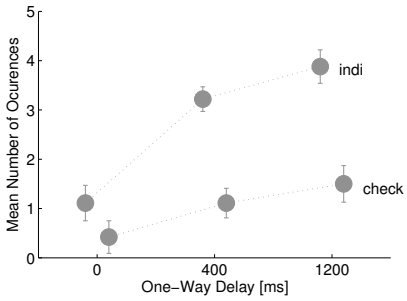
Occurrence values in the A.FB.3 data (Fig. 8.8 (a)-(d)) are approximately double as high as in the V.WB.2 data. This may be due to the fact that in case of three-party calls every participant had two interlocutors and corresponding interactions to both were added together for one person, furthermore, the task differed quite a bit between the two experiments.

For the measure Communication-Management (and auto-positive Feedback) of the A.FB.3 data counted values are reaching considerably higher than for the measures Contact Management and all other indicators for Feedback. The highest increase of the mean number of occurrences with increasing delay is found for partner-communication-management. In general, all mean counts of verbal conversation metrics are increasing substantially from the reference (0) to the 400 ms delay condition and slightly more for the 800 ms delay condition in the A.FB.3 data.

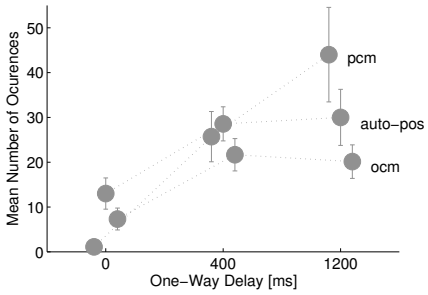
8.3.4 Perceived Fluency

Regarding the perceived fluency of the conversation (FLOW) a full-factorial ANOVA was conducted for the A.WB data set (see Appendix I Table 11.14). Only the main

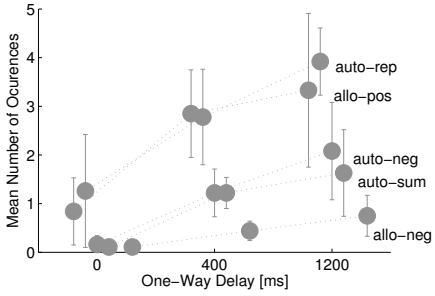
[2] It should be noted that only the symmetrical (A.FB.3) and synchronous (V.WB.2) conditions were annotated and thus included in the analysis.



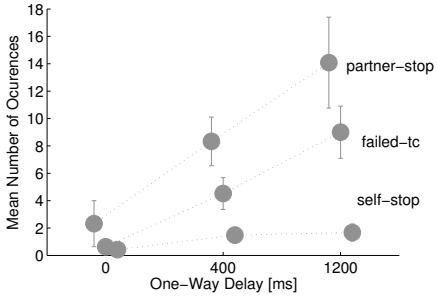
(a) A.FB.3: Contact-Management



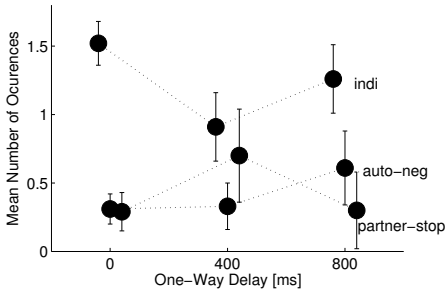
(b) A.FB.3: Communication- Management and Feedback



(c) A.FB.3: Feedback



(d) A.FB.3: Stops



(e) V.WB.2

Fig. 8.9: Mean number of occurrences and 95 % CIs for variables varying significantly with delay conditions; **Contact-Management**: check: contact-check, indi: contact-indication; **Communication-Management**: own-communication-management, pcm: partner-communication-management; **Feedback**: auto-pos: auto-positive, auto-neg: auto-negative, auto-rep: auto-repetition, auto-sum: auto-summary, allo-pos: allo-positive, allo-neg: allo-negative; **Stops**: self-stop: self-induced-stop, partner-stop: partner-induced-stop, failed-tc: failed turn claims

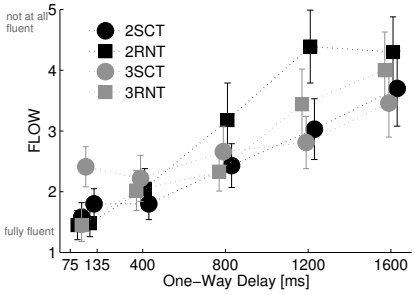


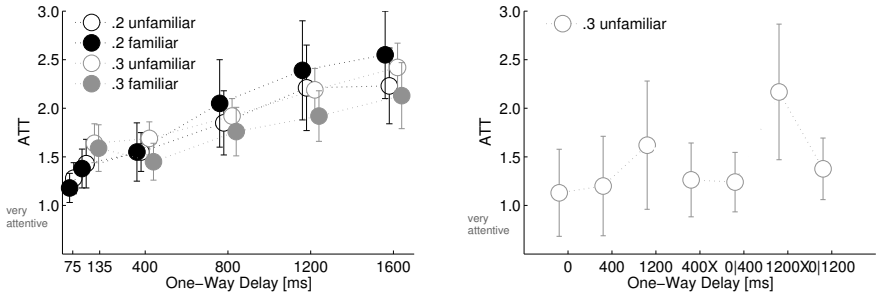
Fig. 8.10: Ratings of the perceived fluency (FLOW) and 95 % CIs across different delay conditions in A.WB

effect for *delay* but not for *task type* or *groups size* turn out to be significant. Post-hoc tests confirm all conditions to be significantly different to each other, except for conditions below 400 ms (75 ms to 135 ms and to 400 ms, 135 ms to 400 ms not significant). The interactions of *task type* and *delay*, *task type* and *group size* as well as *delay* and *groups size* are significant, too, which indicates that lines in the graph are crossing each other.

Figure 8.10 shows that the perceived fluency becomes linearly worse with increasing delay. RNT tasks, especially the 2RNTs, are perceived slightly less fluent for delay conditions higher than 800 ms. In general, the ratings of the three-party experiment are slightly lower than the ones of the two-party experiment for delays higher than 800 ms, indicating the perception of a more fluent conversation. For low delay values in turn the reverse trend can be observed, higher ratings, indicating the perception of less fluent conversation.

8.4 Ratings of Perceived Attributes of the Interaction Partner

For the A.WB data set a full-factorial ANOVA with the factors *task type*, *delay*, *group size* and *familiarity* and the dependant variable rated **perceived inattentiveness (ATT)** of the interlocutors was conducted. It shows a significant main effect for the factor *delay* but not for the other factors. Post-hoc tests on the factor *delay* show that each of the conditions 75 ms, 135 ms and 400 ms was significantly different to the conditions 800 ms, 1200 ms and 1600 ms. Apart from the main effects, the interactions of *task type* and *delay*, *group size* and *familiarity*, as well as the interaction of *task type*, *delay* and *group size* are significant. Since Hypothesis III address the familiarity of participants the focus is put on this factor in Figure 8.11 (a). There, we can see a difference between two-party and three-party ratings for familiar participants but not for unfamiliar ones. If people speak in two-party calls and are familiar they are most critical regarding the attention of the interlocutor the higher the delay. If they speak in a three-party setting and are familiar with each other as well, they



(a) A.WB (b) A.FB.3

Fig. 8.11: Mean rated inattentiveness (ATT) and 95 % CIs for the interlocutors in the different delay conditions; in A.WB by means of an absolute category rating, in A.FB.3 by means of a continues scale

are least critical, the slope still increases but on the lowest level of all slopes. In case of unfamiliar participants, there is only a dependence on the factor *delay* visible but none on the *familiarity* or on the *group size*.

In contrast in experiment A.FB.3, all members of the three-party groups were unfamiliar to each other, the perceived inattentiveness of the interlocutors was rated on a continuous scale and only one rather slow interactive *task type*, the 3CTs, was tested. The one factorial ANOVA conducted shows the factor *delay* to be significant but with a rather low effect size ($\eta^2_{partial}$ = 0.092). Looking at the corresponding graph of the A.FB.3 data (Fig. 8.11 (b)), we can see that only the 1200 ms and 1200X ms condition seem to differ in the ratings (but post-hoc are not significant). When comparing them to the ratings of unfamiliar three-party in the A.WB data on the left side (a), the ratings are generally lower. This might be due to the continuous scale that was used for this experiment or other experimental design factors.

For the A.FB.2 data assessing the **perceived personality (PERSO)** of the interlocutor, a MANOVA regarding the twelve assessed personality scales with the factors *delay* and *speaker* was conducted. It shows a significant effect for the *delay* and for the *speaker* (see Appendix I Table 11.17) but not for the interaction of the two factors.

Looking at the single effects of the factor *delay*, the scales E1 (friendliness), E4 (activity level), E6 (cheerfulness), C1 (self-efficacy), C4 (achievement-striving) and C5 (self-discipline) are found to be rated significantly lower in the 1200 ms compared to the no delay condition (confirmed by post-hoc tests with Bonferroni adjustment of alpha), as visible in Figure 8.12. Concerning the factor *speaker*, all E scales and C1 are found to be statistically different from each other for the three speakers. Speaker I is, in all cases rated highest, speaker II middle and speaker III is given the lowest ratings. Post-hoc tests with Bonferroni adjustment confirm all

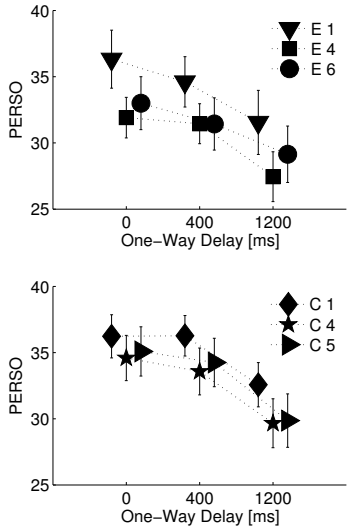


Fig. 8.12: Mean perceived personality ratings (PERSO) and corresponding 95 % CIs on the speakers personality in different delay conditions done by participants, only significant scales shown; E1: friendliness, E4: activity level, E6: cheerfulness, C1: self-efficacy, C4: achievement-striving, C5: self-discipline

speakers to be significantly different from each other with regard to the E-scales and speaker I to be different from speaker III regarding C1.

8.5 Extension of the E-Model

In the following a proposal for extending the E-Model using conversation metrics is presented. It goes beyond the extension presented in Chapter 2. To model the data, MOS values were first transformed to empirical R-scale values, $\mathbf{R}_{emp}$, which can be related to the MOS scores using an S-shaped transformation function (ITU-T Rec. G.107, 2011a). For MOS scores obtained in WB experiments, the corresponding mapping WB function was used (ITU-T Rec. G.107.1, 2011b). As a second step, R-values predicted by the E-Model and WB E-Model, $\mathbf{R}_{E-Model}$, for the tested delay conditions were calculated to be able to obtain the estimated impairment value, **Idd'**, for the empirical data. This was done similar to Raake et al (2013), following the equation:

$$\mathbf{Idd'} = \mathbf{R}_{E-Model} + \mathbf{Idd}_{E-Model} - \mathbf{R}_{emp} \tag{8.1}$$

The initial idea behind this equation was to take possible smaller degradations such as talker echo and line noises into account. The sum of $\mathbf{R}_{E-Model}$ and $\mathbf{Idd}_{E-Model}$ represents the predicted quality corresponding to the NB or WB E-Model when all aspects except the delay impairment are considered. By subtracting $\mathbf{R}_{emp}$ a vector of E-model specific **Idd'** values is obtained.

However, in the experimental investigations the loudness set in the devices was reduced to minimise any remaining possibility for echo. Therefore the default RLR and SLR settings needed to be changed to RLR = 8.8 and SLR = 7 in the E-model prediction. Resulting from this, the maximum achievable $\mathbf{R}_{E-Model}$ value with the adapted loudness settings would have been 84.68 (for 100 ms of delay) in the NB case and 112.12 (for 75 ms of delay) in the WB case.

At the same time, participants had rated the CQ particularly high (>= 4.5 MOS points) when conducting a RNT task with low delay. Most likely this was due to an increased happiness about the possibility to interact very fast in these short delay conditions and thus being closer to win the prize promised to the best team. When transforming these high CQ ratings to **R**-values they took the maximum **R** of 100 for the NB data and 129 for the WB data.

To still be able to follow the proposed approach without receiving negative values for **Idd'** the setting for the *noise floor at the receiver side (Nfor)*, was changed to $Nfor$ = -96 dBm0p. This approach was considered to be reasonable because background noises were not in the focus of participants' attention due to the quite listening lab environment. It only caused a shift of the graph to positive values.

In the WB case the default value for $Nfor$ is -96dBm0p, further enhancement of the value did not increase the $\mathbf{R}_{E-Model}$ substantially and calculated **Idd'** for RNTs were unfortunately still negative. For this reason a further step was taken and Ro, the basic signal-to-noise ratio, was set to the maximum of 129 assuming that the quality regarding the basic signal-to-noise ratio was optimal. This allowed to still include other smaller effects into the calculation.

For the actual modelling, the E-Model formula has been changed as reported earlier (see equation -8.3), corresponding to Raake et al (2013). Three different models were tested, for model #1 the fixed threshold of 100 ms was replaced by the parameter mT considered to represent the minimum perceivable delay and the exponent 6 was multiplied by the new parameter sT indicating the delay sensitivity. In model #2 mT was lowered to 50 and sT was fitted to the data. For model #3, mT was kept constant as in the original version of the E-Model at the value of 100 and only sT was fitted. The so gathered fitting values for model #2 and #3 are listed in Table 8.1. Results for model #1 can be found in Appendix II.

For Ta $<= mT$:

$$Idd = 0 \tag{8.2}$$

For Ta $> mT$:

$$Idd = 25\left\{\left(1+X^{6\cdot sT}\right)^{\frac{1}{6\cdot sT}} - 3\left(1+\left[\frac{X}{3}\right]^{6\cdot sT}\right)^{\frac{1}{6\cdot sT}} + 2\right\} \tag{8.3}$$

$$with: X = \frac{log_{10}\left(\frac{Ta}{mT}\right)}{log_{10}2} \tag{8.4}$$

The results were surprising in respect to the parameter mT. Model #2 which used a lower fixed mT value than model #3 shows better fitting results for the SCT task. For model #1, where mT represented a free parameter as well, corresponding mT values were actually lower in cases where the MOS ratings did not look very critical.

The reason behind the low mT values is the following: **Idd'** values for small delays ($<=$ 400 ms) were comparably high especially for the SCT task type because corresponding MOS scores were rather low (mean usually below 4 MOS points). As a consequence, if mT was a free parameter the fitting function crossed the x-axis before the lowest delay value tested and then increased quickly to fit the first **Idd'** value as well as possible.

For example, in experiment A.NB.3 the mean MOS for 3SCTs was 3.9 at a delay of 100 ms which is rather low for this comparably good condition. As a consequence, the **Idd'** value was rather high with approx. 29. Model #1 calculated a mT value lower 0.00 and an sT value of 0.20 in this case. It is very unlikely that participants were extraordinarily good in picking up the delay impairment here. Thus, the interpretation of "minimal perceivable delay" for the parameter mT is very questionable. Rather mT seems be somehow related to sT, the delay sensitivity.

To keep things easy to understand, it was decided to focus on the two models #2 and #3 using fixed mT values because these models performed better than the previous E-Model as well. The main criterion for judging upon the performance of a model was the Root Mean Square Error (RMSE), but the Pearson correlation coefficient ρ was calculated in addition.

When looking at the RMSE values, as listed in Table 8.1, model #3 was performing better than the E-Model in all cases. Model #2 was better in terms of RMSE in eight out of twelve comparisons. The four cases where model #2 performed worse than the E-Model was in case of RNT task type data (RNTs in A.NB.2b, A.WB.2 and A.WB.3). The value of 100 for mT and a value slightly below 1 for sT represented the best fit and therefore model #3 was performing slightly better than the E-Model and model #2 for RNTs. Looking at the ρ values of RNTs, #2 performed worst for all cases. The ρ values of model #3 were similar to the ones of the E-Model in five cases and better in one case (A.WB.2).

For the SCT task type, model #2 showed better RMSE values than model #3 while both performed better than the E-Model. Regarding ρ the picture was not as clear. Three #2 ρ values of SCTs were lower than the corresponding ρ values of the E-Model. In these cases, the #3 ρ values were, however, slightly better, or in one case similarly good, as the E-Model.

The calculated sT values for the RNT task type are showing that the RNT task type is clearly more delay sensitive than the SCT task type (all RNT sT values are higher than the SCT sT equivalent), supporting Hypothesis Ia.

For the SCT task type and NB, higher sT values were found for the RNT-SCT task-block order in case of two-party interaction (supporting Hypothesis Ib) than for the reverse task-block order. This effect was not found for the three-party data as reported above in Section 8.2.

In general, the delay sensitivity sT was, in most cases, higher for the WB data than for the NB equivalent. This is particularly surprising when looking back to the MOS ratings (Fig. 8.2) where the WB experiments showed clearly less critical ratings especially for high delay values compared to the NB experiments (Fig. 8.1). Some adjustment of the WB E-Model might be necessary here to better capture the effects of delay.

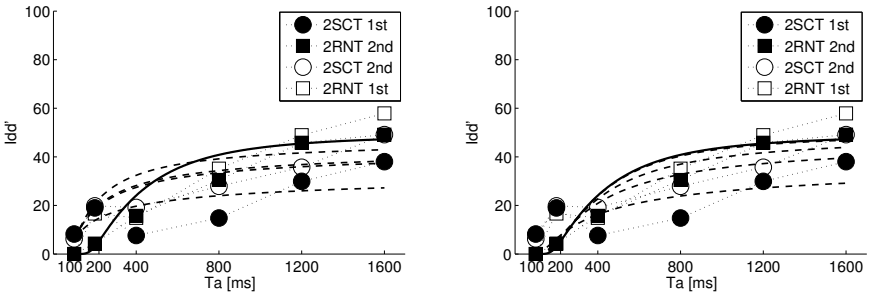
Due to the reported issues with the parameter mT, there will be no recommendations given on this parameter. The interpretation of "minimal perceivable delay" should not be used in the future. Furthermore, it should be checked if similar problems occur for other conversation tasks using a block design and thus having less critical SCT MOS ratings (as for a randomised task-block order, compare Raake et al, 2013).

Regarding the delay sensitivity parameter sT, it was found to maximally achieve 0.55 for the SCT task type but for most NB connections sT was indeed lower (approx. 0.36). Looking at the RNT task type and model #3 results, sT was the highest in case of the two-party WB (2.44) and the lowest in case of three-party WB (0.64). This is to some extend in line with the RNT MOS ratings which were more critical in case of two-party than for three-party especially for delay $>$ 400 ms. On the other hand, the WB MOS ratings were not at all as critical as the NB MOS ratings (see Fig. 8.1 and 8.2). The sT parameter seems to be highly influenced by the model adjustment for WB. In case of NB, the maximum RNT sT value achieved was 1.14 for the SCT-RNT task-block order. For the reverse order, where RNTs were conducted first and were thus not influence by prior task the maximum sT was 0.95. When looking at the corresponding graph in Figure 10.1 (d), it can be seen that the 3RNT 2nd data exceeds the E-Model prediction for 1200 ms but falls below it again for 1600 ms. This is most likely the reason for a $sT > 1$ and can be ascribed to a test specific variation. Thus, a value of 1 for sT can be regarded as reasonable particularly if no prior sensitisation has taken place.

Finally the question arose, if the parameter sT can be related to an automatically accessible conversation measure. If this is the case, it would be possible to predict the delay sensitivity of a particular conversation based on its recordings. The fine tuning of the E-Model prediction would then be possible with the help of conversation analysis, enabling a monitoring of quality including conversation behaviour. From the above discussion of the results, the metric ΔURY seemed to be most related to the MOS ratings because the mean values increase with increasing delay. This increase is big in case of RNTs and much smaller for SCTs. These two aspects

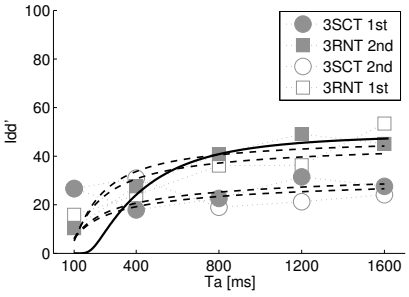
Table 8.1: Curve fitting results for model #2 and #3, mT values were preset and sT values were fitted

Band-width	**Group Size**	**Task Type**	**Order**	**Model**	***mT* [ms]**	***sT***	ρ	**RMSE**
NB	2	SCT	SCT-RNT	EM	100	1	0.66	16.31
				#2	50	0.29	0.69	8.10
				#3	100	0.35	0.70	8.86
		RNT		EM	100	1	0.97	5.51
				#2	50	0.42	0.92	10.10
				#3	100	0.71	0.97	4.75
		SCT	RNT-SCT	EM	100	1	0.87	10.10
				#2	50	0.41	0.88	6.34
				#3	100	0.54	0.89	7.86
		RNT		EM	100	1	0.92	8.45
				#2	50	0.53	0.89	10.37
				#3	100	0.95	0.92	8.44
	3	SCT	SCT-RNT	EM	100	1	0.27	18.23
				#2	50	0.30	0.15	9.51
				#3	100	0.36	0.23	12.11
		RNT		EM	100	1	0.99	5.25
				#2	50	0.57	0.95	4.55
				#3	100	1.14	0.99	5.19
		SCT	RNT-SCT	EM	100	1	0.00	21.70
				#2	50	0.29	0.00	11.00
				#3	100	0.32	0.00	14.06
		RNT		EM	100	1	0.89	8.91
				#2	50	0.48	0.83	7.92
				#3	100	0.77	0.89	8.61
WB	2	SCT	SCT-RNT	EM	100	1	0.83	13.35
				#2	50	0.39	0.72	9.90
				#3	100	0.53	0.83	11.88
		RNT		EM	100	1	0.99	5.81
				#2	50	0.70	0.93	9.18
				#3	100	2.44	0.99	4.27
	3	SCT	SCT-RNT	EM	100	1	0.57	14.56
				#2	50	0.41	0.50	10.01
				#3	100	0.55	0.58	13.28
		RNT		EM	100	1	0.92	8.56
				#2	50	0.42	0.90	11.33
				#3	100	0.64	0.93	7.41

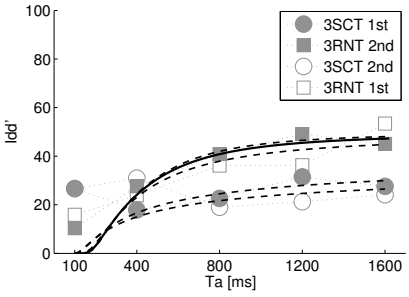


(a) A.NB.2, #2

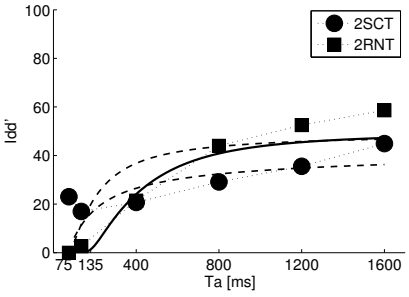
(b) A.NB.2, #3



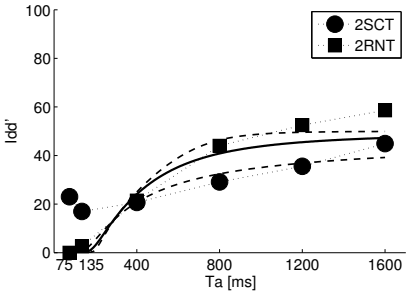
(c) A.NB.3, #2



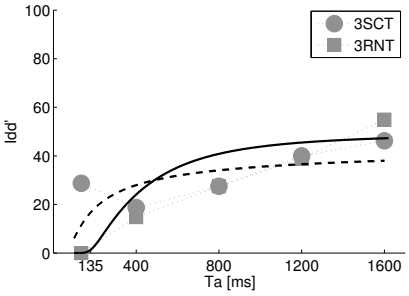
(d) A.NB.3, #3



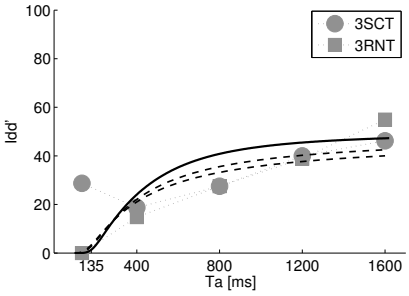
(e) A.WB.2, #2



(f) A.WB.2, #3



(g) A.WB.3, #2



(h) A.WB.3, #3

Fig. 8.13: Curve fitting results; thick black line representing the Idd prediction of the current NB/WB E-Model, and dashed lines representing the fitted models; left column model #2, right column model #3; both models with one free parameter, sT

are in line with the general trends observed for the MOS ratings. Besides, ΔURY is easy to extract without complex definitions of state walks, such as in the case of the metric divergence, and only in need of mono talk spurts. This makes it possible to calculate ΔURY at each individual side for the different interlocutors by recoding the microphone input. The timing of the other interlocutors' speech is not directly of relevance for this parameter which makes it even easier to assess the metric. For these reasons, ΔURY was selected as the conversation metric that should be related to the parameter sT. In a first step for finding the function that relates the two variables to each other, the ΔURY gradients were calculated based on Figure 8.5 (b) and (c). The ΔURY gradient was then plotted over the corresponding sT values that were obtained from the model #3. As explained earlier, #3 performed better than the E-Model in terms of RMSE in all cases. The resulting graph can be seen in Figure 8.14. There seems to be one outlier for the 2RNT A.WB data because the corresponding sT value is especially high with 2.44. As discussed in section 8.3.1, the divergence for RNTs was surprisingly increasing with delay in this data set. This may also indicate that participants had many communication problems. Since there was a fast interaction required and participants most likely had many communication problems due to diverging realities in this experiment they were probably particularly delay sensitive.

Regarding the fitting, the quadratic function fitted to the data performs slightly better with a value of 1.55 for the norm of the residuals than the linear model with a value of 1.62 for the norm of the residuals. However, the linear function might be preferred because it is more simple. Only the data of further investigations can justify whether the linear or rather the quadratic relationship is preferable due to the rather small number of data points here.

Linear Fitting Function:

$$sT = 411.18 * \Delta URY_{gradient} + 0.44058 \tag{8.5}$$

Quadratic Fitting Function:

$$sT = -536800 * \Delta URY^2_{gradient} + 1371.3 * \Delta URY_{gradient} + 0.29159 \tag{8.6}$$

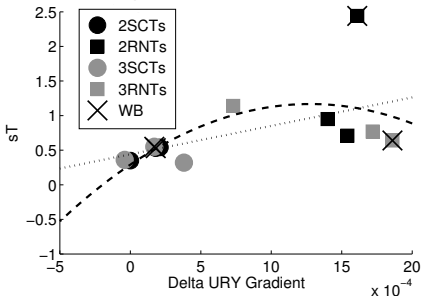


Fig. 8.14: Relationship of the parameter sT obtained from model #3 to ΔURY gradient; Curve fitting: dashed line quadratic model, dotted line linear model

Chapter 9
Discussion

9.1 Hypotheses Ia-d: Conversational Quality

On the basis of the results, it can be confirmed that pure one-way transmission delay is dependant on ratings of Conversational Quality (CQ) assessed in terms of Mean Opinion Score (MOS) (Hypothesis Ia). The direction of the relationship is as expected, the rated MOS is lower the higher the inserted transmission delays are. Delay is found to be the most important factor according to the effect sizes, except for the two experiments in FB bandwidth.

There are different explanations why the effect is not found in the A.FB data. On one hand side, there may be an underlying bandwidth effect. However, due to the design changes between experiments the influence of bandwidth could not be tested statistically. Graphical inspections reveal a tendency towards less critical MOS judgments for high bandwidths. Furthermore, the tested delay conditions reached up to 1200 ms only in both FB experiments and no fast speed, RNT type, task was tested which may also explain why the factor delay was less important in these experiments. Nevertheless, when comparing the A.NB to the A.WB data a difference of MOS ratings can also be detected when comparing the graphs and a bandwidth effect may be a possible explanation. Especially for RNT tasks and high delay values participants rated the quality far less critical. Unfortunately, the experimental design was changed from of A.NB to A.WB experiments as well. In A.WB experiments there was for example an additional rating on the perceived fluency which is why a distinct reason for less critical ratings in WB cannot clearly be identified.

The influence of the task type was also investigated in the present study (Hypothesis Ib) and a difference between SCT tasks and RNT task can be confirmed for experiments in NB bandwidth. There is no significant difference for the RNV task type which is in line with the findings by Egger et al (2010) and Hammer (2006). The RNT tasks is critically different from the RNV task because people are intrinsically motivated to interact quickly by a prize promised to the fastest and most correct team. The difference between RNVs and RNTs also shows that it is not only

the shortness and the possible number of turn switches per time interval that determines the interaction speed but more importantly the speed that is targeted by the interlocutors.

There is no task effect found in the A.WB data. However, the interaction effect of task type and delay turns out to be significant. While there is a task related difference of the MOS ratings in the two-party experiment (A.WB.2) there is no such task related difference in the three-party experiment (A.WB.3) for ratings equal to or above 400 ms. Beyond 400 ms of delay, a decrease related to delay can be detected for both task types and both experiments.

From the literature review, the hypothesis was developed that prior sensitisation to the delay impairment may influence the CQ (Hypothesis Ic). The hypothesis can be supported by different results. Firstly, the results of experiment A.NB.2a show rather critical SCT MOS ratings which can be ascribed to the randomised order of the different tasks types. Secondly, in following experiments the tasks were tested in blocks and the order of the blocks was varied. If participants experienced the RNT task block first, it was assumed that they would be more aware of the impairment for the second, SCT task block, in comparison to the reverse task-block order (SCT-RNT). The results show that this is true for the two-party NB experiment (A.NB.2) but not for the three-party NB experiment (A.NB.3). A more detailed picture can be taken from the graphs. While the 2SCT MOS values are lower in the RNT-SCT than in the SCT-RNT order, the 3SCTs MOS values stay on a rather similar quality level in both orders. This result cannot be explained by a generally lower CQ sensitivity due to the group size because the 3RNT MOS values decrease similarly (with a certain variance) as the 2RNT in both orders.

The fourth hypothesis, Hypothesis Id, states that the bigger the group the less critical people are regarding their quality judgment. It is expected that the conversation itself takes up more cognitive effort for interactions with more than two interlocutors. Thus, less cognitive resources are available for the quality evaluation. The results show that the picture is more complex. For the A.FB data set, a significant effect for the group size can be found. However, the two-party ratings are lower than the three-party ratings which is contrary to what was expected. This could be due to the setup of the A.FB.2 experiment (anechoic chamber) which included additional background noises and may have caused lower CQ ratings. For this reason, this outcome is not helpful to answer the hypothesis.

In the A.NB and the A.WB data only the interaction effects of group size and delay are significant. Three-party groups tend to rate the conversational quality lower for low delay values (up to 400 ms). This is particularly evident for the 2SCTs and 3SCTs. When comparing the CQ ratings of both group sizes for higher delays they do not seem to follow a particular pattern. To fully investigate the hypothesis, it should be investigated further how meaningful this difference for low delay values actually is.

Particular conditions were tested in two experiments. In A.FB.3, asymmetrical delay was implemented and tested in three-party groups. The statistical analysis shows no difference in the MOS ratings regardless of whether there was symmetrical, asymmetrical or no delay at all. Except the aforementioned points addressing both FB experiments, two points particular to experiment A.FB.3 may explain the outcome. Participants in this experiment conducted the 3CTs which took about 7-9 min per call. The scenarios were probably too complex for evaluating the quality at the same time. At least in the A.WB.3 experiment the shorter version of the 3CTs namely the 3SCTs lead to an influence of delay on MOS ratings. The 3SCTs should be investigated in FB as well in the future to see whether the difference can be ascribed to the task format. Another explanation for the outcome may be that only experienced participants were recruited for the A.FB.3 experiment. Unfortunately, the prior experience of the participants was of poor quality and thus their internal reference was low. When being confronted with a FB conference system without any directly perceivable impairments, such as packet loss, echo, or loudness problems participants were convinced that the system was working well in comparison to what they were used to.

In the audio-video experiment, V.WB.2, conditions with asynchronous audio-video delay were investigated. The delay condition is found to be a significant factor in this experiment in contrast to Berndtsson et al (2012) study. Though, it has to be noted that in Berndtsson et al (2012) study both channels were always delayed, only for some conditions one channel was delayed less. In contrast, in experiment V.WB.2, there was no delay added to one of the channels at all. Furthermore, the task of V.WB.2 motivated participants to interact fast and only goups of two conducted the experiment while in Berndtsson et al (2012) groups of four to six interlocutors were tested and free conversations or quiz games (without time pressure) were used.

The results of experiment V.WB.2 show that the CQ is rated best if no delay is inserted (reference condition). When comparing the 800 ms synchronous and asynchronous, audio-only, delay condition to each other, the asynchronous delay is rated worse than the same level of synchronous delay which confirms the findings of Hayashi et al (2007). It is not confirmed for 400 ms synchronous versus asynchronous (audio-only or video-only) delay. The two synchronous delay conditions are not rated with much difference although the one-way delay time was doubled from the lower (400 ms) to the higher condition (800 ms). Maybe asynchrony is easier to be identified as a technical impairment, whereas synchronous delays can also be misinterpreted as based on the conversation behaviour of the interlocutors. Thus, they are not rated as critical. The randomised order of the conditions may have supported a lower sensitivity for the synchronous delay conditions.

9.2 Extension of the E-Model

The present work extents the NB and WB-E-model as proposed by Raake et al (2013) including two new parameters. In model #1 both of these new parameters

were fitted to the data of the four experiments A.NB.2 and .3 and A.WB.2 and .3. As it turned out, the parameter *mT* described as the indicator for the "minimum perceivable delay" by Raake et al (2013) can not be interpreted in this way for the modelled data. The modelling of the second parameter *sT* interpreted as the delay sensitivity seems to be related to the results on *mT*. For this reason, the modelling focuses on the second and third model tested, where *mT* is kept constant (50 for model #2 and 100 for #3, as in the current E-Model) and only *sT* is fitted.

Model #2 turns out to be the best for the SCT task type while model #3 also performs better than the E-Model in terms of RMSE. The ρ values suggest a slight preference for model #3. For the RNT task type #3 performs better or similarly good as the E-Model regarding RMSE and ρ values.

The results furthermore show that the delay sensitivity increases only up to 0.55 for the SCT task. Regarding the RNT task and NB, the maximum value for *sT* does not exceed 1.14, if there is a possible influence by a proceeding SCT task block (SCT-RNT task-block order), and 0.95 if no such influence is present (SCT-RNT task-block order). For the RNT task in the A.WB.2 experiment, *sT* is found to be particularly high. This may be related to a general tendency observed, *sT* values of WB experiments are in any case higher than in the equivalent NB experiment. It may be necessary to adapt the WB E-Model here, because the MOS ratings are actually not very critical. In a second step, the modelling results of the parameter *sT* were related to the gradient of the nonverbal nonvocal conversation metric ΔURY. The found linear functions shows a slight increase of *sT* the higher the ΔURY values are while a quadratic fitting shows a small decrease of *sT* for higher ΔURY values. The found equations should be verified with more data of tests investigating different group sizes and other task types to clarify whether the relationship is rather linear or quadratic.

9.3 Hypothesis IIa: Mediated Interaction Metrics

9.3.1 Nonverbal and Vocal Conversation Metrics

The results of the nonverbal vocal conversation metrics reveal that all three factors, task type, group size and delay determine the conversation structure. The task type and the group size are found to be more important factors than the delay in contrast to the MOS results. The task type constitutes the most important factor, which is not surprising, because the tasks were designed to show a differing conversation structure. It is interesting to confirm that conversations also change with the number of interlocutors per group, even if the concept of the task is kept similar, e.g. 2RNTs versus 3RNTs.

In the literature, a reversed u-shaped development of double talk with increasing delay was reported (Braun, 2003). A similar shape can be observed for the 2SCT

tasks. Up to 400 ms P_{multi}, the probability of multi talk increases but stays constant or even slightly drops again for higher delays. It may be critical here to which extent participants identify the technical origin of the problem. This may determine if they change their interaction strategy leading to a drop of multi talk for delays higher than 400 ms, as it is observed for experiment A.NB.2a, or if they stick to the organisation of the turn-taking as it is observed in all other experiments. For the three-party version of the SCTs P_{multi} simply increases with increasing delay. It seems to have the same effect on the increase of P_{multi} if only one person is delayed (asymmetrical conditions) or all (symmetrical conditions). When comparing synchronous to asynchronous delay conditions, it is interesting to note that P_{multi} increases similarly for video-only delayed conditions as for audio-only or synchronous conditions. Even if this increase is small, it may indicate that people in general are distracted or confused by the asynchrony and thus tend to get problems in managing the turn-taking.

The P_{multi} for the RNT task is normally rather low due to the clear predefined structure and the instructions given, saying that short cuts (and thus multi talk) is not allowed. P_{multi} is sometimes higher for low delayed RNTs. This may be due to longer overall durations of the calls.

Usually it is difficult to determine which degree of P_{multi} is adequate and comfortable for a particular type of conversation and in how far conversations can cope with increased P_{multi}. To find out more about this, participants could have been asked about the acceptability of the amount of double and triple talk. However, the so created awareness for it may change the natural occurrence of it. Regarding the relationship to the CQ ratings, their seems to be no linear relationship for P_{multi} but people may adapt their interaction behaviour if they detect the technical impairment causing less multi talk.

The SARc, Speaker Alternation Rate corrected, can clearly be described as a task differentiating metric. SARc values for SCT tasks are only about one quarter to one third as high as the SARc values of the RNT task. This makes sense because the RNT task is designed to have many short speaker turn changes.

In contrast to Egger et al (2010) and Hammer et al (2004), who looked at the SAR, the now corrected version, the SARc, stays constant with increasing delay and close-to-natural scenarios (SCTs and CTs). Thus, the found decrease in speaker turns per minute in earlier studies was only due the delay causing longer response times.

The SARc values for RNTs have a tendency to decrease with increasing delay, which means that people actually reacted slower, even more than it was necessary due to the delay. The SARc is a good candidate to determine the initial speed of particular type of conversation. However, additional deceleration more than necessary due to delay needs to be taken into account. For determining the initial speed of a conversation the SARc values of two different delay levels, probably one below and one above 400 ms, could be used.

ΔURY, the increase in Utterance RhYthm, seems to be negatively related to the MOS ratings for the RNT task type. While the MOS ratings decrease, ΔURY in-

creases with increasing delay. The frequency of utterances (number and responses) needs to be adapted to longer response times in case of RNTs under delay. For SCTs in turn, there is no need to wait for the utterance of the other (no predefined turn structure). It is interesting to see that people conducting close-to-natural conversations also do not lengthen their time until they utter again, even though delays are present and likely to cause difficulties. Even when participants may realise that delays are distorting the conversations, for instance for very high delay values, it seems to be difficult for them to predict when their speech will arrive at the other side and what effect it may have for their interaction.

The metric divergence, quantifying the degree of differently perceived conversational realities, evolves quite differently for two task types. The proportion of mismatching state walks to matching ones clearly increases with delay for all SCTs and CTs. The 2SCT show more divergence for higher delay values than the 3SCT. Only in the A.NB.2b experiment the rise is not as high. In contrast, the 3CT lead to a growth of divergence about as high as for the 2SCTs. For the audio-video experiment a increase for audio-delay can be observed but not for video-only delay.

In contrast, the divergence is constant across experienced conditions while conducting the RNT task. This means that both participants perceive the same conversational course which is what instructions intended. In the A.WB data, however, an increase of the divergence can be detected being even much higher than the increase observed for the SCTs. Participants obviously did not follow the instructions to wait for the response of the other before taking the turn causing diverging realities and answered very fast on the last syllable of a number. This effect is not necessarily reflected in the SARc values because the number of turns stay similar only the kind of speaker change differs in the two realities (successful interruption at the near-end and alternating silence at the far-end). It is surprising that there seems to be not increase in P_{multi} values for RNT tasks in the A.WB either. Probably, the increase of double talk time due to short-cuts is rather small compared to the increase of the overall duration for higher delays. Thus, in this case it cannot be detected in P_{multi} values that participants did not follow the instructions correctly.

The divergence can also be considered as a measure, which seems to be rather unrelated to the MOS ratings but gives information about how different the conversational realities are and, thus, how complicated it is to keep up a smooth turn-taking system.

9.3.2 Nonverbal and Nonvocal Conversation Metrics

The results of the overall motion (OM) analysis reveal that a recording resolution of 320x240 pixel (two-person video) is necessary for the motion estimation algorithm to work appropriately. In lower resolution (160x240) the algorithm most likely mainly estimates sub-pixel motion which can create too much noise compared to the amount of motion within the pictures. For the higher resolution recordings, it is

found that the motion is highest for synchronous delay. At the same time, there is surprisingly no difference between 400 ms and 800 ms of synchronous delay which is in line with the rated CQ. It would be interesting to see what happens to the detected OM for lower or much higher delay values. It is possible that OM increases in an logarithmical way and saturates at about 400 ms of delay. If such a relationship could be found, the increase of OM may show impatience, moving and shifting the attention away from the screen to bride longer response times.

The OM increases slightly for video-only delay and audio-only delay of 800 ms but stays on a similar level as the reference condition for 400 ms audio-delay. Waiting times were probably not experienced as long in the asynchronous conditions because the information of one channel was delivered in time. The audio channel information seems to be more important in this respect or, in other words, lagging video information seems to be worse. This is in line with tendencies known from acceptability threshold measurements (see e.g. ITU-T RB.1459 1998).

Apart from the OM, the face motion is considered to be interesting because many nonverbal nonvocal behaviours such as head nods or facial expression (kinetics) are expressed by the face. Since the recording resolution for the face area was low, even in the higher resolution case, changes in the facial expression would not have been detected with the algorithm used. Future work could examine the face motion and whether this is the crucial motion to be analysed.

9.3.3 Verbal and Vocal Conversation Metrics

The analysis of the verbal-based measurements show differences between the rather short (2-4 min) and pre-structured CNG task of the V.WB.2 experiment and the more natural and longer (7-9 min) 3CT conversations in the A.FB.3 experiment. Due to the difference in length, the mean numbers of occurrences are found to be much lower in the CNG task. As a further result of the short length and maybe also the pre-defined structure of the CNG conversations, only three scales show significant changes. Contact-indications decrease and partner-induced-stops increase significantly for 400 ms of (synchronous) delay, while no change for 800 ms delay is found. Auto-negative Feedback, on the other hand, increases significantly only for 800 ms of (synchronous) delay. Even though these outcomes are not very strong, it is worth to point out that even with the rather pre-defined structure changes in the verbal communication concerning turn-taking organisation can be observed.

In the A.FB.3 experiment, the outcome is much more clear. All metrics are significantly increasing with increasing (symmetrical) delay. Most metrics start at about zero to one mean number of occurrences for the reference condition. Only three of them are occurring more often already in the reference condition but still increase in their number with delay (partner-induced-stops, own- and partner-communication-management). Highest increases can be found for partner-communication-management, auto-positive Feedback, own-communication management, partner-induced-stop, failed turn claims, contact-indications, auto-repetition

and allo-positive Feedback (in decreasing order). In all of these cases the values increase at least to three times the number of occurrences with 800 ms delay in comparison to the condition without added delay.

These results confirm that the turn-taking becomes unbalanced on a verbal level and that more verbal exchange on the allocation and arrangement of the turn-taking needs to be spent. Especially for the management of the communication a very high increase can be detected. People seem to try to manage the other persons' communication by, for instance, completing the others' sentences. This could also be interpreted as some kind of impatience by the person who completes the sentence of the other. In general, more verbal mistakes and confusion are uttered by the current speaker him- or herself and corrected (own-communication-management). Furthermore, the person currently holding the floor is interrupted more often successfully (partner-induced-stops) which is in line with findings from the nonverbal vocal analysis showing that successful interruptions and especially unintended interruptions happen more often the higher the transmission delay (Egger et al, 2010; Schoenenberg et al, 2014b).

Related to that, turn claims fail more often. This includes the case where interruptions attempts fail and the case where two people begin to talk at the same time and one person leaves the floor to the other person. A third case falling into this category is if person A is breaking and intends to speak on, person B however starts talking and due to delay both happen to start speaking at the same time when looking from a superior perspective. When person B realises the situation difficulty he or she usually suddenly stops talking again.

Auto-positive feedback, such as "mhm", "yea", "ok", "right" or similar, as a sign of showing attention occur much more often with delay. This could be due to an enhanced need of showing the other the own attention because the communication situation is becoming more problematic. On the other hand, due to more difficult turn hand overs with delay the speaking person may hold the floor longer which makes auto-positive feedback more important. Related to this, people more often indicate that they are following the conversation with expressions such as "I'm here", "Hello" or similar. In particular, in the welcome part at the beginning of a conversation such contact-indications are uttered much more often before the actual conversation begins. Similarly, the increased number of allo-positive feedback can be explained by a higher need for reassurance.

The increase of auto-repetitions can easily be explained with the longer response times that are a result of the delay. If no one is taking up the floor the person who talked so far may fill the gap by saying the same thing in other words, to avoid awkward silence.

9.4 Hypothesis IIb: Perceived Fluency

The perceived fluency ratings are clearly dependent on delay. For delays higher than 400 ms, the RNTs lead to less perceived fluency than the SCTs do. Three-party

groups report less fluent conversations for low delay conditions and more fluent conversations for high delay condition (delay $>$ 800 ms) compared to two-party groups.

9.5 Hypotheses III: Perceived Attributes of the Interaction Partners

The Hypothesis III can not be confirmed regarding the distinction made for different levels of familiarity. For all participants in the A.WB data, no matter if they were familiar or unfamiliar to each other, the perceived inattentiveness of the interlocutor increases. The increase is the highest in two-party familiar groups and lowest in three-party familiar groups. In case of unfamiliar groups there is no such difference and the ratings are located in the middle between ratings of two-party and three-party familiar groups, for delays higher than 400 ms. In the A.FB.3 experiment where a continuous scale was used, the wording of the question slightly differed and participants were unfamiliar to each other, the increase of perceived inattentiveness is not as high in the 1200 ms delay condition. However, in the asymmetrical 1200 ms delay condition, where the rating person was delayed but the interlocutors were able to talk without delay to each other the inattentiveness ratings are as high as in the A.WB unfamiliar groups. This result underlines what kind of misinterpretations delay can cause. The own delayed speech is attributed to the inattentiveness of the others. Even though, the familiarity does not play the expected role the current state of the interlocutor is indeed perceived differently when delay is distorting the conversation.

In the A.FB.2 experiment a next step was taken and it was examined in how far people evaluate interlocutors' personality differently in their first contact when delay is or is not inserted in the telephone connection. It is found that on six of the twelve assessed personality scales from the trait extraversion and conscientiousness people rate their unknown interlocutor lower on the corresponding scale with 1200 ms delay than without delay. In particular, if delay was present the person at the far end is rated as less friendly, less active, less cheerful, less self-efficient, less achievement striving and less self-disciplined. It also mattered who the person at the other end was. One of the fixed interlocutor is on average rated highest on all extraversion scales and regarding self-efficacy while another one is on average rated lowest on all these scales. As shown in Schoenenberg et al (2014b), the speaker effect may be due to the different interaction behaviour of the speakers. The speaker who is rated highest is also, for instance, successfully interrupting the participants turns more often.

It can be concluded that communication problems are likely to be misattributed to attributes of the interlocutors. This effect can be explained by a phenomenon widely known in social psychology - the so called *correspondence bias*. Even though the term is related with different interpretations, they all have in common that people have a tendency to infer on the disposition of another person based on situationally

induced behaviour of this person (Gilbert and Malone, 1995). In the case of delay distorted conversations, the communication behaviour of the interlocutors may adapt to the communication conditions after a while. Here, the explanation of correspondence bias would fully apply. Whether non-intended change of the communication behaviour which was caused by the communication medium or caused by situational factors, can still be explained by the correspondence bias is not initially clear. From the point of view of a person using a communication system with delay, he or she usually cannot distinguish why a particular communication behaviour occurred. In most cases, it is unclear if an interlocutor intended to time a particular utterance as it occurred or if this was a result of technical processing by the communication system. Even when users are somewhat aware of delays distorting the connection it is not directly accessible to them why an utterance is placed at a particular time in the conversation. For this reason, it is likely that people can not free themselves from inferring on the interlocutors attributes (current state or personality) in any case. In this sense, it does not matter whether a conversation test is conducted in a laboratory environment or if a real call is investigated - the tendency to attribute the reason for communication difficulties to dispositional aspects of the interaction partners stays and should be considered.

Chapter 10
Conclusion

When planning a communication network, people are usually only concerned about achieving a reasonable **Conversational Quality (CQ)**. As this work shows, there are more components that should be considered to gather a detailed view on the changes that a user experiences. Five components determining the **Quality of Mediated Conversations (QoMC)** are described. Based on empirical investigations these components are examined for mediated conversations under the influence of pure one-way transmission delay.

Clearly, CQ is one important factor in this context. However, the literature does not show a coherent picture about the development of CQ for the presents of transmission delay. Nevertheless, it could be concluded that the interaction speed of a conversation may play a particular role. Beyond that, prior sensitisation and prior expectations of participants seemed to constitute important factors and were thus examined in the present study. The results revealed that the interaction speed that participants aim to have is essential and not the interaction speed a conversation could have. However, the importance of the motivation aspect compared to the degree of pre-structure of a conversation task should be investigated further in the future. For example, people could be asked to accomplish a close-to-natural conversation under time pressure and the results could be compared to the results of a number verification task under time pressure. Concerning prior sensitisation, the results could confirm an influence of prior sensitisation for two-party groups but not for three-party groups. In a different experiment, participants were recruited to be highly experienced with teleconferencing. As their prior experience was unfortunately rather bad the delay impairment had no effect on their CQ judgments which further support the proposed relevance of prior experience or sensitisation.

Moreover, studies examining delayed audio-only communication with more than two interlocutors were lacking in the literature. In the reported experiments, the group size does not affect the CQ ratings per se. However, it is found that three-party groups tend to rate CQ less critical for low delay conditions (< 400 ms), but on a similar level for higher delay values. This effect should be studied further with a more fine graded delay level selection up to 400 or 600 ms of delay. Besides, groups bigger than three participants should be tested. It could be possible, particularly for

audio-only systems, that at a certain group size turn-allocation becomes too difficult, the consequence being that participants have no available attention for assessing the quality anymore.

Beyond CQ, the component **Mediated Interaction (MI)** with the subcomponents *verbal* and *nonverbal* as well as *vocal* and *nonvocal* interaction was considered to constitute an important factor in the context of mediated conversations. As expected, the corresponding measurements are found to be affected by transmission delays besides their expected and intended dependence on the task type and the group size.

On a verbal level, different measurements assessing contact organisation, the level of communication management, feedback behaviour and unplanned stopping of utterances show that with the presence of delay communication becomes more difficult, more communication misunderstandings arrive, more uncertainty arises and more contact- and turn-management is necessary to maintain a conversations. It would be interesting to gain more insights into how in particular turn-management is adapted over time in longer calls in the future.

On a nonverbal and vocal level we saw that the conversational realities of the different ends diverge in close-to-natural conversations the higher the delay. This implies the courses of conversations in terms of timing perceived at the individual ends fit less and less together the higher the delay. It explains why more turn-management is necessary on a verbal level. The degree of divergence is not necessarily related to CQ scores. In contrast, as long as participants can somehow keep their individual utterance speed it seems to be of low relevance that their communication problems may have a technical origin. Only if people want to interact fast but cannot do so, it becomes clear to them that a technical degradation is responsible for a distorted communication.

The nonverbal and nonvocal subcomponent was investigated in an audio-video experiment. Visual interaction, which served as the measurement for this subcomponent, is found to increase when delay is present. Future studies could examine whether people change their focus of attention, e.g. shift their attention away from the screen, with increasing delay or audio-video asynchrony. By means of detailed annotation, it could be examined whether people become impatient when delay is present or if they try to support their speech with gestures more often. On the vocal level, analysing intonation aspects could help to determine the degree of annoyance or boredom.

Connecting the two components CQ and MI, this work enhanced the objective prediction of CQ by taking up the approach of Raake et al (2013) including a conversation related parameter, the delay sensitivity, into the E-Model equation. One of the developed MI metrics, the Utterance RhYthm change (ΔURY), was related to the delay sensitivity parameter to be able to quantify it for network planing purposes. The Δ URY estimates the degree to which people adapt their utterance speed to the impaired communication conditions. In case of a strong adaption, it is more likely that people also detect the delay impairment and reflect this in their CQ ratings.

In addition to the classical CQ scale, a new scale addressing the MI component was tested. The so-called perceived fluency is clearly dependent on delay. For test situations such as a casual or private use the perceived fluency may better cover the impact of delay than the CQ. Future works need to examined the relationship of the CQ and the perceived fluency scale and whether the perceived fluency can also be assessed alone.

The component interaction **Partners** was proposed as a third important factor for the QoMC. It is well known that people have a tendency to attribute the reason for behaviour shown by interaction partners to their disposition and not to situational reasons. In this context, it means that the timing of utterances is likely to be related to the partners current state or their personality and not to the technical impairment transmission delay. Indeed, the results confirm that participant misinterpret the delay induced timing of speech to the attentiveness of the interaction partners or, if interaction partners were unknown, even to their personality.

The component **Experiencer** can be regarded as the baseline because any assessment is dependesnt on the person that is assessing. In the present study it was tried to select random participants, however, due to the recruitment process there was a certain degree of pre-selection on people who saw the corresponding announcement. For this reason mainly students, young adults or retired people took part in the experiments. To enhance the external validity, the results could be repeated with participants in the age range of 35 to 55 years and with different educational backgrounds. It would also be interesting to repeat the experiments in different languages and with people of different cultural backgrounds.

In this work the component **Circumstances** is kept constant by using laboratory experiments which imply rather controlled circumstances. The frame of interaction is thus artificially created by means of task descriptions and assigned roles within the tasks. It would be highly important to examine the QoMC in real life calls with different contexts and different frames, such as private versus business calls.

Finally, this work reveals different components that add to the Quality of Mediated Conversations under transmission delay beyond the sole assessment of Conversational Quality. It was elaborated in which respect the different components are related to each other and how they can serve as distinct sources of information regarding the Quality of Mediated Conversations. Many new questions arose that need to be answered in the future.

APPENDIX I: ANOVA and MANOVA Tables

In all tables, the factor *task type* is abbreviated with "*task*", the factor *group size* is abbreviated with "*group*" and the factor *task-block order* with "*order*".

Conversational Quality Ratings

Table 10.1: Data A.NB.2a: Full-factorial ANOVA for the factors *task* type and *delay* on the MOS; all main effects but only significant interaction effects included in Table

Factor	***df***	***F***	$p_{emp.}$	$\eta^2_{partial}$
task	2	3.79	0.023	0.012
delay	4	143.86	0.000	0.475
task x delay	8	3.28	0.001	0.040
error	637			

Table 10.2: Data A.NB: ANOVA for the factors *task* typ, *delay*, *group* size and task-block *order* on the MOS; all main effects but only significant interaction effects included in Table

Factor	***df***	***F***	$p_{emp.}$	$\eta^2_{partial}$
task	1	20.54	0.000	0.028
delay	5	64.92	0.000	0.313
group	1	0.33	0.569	0.000
order	1	1.22	0.270	0.002
task x delay	5	9.15	0.000	0.060
group x delay	4	7.17	0.000	0.039
group x order	1	4.63	0.032	0.006
error	712			

Table 10.3: Data A.WB: Full-factorial ANOVA for the factors *task* type, *delay* and *group* size on the MOS; all main effects but only significant interaction effects included in Table

Factor	***df***	***F***	$p_{emp.}$	$\eta^2_{partial}$
task	1	0.23	0.634	0.000
delay	5	41.46	0.000	0.196
group	1	2.38	0.124	0.003
task x delay	5	5.80	0.000	0.033
task x group	1	7.88	0.005	0.009
error	848			

Table 10.4: Data A.FB: Full-factorial ANOVA for the factor *delay* condition and *group* size on the MOS; all main effects but only significant interaction effects included in Table; for A.FB.2 data only ratings by participants included in analysis

Factor	***df***	***F***	$p_{emp.}$	$\eta^2_{partial}$
delay	6	0.33	0.919	0.008
group	1	21.79	0.000	0.084
error	237			

Table 10.5: Data V.WB.2: Full-factorial ANOVA for the factor *delay* condition on the MOS

Factor	***df***	***F***	$p_{emp.}$	$\eta^2_{partial}$
delay	6	18.84	0.000	0.235
error	367			

Interaction Measures

Nonverbal and Vocal Conversation Metrics

Table 10.6: Data A.NB.2a: Full-factorial MANOVA for the factors *task* type and *delay* on P_{multi}, SARc, ΔURY, divergence; all main effects but only significant interaction effects included in Table

Multivariate Tests **Factor**	***Wilks-Lambda***	***(df-hypo,df-error)***	***F***	$p_{emp.}$	$\eta^2_{partial}$
task	0.175	(8, 1206)	209.39	0.000	0.581
delay	0.283	(16, 1843)	58.92	0.000	0.271
task x delay	0.598	(32, 2225)	10.41	0.000	0.121

Inbetween-Factors **Factor**	***DV***	***df***	***F***	$p_{emp.}$	$\eta^2_{partial}$
task	P_{multi}	2	70.35	0.000	0.188
	SARc	2	1011.02	0.000	0.769
	ΔURY	2	201.33	0.000	0.399
	divergence	2	7.92	0.000	0.025
delay	P_{multi}	4	16.80	0.000	0.100
	SARc	4	12.76	0.000	0.078
	ΔURY	4	291.33	0.000	0.658
	divergence	4	3.65	0.006	0.023
task x delay	P_{multi}	8	4.74	0.000	0.059
	SARc	8	5.23	0.000	0.065
	ΔURY	8	39.85	0.000	0.345
	divergence	8	11.54	0.000	0.132
error		606			

Table 10.7: Data A.NB: Full-factorial MANOVA for the factors *task* type, *delay* and *group* size on P_{multi}, SARc, ΔURY, divergence; all main effects but only significant interaction effects included in Table

Multivariate Tests **Factor**	***Wilks-Lambda***	***(df-hypo,df-error)***	***F***	$p_{emp.}$	$\eta^2_{partial}$
task	0.752	(4, 743)	562.31	0.000	0.752
delay	0.583	(20, 2984)	21.79	0.000	0.126
delay	0.680	(4, 743)	87.43	0.000	0.320
task x delay	0.821	(20, 2465)	7.54	0.000	0.048
task x group	0.844	(20, 743)	34.38	0.000	0.156
delay x group	0.840	(16, 2270)	8.34	0.000	0.043
task x delay x group	0.885	(16, 2270)	5.77	0.000	0.030

Inbetween-Factors **Factor**	***DV***	***df***	***F***	$p_{emp.}$	$\eta^2_{partial}$
task	P_{multi}	1	467.47	0.000	0.385
	SARc	1	701.65	0.000	0.485
	ΔURY	1	108.23	0.000	0.127
	divergence	1	4.01	0.046	0.005
delay	P_{multi}	5	3.23	0.000	0.021
	SARc	5	16.73	0.000	0.099
	ΔURY	5	27.15	0.000	0.154
	divergence	5	36.77	0.000	0.198
group	P_{multi}	1	168.73	0.000	0.184
	SARc	1	103.49	0.000	0.122
	ΔURY	1	1.38	0.240	0.002
	divergence	1	9.43	0.002	0.012
task x delay	P_{multi}	5	2.22	0.050	0.015
	SARc	5	17.98	0.000	0.108
	ΔURY	5	17.35	0.000	0.104
	divergence	5	2.63	0.023	0.017
task x group	P_{multi}	1	37.68	0.000	0.048
	SARc	1	26.77	0.000	0.035
	ΔURY	1	11.30	0.001	0.015
	divergence	1	9.72	0.002	0.013
delay x group	P_{multi}	4	6.42	0.000	0.033
	SARc	4	27.08	0.000	0.127
	ΔURY	4	0.72	0.577	0.004
	divergence	4	2.40	0.049	0.013
task x delay x group	P_{multi}	4	3.99	0.003	0.021
	SARc	4	15.72	0.000	0.078
	ΔURY	4	0.96	0.427	0.005
	divergence	4	6.81	0.000	0.035
error		746			

Table 10.8: Data A.WB: Full-factorial MANOVA for the factors *task* type, *delay* and *group* size on P_{multi}, SARc, ΔURY, divergence; all main effects but only significant interaction effects included in Table

Multivariate Tests **Factor**	***Wilks-Lambda***	***(df-hypo,df-error)***	***F***	$p_{emp.}$	$\eta^2_{partial}$
task	0.206	(4, 834)	805.29	0.000	0.794
delay	0.584	(20, 2767)	24.32	0.000	0.126
group	0.261	(4, 834)	590.92	0.000	0.739
task x delay	0.584	(20, 2767)	24.33	0.000	0.126
task x group	0.724	(4, 834)	79.57	0.000	0.276
delay x group	0.904	(16, 2548)	5.35	0.000	0.025

Inbetween-Factors **Factor**	***DV***	***df***	***F***	$p_{emp.}$	$\eta^2_{partial}$
task	P_{multi}	1	381.37	0.000	0.313
	SARc	1	2071.10	0.000	0.712
	ΔURY	1	298.43	0.000	0.263
	divergence	1	117.08	0.000	0.123
delay	P_{multi}	5	4.58	0.000	0.027
	SARc	5	17.12	0.000	0.093
	ΔURY	5	64.71	0.000	0.279
	divergence	5	23.44	0.000	0.123
group	P_{multi}	1	33.967	0.000	0.039
	SARc	1	2172.60	0.000	0.722
	ΔURY	1	3.66	0.056	0.004
	divergence	1	66.12	0.000	0.073
task x delay	P_{multi}	5	20.22	0.000	0.108
	SARc	5	53.44	0.000	0.242
	ΔURY	5	43.97	0.000	0.208
	divergence	5	11.12	0.000	0.062
task x group	P_{multi}	1	92.95	0.000	0.100
	SARc	1	246.76	0.000	0.228
	ΔURY	1	5.53	0.019	0.007
	divergence	1	23.01	0.000	0.027
delay x group	P_{multi}	4	10.41	0.000	0.047
	SARc	4	7.69	0.000	0.035
	ΔURY	4	0.62	0.649	0.003
	divergence	4	2.93	0.020	0.014
error		837			

Table 10.9: Data A.FB: Full-factorial MANOVA for the factor *delay* condition and *group* size on P_{multi}, SARc, ΔURY, divergence; all main effects but only significant interaction effects included in Table

Multivariate Tests **Factor**	***Wilks-Lambda***	***(df-hypo,df-error)***	***F***	$p_{emp.}$	$\eta^2_{partial}$
delay	0.575	(24,1358)	9.72	0.000	0.129
group	0.302	(4,389)	224.82	0.000	0.698
delay x group	0.866	(8, 778)	7.27	0.000	0.070
Inbetween-Factors **Factor**	***DV***	***df***	***F***	$p_{emp.}$	$\eta^2_{partial}$
delay	P_{multi}	6	7.61	0.000	0.104
	SARc	6	3.43	0.003	0.050
	ΔURY	6	0.72	0.635	0.011
	divergence	6	31.77	0.000	0.327
group	P_{multi}	1	65.89	0.000	0.144
	SARc	1	801.47	0.000	0.672
	ΔURY	1	1.83	0.177	0.005
	divergence	1	0.20	0.659	0.000
delay x group	P_{multi}	2	8.78	0.000	0.043
	SARc	2	5.15	0.006	0.026
	ΔURY	2	2.55	0.079	0.013
	divergence	2	1.21	0.300	0.006
error		392			

Table 10.10: Data V.WB.2: MANOVA for the factors *delay* condition on P_{multi}, SARc, ΔURY, divergence

Multivariate Tests **Factor**	***Wilks-Lambda***	***(df-hypo,df-error)***	***F***	$p_{emp.}$	$\eta^2_{partial}$
delay	0.493	(24,1271)	11.91	0.000	0.162
Inbetween-Factors **Factor**	***DV***	***df***	***F***	$p_{emp.}$	$\eta^2_{partial}$
delay	P_{multi}	6	2.22	0.041	0.035
	SARc	6	4.02	0.001	0.062
	ΔURY	6	1.18	0.319	0.019
	divergence	6	24.24	0.000	0.284
error		367			

Nonverbal and Nonvocal Conversation Metrics

Table 10.11: Data V.WB.2: ANOVA for the factor *delay* condition on the OM, for LOW video and HIGH video resolution

Factor	***df***	***F***	$p_{emp.}$	$\eta^2_{partial}$
LOW				
delay	6	0.90	0.497	0.020
error	264			
HIGH				
delay	6	4.31	0.001	0.256
error	75			

Table 10.12: Data A.FB.3: MANOVA for the factor *delay* condition on the annotated measures: contact-check, contact-indication, own-communication-management, partner-communication-management, auto-positive, auto-negative, auto-repitition, auto-summary, allo-positive, allo-negative, self-induced-stop, partner-induced-stop, failed turn claims

Multivariate Tests **Factor**	***Wilks-Lambda***	***(df-hypo,df-error)***	***F***	$p_{emp.}$	$\eta^2_{partial}$
delay	0.084	(26, 110)	10.38	0.000	0.710
Inbetween-Factors **Factor**	***DV***	***df***	***F***	$p_{emp.}$	$\eta^2_{partial}$
delay	contact-check	2	10.12	0.000	0.232
	contact-indication	2	82.85	0.000	0.712
	own-communication-management	2	19.62	0.000	0.369
	partner-communication-management	2	33.37	0.000	0.499
	auto-positive	2	14.68	0.000	0.305
	auto-negative	2	7.61	0.001	0.185
	auto-repetition	2	7.67	0.001	0.186
	auto-summary	2	7.17	0.002	0.176
	allo-positive	2	4.88	0.010	0.127
	allo-negative	2	4.78	0.011	0.125
	self-induced-stop	2	5.75	0.005	0.147
	partner-induced-stop	2	22.82	0.000	0.405
	failed turn claims	2	35.87	0.000	0.517
error		67			

Table 10.13: Data V.WB.2: MANOVA for the factor *delay* condition on the annotated measures: contact-check, contact-indication, own-communication-management, partner-communication-management, auto-positive, auto-negative, auto-repitition, auto-summary, allo-positive, allo-negative, self-induced-stop, partner-induced-stop, failed turn claims

Multivariate Tests **Factor**	***Wilks-Lambda***	***(df-hypo,df-error)***	***F***	$p_{emp.}$	$\eta^2_{partial}$
delay	0.713	(26, 346)	2.45	0.000	0.156

Inbetween-Factors **Factor**	***DV***	***df***	***F***	$p_{emp.}$	$\eta^2_{partial}$
delay	contact-check	2	1.031	0.359	0.011
	contact-indication	2	8.926	0.000	0.088
	own-communication-management	2	0.991	0.373	0.011
	partner-communication-management	2	1.040	0.356	0.011
	auto-positive	2	1.128	0.326	0.012
	auto-negative	2	3.440	0.034	0.036
	auto-repetition	2	0.709	0.493	0.008
	auto-summary	2	0.475	0.622	0.005
	allo-positive	2	1.104	0.334	0.012
	allo-negative	2	1.040	0.356	0.011
	self-induced-stop	2	0.697	0.500	0.007
	partner-induced-stop	2	3.540	0.031	0.037
	failed turn claims	2	2.395	0.094	0.025
error		185			

Perceived Fluency of the Conversation

Table 10.14: Data A.WB: Full-factorial ANOVA for the factors *task* type, *delay* and *group* size on the FLOW; all main effects but only significant interaction effects included in Table

Factor	***df***	***F***	$p_{emp.}$	$\eta^2_{partial}$
task	1	2.86	0.091	0.003
delay	5	66.96	0.000	0.283
group	1	1.92	0.166	0.002
task x delay	5	7.11	0.000	0.040
task x group	1	9.52	0.002	0.011
delay x group	4	2.90	0.021	0.014
error	848			

Perceived Attributes of Interaction Partner Ratings

Table 10.15: Data A.WB: Full-factorial ANOVA with the factors *task* type, *delay*, *group* size and *familiarity* regarding the mean perceived ATT, rated on an absolute category rating scale; all main effects but only significant interaction effects included in Table

Factor	***df***	***F***	$p_{emp.}$	$\eta^2_{partial}$
task	1	0.04	0.837	0.000
delay	5	25.98	0.000	0.096
group	1	0.68	0.409	0.001
familiarity	1	0.84	0.361	0.001
task x delay	5	3.10	0.009	0.013
group x familiarity	1	6.18	0.013	0.005
task x delay x group	4	2.48	0.042	0.008
error	1222			

Table 10.16: Data A.FB.3: ANOVA for the factor *delay* condition on the mean perceived ATT, rated on a continuous scale

Factor	***df***	***F***	$p_{emp.}$	$\eta^2_{partial}$
delay	6	2.22	0.045	0.092
error	131			

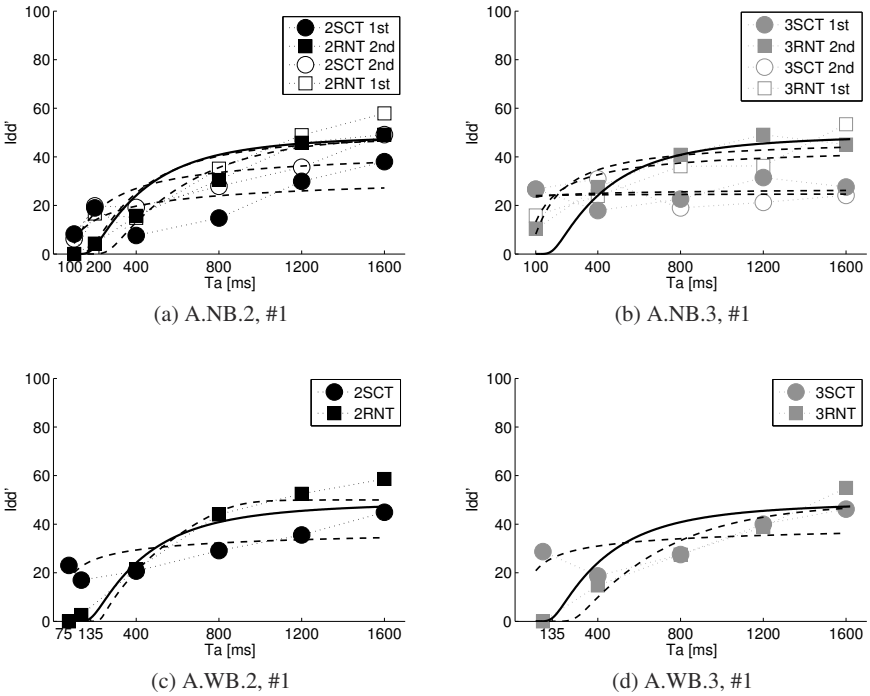
Table 10.17: Data A.FB.2: Full-factorial MANOVA for the factors *delay* condition and *speaker* on the (by the participants) rated PERSO, that are the six facets of extraversion: E1 friendliness, E2 gregariousness, E3 assertiveness, E4 activity level, E5 excitement seeking, E6 cheerfulness, and the six facets of conscientiousness: C1 self-efficacy, C2 orderliness, C3 dutifulness, C4 achievement-striving, C5 self-discipline, C6 cautiousness; all main effects shown below, the interaction effect was not significant and is not included in the Table

Multivariate Tests **Factor**	***Wilks-Lambda***	***(df-hypo,df-error)***	***F***	$p_{emp.}$	$\eta^2_{partial}$
delay	0.709	(24, 224)	1.75	0.019	0.158
speaker	0.570	(24, 224)	3.03	0.000	0.245
Inbetween-Factors **Factor**	***DV***	***df***	***F***	$p_{emp.}$	$\eta^2_{partial}$
delay	E1	2	6.32	0.002	0.093
	E2	2	1.48	0.231	0.024
	E3	2	2.50	0.087	0.039
	E4	2	10.24	0.000	0.143
	E5	2	1.11	0.333	0.018
	E6	2	4.06	0.020	0.062
	C1	2	7.08	0.001	0.103
	C2	2	1.18	0.311	0.019
	C3	2	1.17	0.315	0.019
	C4	2	8.59	0.000	0.123
	C5	2	8.37	0.000	0.120
	C6	2	2.30	0.104	0.036
speaker	E1	2	27.82	0.000	0.311
	E2	2	17.81	0.000	0.225
	E3	2	25.83	0.000	0.296
	E4	2	14.07	0.000	0.186
	E5	2	13.20	0.000	0.177
	E6	2	15.28	0.000	0.199
	C1	2	4.50	0.013	0.068
	C2	2	0.42	0.658	0.007
	C3	2	0.59	0.557	0.009
	C4	2	1.95	0.147	0.031
	C5	2	0.15	0.863	0.002
	C6	2	1.22	0.300	0.019
error		123			

APPENDIX II: Extension of the E-Model - Results for Model With Two Free Parameter

Table 10.18: Curve fitting results for model #1, values should be interpreted with concerns

Band-width	**Group Size**	**Task Type**	**Order**	**Model**	***mT* [ms]**	***sT***	ρ	**RMSE**
NB	2	SCT	SCT-RNT	EM	100	1	0.66	16.32
				#1	49.45	0.29	0.69	8.10
		RNT		EM	100	1	0.97	5.51
				#1	139.97	1.28	0.99	3.16
		SCT	RNT-SCT	EM	100	1	0.87	10.10
				#1	55.61	0.42	0.89	6.27
		RNT		EM	100	1	0.92	8.45
				#1	95.37	0.88	0.91	8.44
	3	SCT	SCT-RNT	EM	100	1	0.27	18.23
				#1	0.06	0.23	0.23	4.51
		RNT		EM	100	1	0.99	5.25
				#1	43.49	0.54	0.95	4.36
		SCT	RNT-SCT	EM	100	1	0.00	21.70
				#1	0.00	0.20	0.00	4.29
		RNT		EM	100	1	0.89	8.91
				#1	34.16	0.44	0.83	7.21
WB	2	SCT	SCT-RNT	EM	100	1	0.83	13.35
				#1	15.80	0.32	0.72	6.64
		RNT		EM	100	1	0.99	5.81
				#1	112.95	4.7	0.99	3.87
	3	SCT	SCT-RNT	EM	100	1	0.57	14.56
				#1	16.41	0.34	0.54	8.22
		RNT		EM	100	1	0.92	8.56
				#1	153.16	1.34	0.96	5.35



(a) A.NB.2, #1

(b) A.NB.3, #1

(c) A.WB.2, #1

(d) A.WB.3, #1

Fig. 10.1: Curve fitting results for model #1 with *mT* and *sT* as free parameter; thick black line representing the Idd prediction of the current NB/WB E-Model, and dashed lines representing the fitted models

APPENDIX III: Informed Consent, Pre-Questionnaire and Ratings

Informed Consent

Telekom Innovation Laboratories

Technische Universität Berlin
Deutsche Telekom Laboratories
Assessment of IP-based Applications
Ernst-Reuter-Platz 7, 10587 Berlin

EINVERSTÄNDNISERKLÄRUNG

Teilnahme an der Studie zur „Telefonieren für die Wissenschaft“

_____
(Vor- und Nachname)

Ich bin ausreichend in mündlicher und schriftlicher Form über die Ziele und Methoden, die möglichen Risiken und den Nutzen der Studie informiert worden. Ich hatte ausreichend Gelegenheit, die Studie mit der Versuchsleitung zu besprechen und Fragen zu stellen. Alle meine Fragen und Bedenken wurden zu meiner Zufriedenheit beantwortet bzw. geklärt.

Ich bin damit einverstanden, dass im Rahmen der Studie persönliche Daten erhoben und meine Gespräche aufgezeichnet werden, und dass diese anonymisiert (d.h. ohne Rückschlussmöglichkeit auf meine Person) gespeichert und ausgewertet werden. Alle im Rahmen der Studie erhobenen Daten werden strikt vertraulich und gemäß dem Datenschutz behandelt. Einer wissenschaftlichen Auswertung der anonymisierten Daten und einer möglichen Veröffentlichung der Studienergebnisse stimme ich zu.

Ich weiß, dass meine Studienteilnahme freiwillig ist und dass ich jederzeit ohne Angabe von Gründen meine Zusage zur Teilnahme zurückziehen kann und mir daraus keine Nachteile entstehen. Insbesondere kann ich zu jedem beliebigen Zeitpunkt die Löschung meiner Daten verlangen. Ich habe ein ID Code genannt bekommen, das mir die Identifikation meiner Daten ermöglicht.

Ich gebe hiermit meine freiwillige Zustimmung zur Teilnahme an dieser Studie.

_____ _____
(Ort, Datum) (Unterschrift)

Pre-Questionnaire

Vorab ein paar kurze Fragen

1. Wie geht es Ihnen gerade so? Bitte setzten Sie ein Kreuz.

ideal sehr schlecht

3	2	1	0	-1	-2	-3

2a. In welchem Land wurden Sie geboren?

_____

2b. Die deutsche Sprache ist ... ? Bitte setzten Sie ein Kreuz.

☐ Ihre Muttersprache
☐ nicht Ihre Muttersprache, aber Sie sprechen fließend Deutsch.
☐ (keins der beiden oberen Optionen) _____

3. Wie alt sind Sie? _____ Jahre

4. Kennen Sie Ihren Gesprächspartner?

☐ Ja ☐ Nein

Ratings for A.NB.2a-b, A.NB.3, A.FB.2, A.V.WB.2

Szenario: [Task Name]

Wie war Ihr persönlicher **Gesamteindruck** von der Verbindung in diesem Gespräch?

Bitte antworten Sie ganz **intuitiv.** Setzten Sie dafür ein Kreuz an die Stelle auf der Skala, die Ihren Gesamteindruck von der **Qualität der Verbindung** am besten beschreibt.

ausgezeichnet	gut	ordentlich	dürftig	schlecht
5	4	3	2	1

Ratings for A.WB.2

Szenario: [Task Name]

Bewertung

Bitte antworten Sie ganz **intuitiv.** Setzten Sie dafür ein Kreuz an die Stelle auf jeder Skala, die Ihren Gesamteindruck am besten beschreibt.

a) Wie war Ihr persönlicher **Gesamteindruck** von der **Qualität** der **Telefonkonferenz-Verbindung** in diesem Gespräch?

ausgezeichnet	Gut	ordentlich	dürftig	schlecht
5	4	3	2	1

b) Wie **aufmerksam** hat sich Ihr **Gesprächspartner** an diesem Gespräch **beteiligt**?

sehr aufmerksam *sehr unaufmerksam*

☐ ☐ ☐ ☐ ☐ ☐ ☐

c) *„**Dieses Gespräch war flüssig.**“*

trifft voll zu — trifft überhaupt nicht zu

O O O O O O O

Ratings for A.WB.3

Szenario: [Task Name]

Bewertung

Bitte antworten Sie ganz **intuitiv.** Setzten Sie dafür ein Kreuz an die Stelle auf jeder Skala, die Ihren Gesamteindruck am besten beschreibt.

a) Wie war Ihr persönlicher **Gesamteindruck** von der **Qualität** der **Telefonkonferenz-Verbindung** in diesem Gespräch?

ausgezeichnet	gut	ordentlich	dürftig	schlecht
5	4	3	2	1

b) Wie **aufmerksam** hat sich der folgende **Gesprächspartner** an diesem Gespräch **beteiligt**?

	sehr aufmerksam						*sehr unaufmerksam*
Person 2 Schubert	☐	☐	☐	☐	☐	☐	☐
Person 3 Manger	☐	☐	☐	☐	☐	☐	☐

c) *„**Dieses Gespräch war flüssig.**“*

trifft voll zu — trifft überhaupt nicht zu

O O O O O O O

Ratings for A.FB.3

Szenario: [Task Name]

Bitte antworten Sie ganz **intuitiv.** Setzten Sie dafür ein Kreuz an die Stelle auf der Skala, die Ihren Eindruck am besten beschreibt.

1. Wie war Ihr persönlicher **Gesamteindruck** von der **Qualität der Verbindung**?

ausgezeichnet	gut	ordentlich	dürftig	schlecht
5	4	3	2	1

2. Wie war Ihr persönlicher **Eindruck** von Ihren **Gesprächspartner in diesem Gespräch**?

a) Wie aufmerksam hat der Gesprächspartner an der Konversation teilgenommen?

Anrufer: **Kirchberg**

sehr angenehm sehr unangenehm

Gesprächspartner 2: **Maibach**

sehr angenehm sehr unangenehm

b) Wie angenehm fanden Sie den Gesprächspartner im Gespräch?

Anrufer: **Kirchberg**

sehr angenehm sehr unangenehm

Gesprächspartner 2: **Maibach**

sehr angenehm sehr unangenehm

Ratings of perceived personality of interlocutor, A.FB.2

Bitte bewerten Sie Ihren Gesprächspartner.

AUSSAGE	STARKE ABLEHNUNG	ABLEHNUNG	NEUTRAL	ZUSTIMMUNG	STARKE ZUSTIMMUNG
ER/SIE FREUNDET SICH SCHNELL MIT LEUTEN AN					
ER/SIE WIRD MIT ANDEREN MENSCHEN SCHNELL WARM.					
ER/SIE FÜHLT SICH IN GESELLSCHAFT ANDERER WOHL.					
ER/SIE KANN GUT MIT ANDERN MENSCHEN UMGEHEN.					
ER/SIE MUNTERT OFT ANDERE LEUTE AUF.					
ES IST SCHWER IHN KENNEN ZU LERNEN.					
ER/SIE FÜHLT SICH OFT UNWOHL IN GESELLSCHAFT.					
ER/SIE VERMEIDET DEN UMGANG MIT ANDEREN.					
ER/SIE INTERESSIERT SICH *NICHT* WIRKLICH FÜR ANDERE MENSCHEN.					
ER/SIE HÄLT ANDERE MENSCHEN AUF DISTANZ.					
ER/SIE LIEBT GROSSE PARTIES.					
AUF PARTIES SPRICHT ER/SIE MIT VIELEN VERSCHIEDENEN LEUTEN.					
ER/SIE GENIESST ES, TEIL EINER GRUPPE ZU SEIN.					
ER/SIE LÄSST ANDERE TEILHABEN, AN DEM WAS ER MACHT.					
ER/SIE LIEBT ÜBERRASCHUNGSPARTIES.					
ER/SIE ZIEHT ES VOR, ALLEIN ZU SEIN.					
ER/SIE MÖCHTE OFT ALLEINE GELASSEN WERDEN.					
ER/SIE MAG *KEINE* ÜBERFÜLLTEN VERANSTALTUNGEN.					
ER/SIE VERMEIDET MENSCHENMENGEN.					
ER/SIE SUCHT OFT DIE RUHE.					
ER/SIE ÜBERNIMMT OFT DIE FÜHRUNG.					
ER/SIE VERSUCHT ANDERE ZU FÜHREN.					
ER/SIE KANN ANDERE GUT ZU ETWAS ÜBERREDEN.					
ER/SIE VERSUCHT, ANDERE ZU BEEINFLUSSEN.					
ER/SIE ÜBERNIMMT GERNE DIE KONTROLLE.					
ER/SIE WARTET DARAUF, DASS ANDERE DIE FÜHRUNG ÜBERNEHMEN.					
ER/SIE HÄLT SICH IM HINTERGRUND.					
ER/SIE HAT WENIG ZU SAGEN.					
ER/SIE ZIEHT *NICHT* GERN AUFMERKSAMKEIT AUF SICH.					
ER/SIE HÄLT SEINE MEINUNG EHER ZURÜCK.					
ER/SIE IST IMMER BESCHÄFTIGT.					
ER/SIE IST IMMER UNTERWEGS.					
ER/SIE UNTERNIMMT VIEL IN SEINER FREIZEIT.					
ER/SIE KANN VIELE DINGE GLEICHZEITIG BEWÄLTIGEN.					
ER/SIE REAGIERT SCHNELL.					
ER/SIE NIMMT ES GERN GELASSEN.					
ER/SIE NIMMT SICH GERNE ZEIT.					
ER/SIE VERFOLGT EINEN GEMÄCHLICHEN LEBENSSTIL.					
ER/SIE LÄSST DEN DINGEN IHREN NATÜRLICHEN LAUF.					
ER/SIE REAGIERT LANGSAM.					
ER/SIE LIEBT AUFREGUNG.					
ER/SIE SUCHT ABENTEUER.					
ER/SIE LIEBT ACTION.					
ER/SIE GENIESST ES, TEIL EINER LAUTEN MENGE ZU SEIN.					
ER/SIE GENIESST ES, WAGHALSIG ZU SEIN.					
ER/SIE VERHÄLT SICH OFT WILD UND VERRÜCKT.					
ER/SIE IST GEWILLT, ALLES MAL ZU PROBIEREN.					
ER/SIE SUCHT DIE GEFAHR.					
ER/SIE WÜRDE NIEMALS DRACHENFLIEGEN ODER BUNGEE- JUMPING GEHEN.					
ER/SIE MAG *KEINE* LAUTE MUSIK.					

Ratings of perceived personality of interlocutor, A.FB.2

Bitte bewerten Sie Ihren Gesprächspartner.

AUSSAGE	STARKE ABLEHNUNG	ABLEHNUNG	NEUTRAL	ZUSTIMMUNG	STARKE ZUSTIMMUNG
ER/SIE STRAHLT FREUDE AUS.					
ER/SIE HAT VIEL SPAß.					
ER/SIE FREUT SICH MANCHMAL WIE EIN KLEINES KIND.					
ER/SIE LACHT SICH DURCHS LEBEN.					
ER/SIE LIEBT DAS LEBEN.					
ER/SIE SIEHT DAS LEBEN VON SEINER SCHOKOLADENSEITE.					
ER/SIE LACHT LAUT.					
ER/SIE ERHEITERT SEINE FREUNDE.					
ER/SIE IST *NICHT* LEICHT ZU ERHEITERN.					
ER/SIE MACHT SELTEN WITZE.					
ER/SIE ERLEDIGT AUFGABEN ERFOLGREICH.					
ER/SIE IST HERVORRAGEND IN DEM, WAS ER TUT.					
ER/SIE ERLEDIGT AUFGABEN REIBUNGSLOS.					
ER/SIE IST SICH SEINER SACHEN SICHER.					
ER/SIE WARTET MIT GUTEN LÖSUNGEN AUF.					
ER/SIE IST GUT DARIN, DINGE ZU ERLEDIGEN.					
ER/SIE SCHÄTZT SITUATIONEN OFTMALS FALSCH EIN.					
ER/SIE VERSTEHT SACHEN OFTMALS *NICHT.*					
ER/SIE HAT OFTMALS WENIG BEIZUTRAGEN.					
ER/SIE DENKT *NICHT* AN DIE KONSEQUENZEN VON DINGEN ODER TATEN.					
ER/SIE MAG ORDNUNG.					
ER/SIE RÄUMT GERNE AUF.					
ER/SIE MÖCHTE, DASS ALLES PERFEKT IST.					
ER/SIE LIEBT ORDNUNG UND REGELMÄßIGKEIT.					
ER/SIE GEHT PLANMÄßIG VOR.					
ER/SIE VERGISST OFT, DINGE WIEDER AN DEN RICHTIGEN PLATZ ZURÜCK ZU BRINGEN.					
ER/SIE HINTERLÄSST UNORDNUNG IN SEINEM ZIMMER.					
ER/SIE LÄSST SEINE SACHEN HERUMLIEGEN.					
ER/SIE FÜHLT SICH VON UNORDENTLICHEN LEUTEN *NICHT* GESTÖRT.					
ER/SIE FÜHLT SICH DURCH UNORDNUNG *NICHT* GESTÖRT.					
ER/SIE VERSUCHT REGELN ZU FOLGEN.					
ER/SIE HÄLT SEINE VERSPRECHEN.					
ER/SIE ZAHLT SEINE RECHNUNGEN PÜNKTLICH.					
ER/SIE SAGT DIE WAHRHEIT.					
ER/SIE HÖRT AUF SEIN GEWISSEN.					
ER/SIE BRICHT REGELN.					
ER/SIE BRICHT SEINE VERSPRECHEN.					
ER/SIE BRINGT ANDERE DAZU SEINE PFLICHTEN ZU ERFÜLLEN.					
ER/SIE TUT DAS GEGENTEIL VON DEM, WAS GEFRAGT IST.					
ER/SIE STELLT DIE FAKTEN FALSCH DAR.					
ER/SIE STEUERT SEIN ZIEL OHNE UMWEGE AN.					
ER/SIE ARBEITET HART.					
ER/SIE WANDELT PLÄNE IN TATEN UM.					
ER/SIE ERLEDIGT AUFGABEN MIT GANZEM HERZEN.					
ER/SIE MACHT MEHR, ALS VON IHM ERWARTET WIRD.					
ER/SIE SETZT FÜR SICH UND ANDERE HOHE STANDARDS.					
ER/SIE VERLANGT QUALITÄT.					
ER/SIE IST *NICHT* HOCH MOTIVIERT ERFOLG ZU HABEN.					

Ratings of perceived personality of interlocutor, A.FB.2

Bitte bewerten Sie Ihren Gesprächspartner.

AUSSAGE	STARKE ABLEHNUNG	ABLEHNUNG	NEUTRAL	ZUSTIMMUNG	STARKE ZUSTIMMUNG
ER/SIE MACHT GERADE GENUG ARBEIT UM DURCHZUKOMMEN.					
ER/SIE STECKT NUR WENIG ZEIT UND AUFWAND IN SEINE ARBEIT.					
ER/SIE ERLEDIGT HAUSARBEIT SOFORT.					
ER/SIE IST IMMER VORBEREITET.					
ER/SIE FÄNGT MIT AUFGABEN SOFORT AN.					
ER/SIE MACHT SICH UMGEHEND AN DIE ARBEIT.					
ER/SIE FÜHRT SEINE PLÄNE AUS.					
ER/SIE FINDET ES SCHWIERIG SICH AN DIE ARBEIT ZU MACHEN.					
ER/SIE VERSCHWENDET ZEIT.					
ER/SIE BRAUCHT DRUCK UM ANZUFANGEN.					
ER/SIE HAT SCHWIERIGKEITEN AUFGABEN ANZUFANGEN.					
ER/SIE VERSCHIEBT ENTSCHEIDUNGEN.					
ER/SIE VERMEIDET FEHLER.					
ER/SIE WÄHLT SEINE WÖRTER SORGFÄLTIG AUS.					
ER/SIE HÄLT AN SEINEM GEWÄHLTEN WEG FEST.					
ER/SIE STÜRZT SICH IN DINGE OHNE NACHZUDENKEN.					
ER/SIE TRIFFT VOREILIGE ENTSCHEIDUNGEN.					
ER/SIE HANDELT GERN SPONTAN.					
ER/SIE HANDELT ÜBERSTÜRZT.					
ER/SIE MACHT VERRÜCKTE SACHEN.					
ER/SIE HANDELT OHNE NACHZUDENKEN.					
SEINE/IHRE PLÄNE GESTALTET ER/SIE OFT AUF DEM LETZTEN DRÜCKER					

References

Aldridge R, Davidoff J, Ghanbari M, Hands D, Pearson D (1995) Measurement of scene-dependent quality variations in digitally coded television picture. In: Vision, Image and Signal Processing, IEEE, pp 149–154

Apperley M, Masoodian M (1995) An experimental evaluation of video support for shared work-space interaction. In: Conference Companion on Human Factors in Computing Systems, ACM, pp 306–307

Apple W, Streeter L, Krauss R (1979) Effects of pitch and speech rate on personal attributions. Journal of Personality and Social Psychology 37(5):715–727

Aran O, Gatica-Perez D (2011) Computer analysis of human behavior, Springer, chap Analysis of group conversations: Modeling social verticality, pp 293–322

Argyle M, Lalljee M, Cook M (1968) The effects of visibility on interaction in a dyad. Human Relations 21(1):3–17

Arnauld A, Nicole P (1662) La logique ou l'art de penser [Engl.: Port-royal logic]

Battersby SA (2011) Moving together: The organisation of non-verbal cues during multiparty conversation. PhD thesis, Queen Mary University of London

Beigbeder T, Coughlan R, Lusher C, Plunkett J, Agu E, Claypool M (2004) The effects of loss and latency on user performance in unreal tournament 2003. In: Workshop on Network and System Suppot for Games, ACM

Belmudez B, Möller S (2013) Audiovisual quality integration for interactive communications. EURASIP Journal on Audio, Speech, and Music Processing 2013(1):1–24

Berndtsson G, Folkesson M, V VK (2012) Subjective quality assessment of video conferences and telemeetings. In: International Packet Video Workshop, pp 25–30

Bodden M, Jekosch U (1996) Entwicklung und Durchführung von Tests mit Versuchspersonen zur Verifizierung von Modellen zur Berechnung der Sprachübertragungsqualität., final project report, Institute of Communication Acoustics, Ruhr University, Bochum, Germany

Bouch A, Sasse MA, DeMeer H (2000) Of packets and people: A user-centered approach to quality of service. In: International Workshop on Quality of Service, IEEE, pp 189–197

Brady PT (1965) A technique for investigating on-off patterns of speech. Bell System Technical Journal 44(1):1–22

Brady PT (1968) A statistical analysis of on-off patterns in sixteen conversations. Bell System Technical Journal 47(1):73–91

Brady PT (1971) Effects of transmission delay on conversational behavior on echo-free telephone circuits. Bell System Technical Journal 50(1):115–134

Brander E, Mark G (2001) Social presence with video and application sharing. In: International Conference on Supporting Group Work, ACM, pp 154–161

Bräuer F, Ehsan MS, Kubin G (2008) Subjective evaluation of conversational multimedia quality in IP networks. In: Workshop of Multimedia Signal Processing, IEEE, pp 872–876

Braun AM (2003) Qualitätsaspekte multimodaler Kommunikation: Subjektive und objektive Messungen [engl.: Qualityaspects of multimodal communication: Subjective and objective measurements]. PhD thesis, Eidgenössische Technische Hochschule Zürich

Bühner M, Ziegler M (2009) Statistik fr Psychologen und Sozialwissenschaftler. Pearson Deutschland GmbH

Bunt H, Alexandersson J, Carletta J, Choe JW, Fang AC, Hasida K, Lee K, Petukhova V, Popescu-Belis A, Romary L, Soria C, Traum D (2010) Towards an ISO standard for dialogue act annotation. In: Conference on International Language Resources and Evaluation

Burgoon JK, Buller DB, Woodall WG (1996) Nonverbal communication. The unspoken dialogue. McGraw-Hill

Cermak GW (2002) Subjective quality of speech over packet networks as a function of packet loss, delay and delay variation. International Journal of Speech Technolog 5(1):65–84

Cermak GW (2005) Multimedia quality as a function of bandwidth, packet loss, and latency. International Journal of Speech Technology 8(3):259–270

Cramton CD (2001) The mutual knowledge problem and its consequences for dispersed collaboration. Organization Science 12(3):346–371

Crawford C (2003) The art of interactive design: A euphonious and illuminating guide to building successful software. No Starch Press

Cutler A (1994) The perception of rhythm in language. Cognition 50(1):79–81

Darft RL, Lengel RH (1986) Organizational information requirements, media richness and structural design. Management Science 32(5):554–571

Dennis AR, Kinney ST (1998) Testing media richness theory in the new media: The effects of cues, feedback, and task equivocality. Information Systems Research 9(3):256–274

Deutsches Institut für Normierung (1989-03) DIN 55350 Part 12: Begriffe der Qualitätssicherung und Statistik. Merkmalsbezogene Begriffe. [Engl.: Terms of quality control and statistics. Terms related to features.]. Beuth

Deutsches Institut für Normierung (2008-05) DIN 55350 Part 11: Begriffe der Qualitätssicherung und Statistik. Begriffe des Qualitätsmanagements. [Engl.: Terms of quality control and statistics. Terms of quality management.]. Beuth

Dixon N, Spitz L (1980) The detection of auditory visual desynchrony. Perception 9(6):719–721

Doherty-Sneddon G, Anderson A, O'Malley C, Langton S, Garrod S, Bruce V (1997) Face-to-face and video-mediated communication: A comparison of dialogue structure and task performance. Journal of Experimental Psychology: Applied 3(2):105–125

Dourish P, Adler A, Bellotti V, Henderson A (1996) Your place or mine? Learning from long-term use of audio-video communication. Computer Supported Cooperative Work 5(1):33–62

Egger S, Reichl P (2009) A nod says more than thousand uhmms: Towards a framework for measuring audio-visual interactivity. In: COST298 Conference

Egger S, Schatz R, Scherer S (2010) It takes two to tango - Assessing the impact of delay on conversational interactivity on perceived speech quality. In: Annual Conference of the Speech Communication Association, pp 1321–1324

Egger S, Schatz R, Schoenenberg K, Raake A, Kubin G (2012) Same but different? - Using speech signal features for comparing conversational VoIP quality studies. In: International Conference on Communications, IEEE, pp 1320 – 1324

Egger S, Reichl P, Schoenenberg K (2014) Quality of Experience, Springer, chap Quality of Experience and interactivity, pp 151–163

Fish RS, Kraut RE, Root R, Rice R (1993) Video as a technology for informal communication. Communications of the ACM 36(1):48–61

Friebel M, Loenhoff J, Schmitz HW, Schulte O (2003) Siehst Du mich? Hörst Du mich? Videokonferenzen als Gegenstand kommunikationswissenschaftlicher Forschung. kommunikation@gesellschaft 4(1):1–23

Gale S (1991) Adding audio and video to an office environment. Studies in Computer Supported Cooperative Work: Theory, Practice and Design pp 49–62

Gatica-Perez D (2006) Analyzing group interactions in conversations: A review. In: International Conference on Multisensor Fusion and Integration for Intelligent Systems, IEEE, pp 41–46

Geelhoed E, Parker A, Williams DJ, Groen M (2009) Effects of latency on telepresence. Tech. rep., HPL-2009-120, HP Laboratories

Gellri P, Kanning UP (2007) Handbuch der Arbeits- und Organisationspsychologie [Engl.: Handbook of work- and organizational psychology], Hogrefe, chap Kommunikation und Interaktion [Engl.: Communication and interaction], pp 331–338

Gibbs RW (1987) Mutual knowledge and the psychology of conversational inference. Journal of Pragmatics 11(5):561–588

Gilbert DT, Malone PS (1995) The correspondence bias. Psychological Bulletin 117(1):21–38

Glass L (2001) Synchronization and rhythmic processes in physiology. Nature 410(6825):277–284

Goffman E (1961) The presentation of self in everyday life. Doubleday Anchor

Goffman E (1963) Behavior in public spaces. Notes on the social organization of gatherings. The Free Press

Goffman E (1974) Frame analysis: An essay on the organization of experience. Harvard University Press

Goldberg LR (1992) The development of markers for the Big-Five factor structure. Psychological Assessment 4(1):26–42

Goldberg LR, Johnson JA, Eber HW, Hogan R, Ashton MC, Cloninger CR, Gough HG (2006) The international personality item pool and the future of public-domain personality measures. Journal of Research in Personality 40(1):84–96

Goodwin C, Heritage J (1990) Conversation analysis. Annual Review of Anthropology 19:283–307

Guguin M, Bouquin-Jeans RL, Gautier-Turbin V, Faucon G, Barriac V (2008) On the evaluation of the conversational speech quality in telecommunications. EURASIP Journal on Advances in Signal Processing 2008:1–15

Hammer F (2006) Quality aspects of packet-based interactive speech communication. PhD thesis, Technical University Graz
Hammer F, Reichl P, Raake A (2004) Elements of interactivity in telephone conversations. In: International Conference Spoken Language, pp 1741–1744
Hashimoto Y, Ishibashi Y (2006) Influences of network latency on interactivity in networked rock-paper-scissors. In: Workshop on Network and System Support for Games, ACM, pp 23–27
Hayashi T, Yamagishi K, Tominaga T, Takahashi A (2007) Multimedia quality integration function for videophone services. In: Global Telecommunications Conference, IEEE, pp 2735–2739
Helder GK (1966) Customer evaluation of telephone circuits with delay. Bell System Technical Journal 45(7):1157–1191
Hoeldtke K, Raake A (2011) Conversation analysis of multi-party conferencing and its relation to perceived quality. In: International Conference on Communications, IEEE, pp 1–5
Hoffman D (2000) Visual intelligence: How we create what we see. WW Norton & Company
Holub J, Tomiska O (2009) Delay effect on conversational quality in telecommunication networks: Do we mind? In: Wireless Technology, pp 91–98
Huynh-Thu Q, Barkowsky M, Callet PL (2011) The importance of visual attention in improving the 3d-tv viewing experience: Overview and new perspectives. IEEE Transactions on Broadcasting 57(2):421–431
Iai S, Kurita T, Kitawaki N (1993) Quality requirements for multimedia communication services and terminals-interaction of speech and video delays. In: Global Telecommunications Conference, IEEE, pp 394–398
International Organization for Standardization, ISO (2005) ISO 9000: Quality management systems. fundamentals and vocabulary
International Organization for Standardization, ISO (2012) ISO 24617-2: Language resource management - semantic annotation framework (semaf) - Part 2: Dialogue acts
International Telecommunication Union, ITU-R (1998) Rec. BT.1359-1: Relative timing of sound and vision for broadcasting
International Telecommunication Union, ITU-T (1996) Rec. P.800: Methods for subjective determination of transmission quality
International Telecommunication Union, ITU-T (2000) Rec. P.920: Interactive test methods for audiovisual communications
International Telecommunication Union, ITU-T (2001) Rec. 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 Telecommunication Union, ITU-T (2003a) Rec. G.114: One-way transmission time
International Telecommunication Union, ITU-T (2003b) Rec. G.131: Talker echo and its control
International Telecommunication Union, ITU-T (2007a) Rec. P.79: Calculation of loudness ratings for telephone sets

International Telecommunication Union, ITU-T (2007b) Rec. P.805: Subjective evaluation of conversational quality
International Telecommunication Union, ITU-T (2008) Rec. E.800: Definitions of terms related to quality of service
International Telecommunication Union, ITU-T (2011a) Rec. G.107: The E-Model: A computational model for use in transmission planning
International Telecommunication Union, ITU-T (2011b) Rec. G.107.1: Wideband E-Model
International Telecommunication Union, ITU-T (2012a) P-series supplement p.sup26 (06/12) - scenarios for the subjective quality evaluation of audio and audiovisual multiparty telemeetings
International Telecommunication Union, ITU-T (2012b) Rec. G.1070: Opinion model for video-telephony applications
International Telecommunication Union, ITU-T (2012c) Rec. P.1301: Subjective quality evaluation of audio and audiovisual multiparty telemeetings
International Telecommunication Union, ITU-T (2014) Rec. P.863: Perceptual objective listening quality assessment (polqa)
International Telecommunication Union, ITU-T, AT&T (2004) Delayed contribution 214, echo-free delay, voip speech quality and the E-Model. Tech. rep., ITU-T Study Group 12
International Telecommunication Union, ITU-T, Bell-Northern Research (1990) Delayed contribution 21, the subjective impact of pure delay on voice connections. Tech. rep., ITU-T Study Group 12
International Telecommunication Union, ITU-T, Deutsche Telekom (2013) E-Model update regarding delay and interactivity (rec. itu-t g.107). Tech. rep., ITU-T Study Group 12
International Telecommunication Union, ITU-T, Telekom Canada (1990) Delayed contribution 22, subjective impact of ddelay in the absence of echo. Tech. rep., ITU-T Study Group 12
Isaacs EA, Tang JC (1994) What video can and cannot do for collaboration: A case study. Multimedia Systems 2(2):63–73
Ishibashi Y, Tasaka S (2003) Causality and media synchronization control for networked multimedia games: Centralized versus distributed. In: NetGames, ACM, pp 34–43
Issing J, Nikolaus N (2012) Conversational quality as a function of delay and interactivity. In: International Conference on Software, Telecommunications and Computer Networks, IEEE, pp 1–5
Jaffe J, Beebe B, Feldstein S, Crown CL, Jasnow MD, Rochat P, Stern DN (2001) Rhythms of dialogue in infancy: Coordinated timing in development, vol 66. Monographs of the Society for Research in Child Development
Jekosch U (2005) Voice and speech quality perception. Assessment and evaluation. Springer
Jokinen K (2011) Turn taking, utterance density, and gaze patterns as cues to conversational activity. In: International Conference on Multimodal Interaction, Workshop on Multimodal Corpora for Machine Learning, pp 31–36

Jones E, Gallois C, Callan V, Barker M (1999) Strategies of accommodation: Development of a coding system for conversational interaction. Journal of Language and Social Psychology 18(2):123–152

Jones SE, LeBaron CD (2002) Research on the relationship between verbal and nonverbal communication: Emerging integrations. Journal of Communication 52(3):499–521

Jordan B, Henderson A (1995) Interaction analysis: Foundations and practice. The Journal of the Learning Sciences 4(1):39–103

Kahneman D (1973) Attention and effort. Prentice Hall

Kahneman D (2003) Well-Being: The foundations of hedonic psychology, Russell Sage Foundation, chap Objective happiness, pp 3–25

Karis D (1991) Evaluating transmission quality in mobile telecommunication systems using conversation tests. In: Annual Meeting of the Human Factors Society, pp 217–221

Kendon A (1967) Some function of gaze direction in social interaction. Acta Psychologica 26:22–63

Kendon A (1972) Review of the book kinesics and context by Ray Birdwhistell. American Journal of Psychology 85:441–445

Keysar B, Barr DJ, Balin JA, Paek T (1998) Definite reference and mutual knowledge: Process models of common ground in comprehension. Journal of Memory and Language 39(1):1–20

Keysar B, Barr DJ, Balin JA, J S Brauner JS (2000) Taking perspective in conversation: The role of mutual knowledge in comprehension. Psychological Science 11(1):32–38

Kiousis S (2002) Interactivity: A concept explication. New Media & Society 4(3):355–383

Kitawaki N, Itoh K (1991) Pure delay effects on speech quality in telecommunications. Journal on Selected Areas in Communications, IEEE 9(4):586–593

van der Kleij R, Schraagen JM, Werkhoven P, DeDreu CK (2009) How conversations change over time in face-to-face and video-mediated communication. Small Group Research 40(4):355–381

Klemmer ET (1967) Subjective evaluation of transmission delay in telephone conversations. Bell System Technical Journal 46(6):1141–1147

Knapp ML, Hall J (2010) Nonverbal communication in human interaction, Cengage, chap Nonverbal communication: Basic perspective, pp 3–30

Krauss RM (2001) International encyclopaedia of the social and behavioural sciences, Elsevier, chap The psychology of verbal communication, pp 16.161–16.165

Krauss RM, Bricker PE (1967) Effects of transmission delay and access delay on the efficiency of verbal communication. Journal of the Acoustical Society of America 41(2):286–292

Krauss RM, Garlock C, Bricker PD, McMahon L (1977) The role of audible and visible back-channel responses in interpersonal communication. Journal of Pesonality and Social Psychology 35(7):523–529

Kraut R, Lewis S, Swezey L (1982) Listener responsiveness and the coordination of conversation. Journal of Personality and Social Psychology 43(4):718–731
Kurita T, Lai S, Kitawaki N (1994) Effects of transmission delay in audiovisual communication. Electronics and Communications in Japan (Part I: Communications) 77(3):63–74
Lakaniemi A, Rosti J, Raisanen V (2001) Subjective VoIP speech quality evaluation based on network measurements. In: International Conference on Communications, pp 748 – 752
Laver J, Hutcheson S (1972) Communication in face-to-face interaction. Penguin
Lay CH, Burron BF (1968) Perception of the personality of the hesitant speaker. Perceptual and Motor Skills 26(3):951–956
Mausfeld R (2002) Perception and the physical world, Wiley, chap The physicalistic trap in perception theory, pp 75–114
Mausfeld R (2003) Looking into pictures: An interdisciplinary approach to pictorial space, MIT Press, chap Conjoint representations and the mental capacity for multiple simultaneous perspectives, pp 17–60
Mausfeld R (2010a) Cognition and neuropsychology: International perspectives on psychological science, vol 1, Psychology Press, chap Intrinsic multiperspectivity: On the architectural foundations of a distinctive mental capacity, pp 95–116
Mausfeld R (2010b) Perception beyond inference. The information content of visual processes, MIT Press, chap The perception of material qualities and the internal semantics of the perceptual system, pp 159–200
Mausfeld R (2011) Interdisciplinary anthropology. Continuing evolution of man, Springer, chap Intrinsic multiperspectivity: Conceptual forms and the functional architecture of the perceptual system, pp 19–54
Mausfeld R (2013) Handbook of experimental phenomenology. Visual peception of shape, space and appearance, Wiley, chap The attribute of realness and the internal organization of perceptual reality, pp 91–118
Max Planck Institute for Psycholinguistics, The Language Archive (2008) EUDICO Linguistic Annotator. URL http://tla.mpi.nl/tools/tla-tools/elan/
McGrath JE (1984) Groups: Interaction and performance. Englewood Cliffs, NJ:Prentice-Hall
Miller N, Maruyama G, Beaber RJ, Valone K (1976) Speed of speech and persuasion. Journal of Personality and Social Psychology 34(4):615–624
Miner N, Caudell T (1998) Computational requirements and synchronization issues of virtual acoustic displays. Presence 7(4):396–409
Möller S (2000) Assessment and prediction of speech quality in telecommunications. Kluwer Academic Publishers
Möller S, Raake A (2002) Telephone speech quality prediction: Towards network planning and monitoring models for modern network scenarios. Speech Communication 38(1):47–75
Möller S, Raake JBA, Wältermann M, Weiss B (2011) A new dimension-based framework model for the quality of speech communication services. In: International Workshop on Quality of Multimedia Experience, pp 107 – 112

Möller S, Callet PL, Perkis A (eds) (2013) Qualinet white paper on definitions of Quality of Experience. European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Lausanne, Switzerland

Nöth W (2000) Handbuch der Semiotik. 2., vollständig neu bearbeitete und erweiterte Auflage [Engl.: Handbook of Semiotics]. Metzler

O'Conaill B, Whittaker S, Wilbur S (1993) Conversations over videoconferences: an evaluation of the spoken aspects of video mediated interaction. Human Computer Interaction 8(4):389–428

Olson JS, Olson GM, Meader DK (1995) What mix of video and audio is useful for small groups doing remote real-time design work? In: Human Factors in Computing Systems, ACM, pp 362–368

Osgood CE, Suci G, Tannenbaum P (1957) The measurement of meaning. University of Illinois Press

Pantel L, Wolf LC (2002) On the impact of delay on real-time multiplayer game. In: International Workshop on Network and Operating Systems Support for Digital Audio and Video, pp 23–29

Petty RE, Wegener DT (1998) The handbook of social psychology, Vol I, University Press, chap Attitude change: Multiple roles for persuasion variables, pp 343–366

Pomerantz A (1984) Structures of social action: Studies in conversation analysis., Cambridge University Press, chap Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shapes., pp 57–101

Poyatos F (1976) Man beyond words. New York State University

Pressing J (1999) The referential dynamics of cognition and action. Psychological Review 106(4):714–747

Puckette M (2007) The theory and technique of electronic music. URL http://puredata.info/

Pye R, Williams E (1977) Teleconferencing: is video valuable or is audio adequate? Telecommunications Policy 1(3):230–241

Quax P, Monsieurs P, Lamotte W, Vleeschauwer DD, Degrande N (2004) Objective and subjective evaluation of influence of small amounts of delay and jitter on a recent first person shooter game. In: Network and System Support for Gamens, ACM, pp 152 – 156

Raake A (2006) Speech quality of VoIP: Assessment and prediction. John Wiley & Sons

Raake A, Egger S (2014) Quality of Experience: Advanced concepts, applications and methods, Springer, chap Quality and Quality of Experience, pp 11–34

Raake A, Schlegel C (2010) 3cts - 3-party conversation test scenarios for conferencing assessment. DOI 10.5281/zenodo.16129, URL https://zenodo.org/record/16129

Raake A, Hoeldtke K, Schlegel C, Ahrens J, Geier M (2010) Listening and conversational quality of spatial audio conferencing. In: International Conference of the Audio Engineering Society, pp 4–7

Raake A, Schoenenberg K, Skowronek J, Egger S (2013) Predicting speech quality based on interactivity and delay. In: Annual Conference of the International Speech Communication Association, pp 1384–1388

Rafaeli S (1988) Interactivity: From new media to communication. Annual Review of Communication Research: Advancing Communication Science 16:110134
Repp B (2005) Sensorimotor synchronization: A review of the tapping literature. Psychonomic Bulletin & Review 12(6):969–992
Rice RE (1987) The new media: Communication, research, and technology, Sage, chap Mediated group communication, p 107120
Rice RE (1992) Task analysability, use of new media, and effectiveness. a multisite exploration of media richness. Organization Science 3(1):475–500
Richards DL (1962) Conversational performance of speech links subject to long propagation times. In: International Conference on Satellite Communication, IEEE, pp 955–963
Riesz RR, Klemmer ET (1963) Subjective evaluation of delay and echo suppressors in telephone communications. Bell System Technical Journal 42(6):2919–2941
Ross L (1977) Advances in experimental social psychology, vol 10, Academic Press, chap The intuitive psychologist and his shortcomings, pp 173–220
Ruhleder K, Jordan B (1999) Meaning-making across remote sites: how delays in transmission affect interaction. In: European Conference on Computer-Supported Cooperative Work, pp 411–429
Ruhleder K, Jordan B (2001) Co-constructing non-mutual realities: Delay-generated trouble in distributed interaction. Computer Supported Cooperative Work 10(1):113–138
Sacks H, Schegloff EA, Jefferson G (1974) A simplest systematics for the organization of turn-taking for conversations. Language 50(4):696–734
Sassenberg K (2004) Electronic human ressource im inter-und intranet [Engl.: Electronic human ressource in the inter-und intranet], Hogrefe, chap Formen und Bedeutung elektronischer Kommunikation in Unternehmen [Engl.: The format and meaning of electronical communication in businesses], pp 92–109
Sat B, Wah BW (2009) Analyzing voice quality in popular voip applications. MultiMedia, IEEE 16(1):46–59
Sat B, Huang Z, Wah BW (2007) The design of a multi-party VoIP conferencing system over the internet. In: International Symposium on Multimedia, IEEE, pp 3–10
Schaefer C, Enderes T, Ritter H, Zitterbart M (2002) Subjective quality assessment for multiplayer real-time games. In: Network and System Support for Games, ACM, pp 74–78
Schegloff EA, Jefferson G, Sacks H (1977) The preference for selfcorrection in the organization of repair in conversation. Language 53(2):361–382
Schmitt M, Gunkel S, Cesar P, Bulterman D (2014) The influence of interactivity patterns on the quality of experience in multi-party video-mediated conversations under symmetric delay conditions. In: International Workshop on Socially-Aware Multimedia, pp 13–16
Schoenenberg K, Raake A (2011) The effect of gender and mood change on the perceived integral quality, besides technical conditions in teleconferencing. In: Fortschritte der Akustik, DEGA, pp 263–264

Schoenenberg K, Schmieder M (2014) 3rnt - 3-party random number verification (timed version) task. DOI 10.5281/zenodo.16133, URL https://zenodo.org/record/16133

Schoenenberg K, Raake A, Egger S, Schatz R (2014a) On interaction behaviour in telephone conversations under transmission delay. Speech Communication 63:114

Schoenenberg K, Raake A, Koeppe J (2014b) Why are you so slow? - misattribution of transmission delay to attributes of the conversation partner at the far-end. Internationsl Journal of Human-Computer Studies 72(5):477–487

Schrire S (2004) Interaction and cognition in asynchronous computer conferencing. Instructional Science 32(6):475–502

Sellen AJ (1995) Remote conversations: The effects of mediating talk with technology. Human-Computer Interaction 10(4):401–444

Short J, Williams E, Christie B (1976) The social psychology of telecommunications. John Wiley & Sons, London

da Silva APC, Varela M, de Souza e Silva E, ao RMML, Rubino G (2008) Quality assessment of interactive voice application. Computer Networks 52(6):1179–1192

Skowronek J, Raake A, Hoeldtke K, Geier M (2011) Speech recordings for systematic assessment of multi-party conferencing. In: Forum Acusticum, pp 111–116

Skowronek J, Herlinghaus J, Raake A (2013) Quality assessment of asymmetric multiparty telephone conferences: A systematic method from technical degradations to perceived impairments. In: Annual Conference of the International Speech Communication Association, pp 2604–2608

Skowronek J, Schiffner F, Schoenenberg K (2014) 3sct 3-party short conversation test scenarios for conferencing assessment (version 03). DOI 10.5281/zenodo.16136, URL https://zenodo.org/record/16136

Skype (2010) URL blogs.skype.com/en/2010/09/the_power_of_silk.html

Streeck J, Knapp ML (1992) Advances in nonverbal communication, Benjamins, chap The interaction of visual and verbal features in human communication, pp 3–23

Stromer-Galley J (2004) Interactivity-as-product and interactivity-as-process. The Information Society: An International Journal 20(5):391–394

Tam J, Carter E, Kiesler S, Hodgins J (2012) Video increases the perception of naturalness during remote interactions with latency. In: Human Factors in Computing Systems, ACM, pp 2045–2050

Thomsen G, Jani Y (2000) Internet telephony: Going like crazy. Spectrum, IEEE 37(5):52 – 58

Trager GL (1958) Paralanguage: A first approximation. Studies in Linguistics 13(1-2):1–10

Trevarthen C (1993) The self born in intersubjectivity: An infant communicating, Cambridge University Press, pp 121–173

Vartabedian AG (1966) The effects of transmission delay in four-wire teleconferencing. Bell System Technical Journal 45(10):1673–1688

Wah BW, Sat B (2009) The design of VoIP systems with high perceptual conversational quality. Journal of Multimedia 4(2):49–62
Wältermann M (2013) Dimension-based quality modeling of transmitted speech. Springer
Wang J, Yang F, Xie Z, Wan S (2010) Evaluation on perceptual audiovisual delay using average talkspurts and delay. In: International Congress on Image and Signal Processing, p 125128
Ward N, Nakagawa S (2004) Automatic user-adaptive speaking rate selection. International Journal of Speech Technology 7(4):259–268
Watson A (2001) Assessing the quality of audio and video components in desktop multimedia conferencing. PhD thesis, University of London
Weiss B, Schoenenberg K (2014) Conversational structures affecting auditory likeability. In: Annual Conference of the International Speech Communication Association, p 17911795
Werlberger M, Pock T, Bischof H (2010) Motion estimation with non-local total variation regularization. In: Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, pp 2464–2471
Whittaker S (1995) Rethinking video as a technology for interpersonal communication: Theory and design implication. International Journal of Human-Computer Studies 42(5):501–529
Wollermann C, Lasarcyk E (2007) Modeling and perceiving of (un)certainty in articulatory speech synthesis. In: Workshop on Speech Synthesis, pp 40–47
Yamagishi K, Hayashi T (2006) Opinion model using psychological factors for interactive multimodal services. In: Transactions on Communications, IEICE, pp 281–288
Zuberbühl HJ (2003) Quality aspects of multimodal communication: User perception and acceptance thresholds. PhD thesis, Swiss Federal Institute of Technology, Zürich