Simulating Conversations for the Prediction of Speech Quality

vorgelegt von
Thilo Michael, M.Sc.
ORCID: 0000-0002-1086-6882

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

Doktor der Ingenieurwissenschaften
– Dr.-Ing. –

genehmigte Dissertation

Promotionsausschuss
Vorsitzender: Prof. Dr. Henning Sprekeler
Gutachter: Prof. Dr.-Ing. Sebastian Möller
Gutachter: Prof. Dr. Bernd Möbius
Gutachter: Dr.-Ing. Florian Hammer

Tag der wissenschaftlichen Aussprache: 02. November 2022

Berlin 2022

Abstract

The measurement and prediction of speech quality are crucial planning tools for Voice over Internet Protocol (VoIP) communication providers. Current instrumental models that predict the quality of speech in a conversation scenario mainly rely on parameters of the transmission system for their prediction. However, for some degradations, it has been shown that the impact on the conversation, and thus the perceived quality, cannot be modeled by the parameters of the transmission alone. The effect of transmission delay on a telephone conversation depends on conversational interactivity, as the delayed speech signal slows down the turn-taking of the conversation partners. The impact of packet loss, while being audible in a listening situation, is also dependent on the part of transmitted information that is lost and, thus, whether the conversation partner needs to resolve a misunderstanding with additional repairing dialogue. In conversations where these impairments co-occur, interactivity effects may arise, as the meta-communication due to lost packets is, in turn, affected by transmission delay. As current instrumental quality prediction models do not consider these factors and their interaction, they cannot account for them.

This thesis introduces conversation simulation as a new approach to the instrumental prediction of conversational quality. A simulation architecture is described based on incremental spoken dialogue processing that can model standardized conversation scenarios on the concept, turn-taking, and speech signal level. Especially the changes in turn-taking during delayed transmission and the retransmission of information due to packet loss are modeled and evaluated based on empirical conversations. The resulting simulated conversations are assessed with methods from the field of spoken dialogue systems and speech quality, resulting in parameters that represent the changes in conversations due to delay and packet loss. The fullband E-model, a standardized parametric model, is extended for conversational interactivity and bursty packet loss to utilize the parameters extracted from the conversations. Finally, the conversational quality is predicted based on the extended E-model and the parameters from the simulated conversations.

Zusammenfassung

Die Messung und Vorhersage der Sprachqualität ist ein wichtiges Planungsinstrument für Anbieter von Voice-over-Internet-Protocol-Diensten. Aktuelle instrumentelle Modelle, die die Sprachqualität in einem Gesprächsszenario vorhersagen, stützen sich hauptsächlich auf Parameter des Übertragungssystems für ihre Vorhersage. Es hat sich jedoch gezeigt, dass die Auswirkungen auf das Gespräch und damit auf die wahrgenommene Qualität bei einigen Störungen nicht allein durch die Parameter der Übertragung modelliert werden können. Die Auswirkung einer Übertragungsverzögerung auf ein Telefongespräch hängt von der Interaktivität des Gesprächs ab, da das verzögerte Sprachsignal die Gesprächsteilnehmer in ihrem Redefluss bremst. Die Auswirkungen von Paketverlusten sind zwar in einer Hörsituation erkennbar, hängen aber auch davon ab, welcher Teil der übertragenen Informationen verloren geht und ob der Gesprächspartner ein Missverständnis durch einen zusätzlichen Reparaturdialog aufklären muss. In Gesprächen, in denen diese Beeinträchtigungen gleichzeitig auftreten, kann es zu Interaktivitätseffekten kommen, da die Metakommunikation aufgrund verlorener Pakete wiederum durch die Übertragungsverzögerung beeinträchtigt wird. Da die derzeitigen Modelle zur Qualitätsvorhersage diese Faktoren und ihre Wechselwirkung nicht berücksichtigen, können sie diese nicht in die Vorhersage mit einbeziehen.

In dieser Arbeit wird die Konversationssimulation als neuer Ansatz für die instrumentelle Vorhersage der Gesprächsqualität vorgestellt. Es wird eine Simulationsarchitektur beschrieben, die auf der inkrementellen Verarbeitung gesprochener Dialoge basiert und standardisierte Gesprächsszenarien auf Konzept-, Turn-Taking- und Sprachsignalebene modellieren kann. Insbesondere werden die Veränderungen im Turn-Taking bei verzögerter Übertragung und die erneute Übertragung von Informationen aufgrund von Paketverlusten modelliert und anhand von empirischen Gesprächen bewertet. Die daraus resultierenden simulierten Gespräche werden mit Methoden aus dem Bereich der gesprochenen Dialogsysteme und der Gesprächsanalyse ausgewertet, so dass sich Parameter ergeben, die die Veränderungen in Gesprächen aufgrund von Verzögerungen und Paketverlusten darstellen. Das Vollband-E-Modell, ein standardisiertes parametrisches Modell, wird für Gesprächsinteraktivität und Paketverluste erweitert, um die aus den Gesprächen extrahierten Parameter zu nutzen. Schließlich wird die Konversationsqualität auf der Grundlage des erweiterten E-Modells und der Parameter aus den simulierten Konversationen vorhergesagt.

Acknowledgements

This thesis is the result of over 5 years of work at the Quality and Usability Lab and would not have been possible without the support of friends, colleagues, and family. First, I would like to thank my supervisor, mentor, and friend Prof. Dr. Sebastian Möller, for all the scientific advice and discussions, but also for the moments we didn't talk about work. Special thanks go to Dr. Stefan Hillmann and Dr. habil. Benjamin Weiss, who introduced me to the scientific world of speech quality and spoken dialogue systems and motivated me to start my doctoral studies. I would like to thank Prof. Dr. Bernd Möbius and Dr. Florian Hammer for reviewing my dissertation and for serving on my doctoral committee. I want to give special thanks to Irene Hube-Achter and Yasmin Hillebrenner, who keep the lab running and were always a great support in any regard.

I would like to thank my student workers, Jannik Reichert, Jana Müller, and especially Elisabeth von Oswald, who helped me with my experiments and were a pleasure to work with. A huge thanks also to all my colleagues for constructive discussions, but most importantly, for the great company and talks over some good coffee, including Robert Spang, Carola Trahms, Wafaa Wardah, Tanja Kojic, Vera Schmidt, Maurizio Vergari, Gabriel Mittag, Steven Schmidt, Saman Zadtootaghaj, Babak Naderi, Philine Görzig, Britta Hesse, Salar Mohtaj, Sai Sirisha Rallabandi, Benjamin Bähr, Friedemann Köster, and all the others! Additionally, I would like to thank all the people from the Freitagsrunde and the Minitiative.

A huge thank you goes to my emotional support group, which also happens to be my family: Klaus, Uta, Jane, Ali, Hanna, Mona, Inga, Antje, and Fine.

Finally, I want to thank my wife Milena for always supporting me and having the patience to listen to my rambles about work, for reading my dissertation multiple times, for listening to my defense talk over and over again, and for being the best person I know.

Contents

Acronyms

ACR	Absolute Category Rating
AI	Articulation Index
AIR	Active Interruption Rate
ASR	Automatic Speech Recognitino
CCR	Comparison Category Rating
CDR	Conversation Disruption Rate
DA	Dialogue Act
DCR	Degradation Category Rating
DIAL	Diagnostic Instrumental Assessment of Listening quality
DM	Dialogue Manager
DT	Double Talk
DTR	Double Talk Rate
ECS	Extended Continuous Scale
ETSI	European Telecommunication Standard Institute
EVS	Enhanced Voice Services
FB	Fullband
HCI	Human-Computer-Interaction
IDCQ	Instrumental Diagnostic Conversational Quality
IIR	Intended Interruption Rate
InproTK	Incremental Processing Toolkit
IR	Interruption Rate
IU	Incremental Unit
KL	Kullback-Leibler
MOS	Mean Opinion Score
MS	Mutual Silence
NB	Narrowband
NI	Non-successful Interruption
NISQA	Non-instrusive Speech Quality Assessment
NLG	Natural Language Generation
NLU	Natural Language Understanding
OCSQ	Objective Conversational Speech Quality
OTT	Over-The-Top
P-CA	Parametric Conversation Analysis
PCM	Pulse Code Modulation

PESQ Perceptual Evaluation of Speech Quality
PESQM Perceptual Echo and Sidetone Quality Measure
PIR Passive Interruption Rate
PLC Packet Loss Concealment
POLQA Perceptual Objective Listening Quality Assessment
POTS Plain Old Telephone Service
PR Pause Rate
PSTN Public Switched Telephone Networks
RMSE Root-Mean-Square Error
RNV Random Number Verification
SA Speaker A
SAR Speaker Alternation Rate
SARc Corrected Speaker Alternation Rate
SB Speaker B
SCT Short Conversation Test
SD Standard Deviation
SDS Spoken Dialogue Systems
SI Successful Interruption
SII Speech Intelligibility Index
SUS Semantically Unpredictable Sentences
SWB Superwideband
TTS Text-To-Speech
UIR Unintended Interruption Rate
VAD Voice Activity Detection
VoIP Voice over Internet Protocol
WB Wideband
WER Word Error Rate

Chapter 1
Introduction

1.1 Motivation

Conversations over the telephone have been an integral part of our everyday lives for over a century. In recent years, with the rise of smartphones and their ubiquitous internet access, the ways of communication have shifted. While previously, the interaction was dominated by voice-only telephone calls over publicly switched or mobile networks, calls have now moved towards an internet-driven, rich multimedia experience. This shift has not only been restricted to private calls. Also, in the business telecommunication area, distributed and remote teams rely on flexible voice and multimedia communication over the internet. Especially the COVID-19 pandemic has given a push to video-conferencing, Over-The-Top (OTT) speech communication services, and remote working communication software, as the remote communication scenario is now a daily occurrence for many people.

Thus, especially now, it is essential for speech communication service providers to plan, measure and monitor their networks and services to provide an overall satisfying experience for their customers. The quality of transmitted speech can be measured by asking users for their subjective ratings (subjective methods), or it can be predicted to plan or monitor speech transmission networks (objective methods). Based on those ratings and predictions, service providers optimize the quality while using the least resources. Subjective methods for measuring the speech quality include listening, speaking, and conversation tests, where participants directly rate prepared speech samples, speaking situations, or even whole conversations using the system under study. This allows researchers to quantify the quality of the transmitted speech as the end-users perceive it. Objective (or instrumental) methods try to estimate results of the subjective tests by predicting how end-users would rate a given speech sample or speech communication service.

With the move towards using the internet to transmit speech with the Voice over Internet Protocol (VoIP), speech communication providers have increased the bandwidth of transmission and made telephony much more flexible. Also, many OTT services have emerged that utilize the network capabilities of computers and smartphones to provide speech and video communication services over the internet. This packet-based transmission of the speech signal has also changed the types of degradation that speech communication services deal with. While for the Plain Old Telephone Service (POTS), problems with loudness, circuit noise, and a small bandwidth were prevalent, impair-

ments in today's VoIP telephony are mostly packet related. With speech samples being split into small packages and transmitted over distributed networks, packets might get lost during the transmission or arrive too late to be used for the speech decoding process. This packet loss results in an audible dropping out of the speech. Current coding algorithms employ Packet Loss Concealment (PLC) by reconstructing the speech signal of the missing packet based on the previous and potentially the following packets. However, depending on the timing and frequency of the packet loss, artifacts can still be perceived by the end-user. In order to mitigate packet loss, speech communications services make use of a jitter buffer that buffers a flexible amount of packets to make the connection robust against packets that arrive too late or even allow for the time to resent a missing or corrupted packet.

Especially in real-time services like online telephony, the delay caused by the jitter buffer, the general uncertain latency of the network, and delays due to speech signal processing on the user's device influence the interactivity of conversations. The end-to-end transmission delay that is often present in IP-based communication cannot be perceived audibly. Nonetheless, it affects the interaction by slowing down the speed of speaker changes. Furthermore, it often causes unwanted interruptions and miscommunication because the orderly turn-taking present in delay-free conversations is no longer possible, as turn-taking cues in the signal arrive too late. However, it has been shown that these effects on the interactivity depend not only on the overall delay but also on the type of conversation being carried out. For example, a highly interactive discussion has faster and denser speaker changes and is thus more affected by transmission delay. In contrast, a slow conversation with limited speaker alternations will have fewer interruptions and turn-taking problems. This difference in the smoothness of the conversation is also reflected in the overall conversational quality ratings of the interlocutors. For packet loss, no direct relationship between interactivity and quality ratings has been measured. However, in conversational contexts, packet loss can cause important information to not be correctly transmitted, resulting in additional "repairing" communication to retransmit the lost information. In transmission scenarios where both delay and packet loss are present, the way one impairment affects the conversation might impact the perception of the other.

Current signal-based speech quality models like NISQA (Mittag et al., 2021) or POLQA (ITU-T Recommendation P.863, 2014) predict the listening quality and the underlying perceptual dimensions: coloration, continuity, noisiness, and loudness. However, this approach cannot model the impact of transmission delay because this impairment is not audible in listening-only tests. Due to the realistic and interactive nature of conversations, conversational quality is able to capture all effects of transmission impairments on the conversation. For its subjective evaluation, conversation tests are used, where two participants are connected through a simulated telephone network and carry out standardized conversation scenarios. Compared to listening quality experiments, the assessment of conversational quality is more time-intensive and costly. The resulting conversations can be used to perform a conversation analysis that results in metrics for measuring the interactivity and the smoothness of turn-taking in the conversation. For conversational quality, there exists no standardized signal-based model. The E-model, a parametric model for predicting conversational quality, is able to include impairments due to delayed transmission. The narrowband version of the E-model includes parame-

ters that incorporate the interactivity of the conversation. However, there are limits on how well the conversational quality can be predicted with this approach: the interactivity parameters require empirical data from previously recorded conversations in order to be calculated. Also, interactivity effects between delay and other impairments cannot be calculated, as the E-model assumes degradations to be independent of each other. These shortcomings limit the network planning capabilities, as no model exists that is able to include the turn-taking interactions of a conversation.

The simulation of dialogues has been a research area in the fields of computational linguistics and Spoken Dialogue Systems (SDS) for many years. Here, conversations are usually simulated by a semantic representation of turns that are exchanged in a dialogue between a human and a dialogue system. The resulting conversations are either analyzed to predict the interaction quality of the dialogue system under study or used to train the dialogue managing component of the dialogue system to create correct and efficient dialogues. Newer research in the area of SDS also focuses on the turn-taking of dialogues to increase the interactivity and naturalness of Human-Computer-Interaction (HCI).

In this work, the concepts of dialogue and user simulation in the domain of HCI are applied to a human-to-human conversation simulation to predict speech quality. Specifically, a simulation is designed to model conversations between two interlocutors that interact with each other based on conversation scenarios standardized in ITU-T Recommendation P.805 (2007). To model behavior during conversations with transmission delay and packet loss, the simulated agents perform smooth turn-taking and simulate the understanding of incoming information based on incoming packet loss patterns. Between the two interlocutors, a simulated network introduces both delay and packet loss, which incites changes in the behavior of the virtual agents. The resulting conversations are recorded and analyzed, forming the basis for a conversational quality prediction.

To predict the conversational quality from the simulated conversations, the E-model is being adapted to include interactivity parameters and bursty packet loss. These extensions are validated with empirical conversations collected in a conversation test. Additionally, an E-model-independent quality model that is based on the parameters of the conversations is created and evaluated. In the last step, conversations with combinations of different interactivity levels, transmission delay, and packet loss probabilities are simulated and used to predict the conversational quality with the extended E-model as well as the parameter-based model.

1.2 Objective and Research Questions

So far, predictive models for speech quality either do not consider the interactive aspect of a conversation or include them as parameters that have to be observed empirically beforehand. This limits the usefulness of such models for network planning. This work will explore the use of concepts from the area of Spoken Dialogue Systems to simulate a conversation, allowing for quality models that explicitly include the interactivity of conversations and take into account the side-effects of combinations of impairments. While the simulation approach can model changes in ratings due to different interactivity, it can also model a variety of conversation strategies, thus resulting in a range of distinct conversations and multiple quality predictions for a single experimental condition.

Because a simulation of human-to-human conversations has not been a research topic before, the work is split into two main parts. First, a simulation environment is being designed, that is able to simulate different conversation types and to model turn-taking and overall the interactivity of conversations. This simulation environment is then extended to cater for specific behavior required for conversations under the influence of packet loss and transmission delay. Secondly, a quality model is created and the existing parametric E-model is extended to be able to incorporate the additional information from the simulated conversations. Finally, the simulation is evaluated by predicting the quality from parameters of the recorded audio and transcriptions of the simulated conversations.

Based on this approach, the following five research questions are answered in this thesis:

1. **Simulation of Conversations:** How can the models and methods of dialogue and user simulation from the area of Spoken Dialogue Systems be applied to the simulation of conversations between two humans?
2. **Simulation of Turn-Taking:** How can the smooth taking of turns in natural VoIP conversations of different levels of interactivity be replicated in a simulation?
3. **Turn-Taking during Delayed Speech Transmission:** How is turn-taking affected by transmission delay, and what rules and models can be employed to replicate these changes in a simulation?
4. **Understandability and Packet Loss:** What impact has bursty packet loss on the understandability of speech in a conversation, and how can it be modeled in a simulation?
5. **Conversational Quality Prediction:** How well can conversational parameters and the overall quality be predicted with this new approach?

1.3 Thesis Structure

Chapter 2 of this thesis details the fundamental and related work that is needed for the proposed conversation simulation approach. It introduces the main concepts of speech transmission and quality. It gives an overview of the standards of conversational quality and introduces the major conversational parameters that can be extracted with Parametric Conversation Analysis (P-CA). Afterwards, an overview of instrumental quality prediction is given, with a focus on the E-model. Lastly, related work in the area of dialogue and user simulation, as well as incremental dialogue systems, is highlighted.

In Chapter 3, the datasets used in this thesis are presented, the technical design and architecture of the simulation environment are described, and the simulation is evaluated. The chapter starts with a description of the datasets and their annotations. This data is used throughout this work to train and model the simulation, as well as to evaluate the quality prediction models. Then, a brief overview of the *retico* incremental processing framework, which was designed to provide a foundational programming framework for the implementation of the simulation, is given. Following this, the incremental

simulation setup is described, detailing all components that are used in the simulation. Lastly, the semantics and interactivity of the resulting simulated conversations are analyzed and compared to the above baseline dataset.

Chapter 4 firstly analyzes the changes in turn-taking patterns in conversations with different levels of interactivity. Based on this, a turn-taking model is created, implemented into the simulation from Chapter 3, and finally evaluated. Following this, the impact of delay on conversations with distinct interactivity levels is analyzed, and the turn-taking model is evaluated in regards to the changes occurring with transmission delay. Next, necessary adaptions to the turn-taking model are explained and implemented, followed by a final evaluation and discussion of the adapted turn-taking model.

In Chapter 5, conversation disruptions, which are a measure of misunderstandings due to packet loss, are introduced. A model is created to simulate the behavior inside the simulation and is integrated into the simulation of Chapter 3. The resulting simulated conversations are then analyzed and compared to empirical conversations and the results are discussed.

Chapter 6 adapts the E-model towards interactivity, delay, and bursty packet loss and introduces a new model that predicts conversational quality based on interactivity parameters of a conversation. Then, these models are used to predict speech quality from the simulated conversation, validating the simulation approach.

Finally, Chapter 7 recapitulates the results of the simulation with the delay and packet loss extensions, as well as its quality prediction capabilities, and an outlook on future work is presented.

Chapter 2
Fundamentals

The conversation simulation approach explored in this thesis draws mainly from two research fields: the technology and groundwork for the simulation are based on the methods in the area of Spoken Dialogue Systems, while the analysis of conversational structures and the prediction of conversational quality falls into the area of Speech Quality.

This chapter describes the foundations and related work of speech quality, dialogue systems, and dialogue simulation. The fundamentals in speech quality focus on the technical impairments delay and packet loss, as well as the assessment of conversational quality, which includes the parametric conversation analysis method. Also, an introduction to instrumental speech quality prediction, with a focus on the E-model, will be given. The fundamentals in SDS will outline the use cases of dialogue simulation and the architectural details needed for the implementation of a conversation simulation.

2.1 Speech Transmission

Currently used speech transmission services fall into three types of categories: the landline telephone network, the mobile network, and so-called Over-The-Top (OTT) speech services. Although the types of transmission networks differ in these services, they all use digital codecs to transmit the speech nowadays.

The landline telephone network, retroactively also called the Plain Old Telephone Service (POTS), is the oldest speech transmission network. While in the early days of the telephone, the speech was transmitted as an analog signal over copper wires, landline networks mostly use digital technology for transmission nowadays. One of the most widely used digital codecs in this type of network is the one described in ITU-T Recommendation G.711 (1988). While the human hearing capabilities range from around 20–20,000 Hz, this codec transmits speech signal in the frequency range between 300–3,400 Hz, which is also called Narrowband (NB). This small bandwidth results in a muffled and colored speech, for which a typical landline telephone call is known. In the newer codec standardized in ITU-T Recommendation G.722 (2012), speech is transmitted in Wideband (WB), which covers the frequency range of 100–7,000 Hz.

Mobile networks have become the main way to perform telephone conversations in more recent history. Here, the speech has to be transmitted via radio waves to the nearest cellular base station. This form of speech transmission entails its own types of impairments, like transmission errors and hand-overs of calls between cell towers. The most common codecs in these types of networks are AMR-NB (3GPP Technical Specification 26.071, 1999) for narrowband communication and AMR-WB (3GPP Technical Specification 26.171, 2001; ITU-T Recommendation G.722.2, 2003) for wideband communication. Most recently, also the Enhanced Voice Services (EVS) codec (3GPP Technical Specification 26.441, 2014) has been standardized, which supports Superwideband (SWB) at 50–14,000 Hz and Fullband (FB) at 20–20,000 Hz speech transmission.

With the rise of the internet, so-called Over-The-Top speech services have become common. These services provide third-party speech transmission connections over the internet without controlling and having access to the underlying network. In contrast to landline telephony and mobile networks, they are generally not directly connected to the Public Switched Telephone Networks (PSTN). OTT speech communication solutions range from services mostly targeted to businesses (e.g., Zoom, Webex, Microsoft Teams), to social media apps that include voice communication (e.g., WhatsApp, Facebook Messenger, Instagram), to dedicated speech and video communication applications (e.g., Skype, Facetime). Codecs that are used with these services include EVS (3GPP Technical Specification 26.441, 2014) and OPUS (RFC 6716, 2012), which both provide SWB and FB speech transmission. Because these types of services only require basic internet protocol functionality and transmit the speech data on the application layer, there is no standard protocol for OTT services. However, one standard that has developed as a framework for many web-based speech and video telephony services is WebRTC (W3C Recommendation WebRTC, 2021), which provides real-time communication on web-based platforms.

Nowadays, landline telephony, mobile telephony and OTT services rely mostly on the packet-switched Voice over Internet Protocol (VoIP) transmission technique. While previously, in circuit-switched networks, a dedicated communication channel was available for each call, there are no such guarantees in internet-based telephony. With VoIP, the speech is coded and transmitted in small data packets. Each packet might take a different route through the network until it arrives at its destination. This entails that the order of arrival might not be the same order the packets were sent in, and disruptions in the network might affect only some of the packets. At the receiving end, the packets have to be put into the right order and decoded to recreate the speech signal. Because of the uncertainty in the routing process, a "jitter buffer" is employed, which handles packets that arrive in the wrong order, are too late, are duplicates, or, in certain cases, even request a retransmission of a lost packet. If packets arrive more slowly than expected, the jitter buffer can use the buffered packets to stretch the available speech signal and increase the size of the buffer. It can also speed up the playback of the buffered packets to decrease the size of the buffer. To keep the buffer to a reasonable length, the jitter buffer might also drop a packet when it arrives too late, or is corrupted. This *packet loss* leads to a short segment of speech missing, which has to be bridged. In the simplest case, the codec inserts "zeros" (i.e., silence) into the speech output, resulting in an audible cut in the speech. However, most modern codecs deploy some sort of

Packet Loss Concealment (PLC), which tries to approximate the lost speech with the previous (and sometimes next) packets (Lecomte et al., 2015). While these concealment methods work well when a single discarded packet has to be concealed, the algorithms cannot bridge longer gaps of missing packets, resulting in robotic-sounding artifacts in the speech.

Increasing the buffer size makes the connection robust against packet loss. However, it also increases the overall delay with which the speech arrives at the receiver's ear. While a high transmission delay is not audible, it affects the interaction between the conversation partners, making it harder to take turns properly. This results in a slower conversation with more unwanted interruptions.

2.2 Speech Quality and Assessment

The previous section briefly introduced current speech transmission networks and technology. Current speech transmission technologies mostly provide intelligible and clear speech transmission. However, fundamental information used for the planning, monitoring, and evaluation of speech transmission systems is the user's perception of the *quality* of the system. In ITU-T Recommendation P.10 (2017), the ITU-T defines *speech quality* as:

> "The quality of spoken language as perceived when acoustically displayed. The result of a perception and assessment process, in which the assessing subject establishes a relationship between the perceived characteristics, i.e., the auditory event, and the desired or expected characteristics."

Furthermore it defines the *speech transmission quality* as:

> "The speech quality related to the performance of a communication system, in general terms."

Jekosch (2005) describes the quality perception process as a judgment based on a comparison between the perceived quality features and the listener's internal desired quality features. ITU-T has standardized methods for the subjective determination of the transmission quality (ITU-T Recommendation P.800, 1996). There, a clear distinction between *listening-opinion* and *conversation-opinion* tests is made. During a listening-opinion test, participants usually listen to short speech samples and rate their perceived quality on a 5-point Absolute Category Rating (ACR) scale, shown in Table 2.1. Averaging the ratings of a sample over all participants produces the Mean Opinion Score (MOS). Usually, there are multiple speech samples for a particular set of impairments, whose ratings can be combined into a MOS per condition. Other types of listening-opinion test methods defined in ITU-T Recommendation P.800 (1996) are the Degradation Category Rating (DCR), the Comparison Category Rating (CCR), and the Threshold Method.

Table 2.1 5-point Absolute Category Rating (ACR) scale recommended by ITU-T Recommendation P.800 (1996).

Label	Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

ITU-T Recommendation P.800.1 (2016) defines the terminology of the MOS in different contexts. For audio applications, the MOS is defined for listening-only, conversational, and talking contexts. Also, a MOS can result from a subjective experiment, predicted by an objective model, or it can be estimated with a parametric model. These different variants of MOS are shown in Table 2.2.

Table 2.2 Different identifiers for MOS as defined by ITU-T Recommendation P.800.1 (2016)

	Listening-only	Conversational	Talking
Subjective	MOS-LQS	MOS-CQS	MOS-TQS
Objective	MOS-LQO	MOS-CQO	MOS-TQO
Estimated	MOS-LQE	MOS-CQE	MOS-TQE

The listening-opinion test method produces one averaged quality judgment on a single rating scale for each rated sample. While this type of quality assessment gives an accurate overall rating of a speech sample, it does now allow for the identification of underlying causes of the degradation or for the classification of degradations. Wältermann (2012) stated that the perceived quality could be represented as a multidimensional coordinate system with different perceptual dimensions. In Wältermann et al. (2010) the three main listening-quality dimension for NB and WB were defined as *Discontinuity*, *Noisiness*, and *Coloration*, and a fourth dimension *Loudness* was introduced in Côté et al. (2007). A different set of 6 perceptual dimensions were identified by Sen and Lu (2012) and standardized in ITU-T Recommendation P.806 (2014), consisting of dimensions for slow-varying degradations, fast-varying degradations, low-frequency coloration, high-frequency coloration, level of background noise, and variability of background noise.

Considering the practical execution of an experiment, the listening-opinion test method can provide many data points (i.e., user ratings) that very accurately and reproducibly describe any given condition. Because one speech sample is oftentimes no longer than 10 seconds, many quality ratings can be obtained from a single participant, and it is nowadays also used in crowdsourcing (ITU-T Recommendation P.808, 2021). Because of this, many quality prediction models predict the listening-only quality (ITU-T Recommendation P.862, 2001; ITU-T Recommendation P.863, 2014).

Despite its predominant use as a quality indicator for speech communication services, the listening quality has drawbacks. Since participants are only listening to a speech signal, impairments to speaking (e.g., due to echo) or to the interactivity of a conversation

(e.g., due to transmission delay) cannot be captured by this assessment method. Also, the participants in listening-opinion tests might focus on degradations more than in a real conversational scenario, yielding a difference in the rating of the perceived quality.

2.3 Conversational Quality

In ITU-T Recommendation P.10 (2017), *conversational quality* is defined as

> "The quality of a bi- or multidirectional conversation as perceived by a communication partner."

In contrast to the speech quality obtained in a listening-only scenario, this quality definition is more focused on the user's experience rather than the user's perception of a speech signal. The ITU-T Recommendation P.800 (1996) on methods for subjective determination of transmission quality defines both listening-opinion and conversation-opinion tests, stating:

> "Listening-opinion tests are not expected to reach the same standard of realism as conversation tests, and the restrictions are therefore less severe in some respects; but the artificiality that has to be accepted brings with it a necessity for strict control of many things which in conversation tests are allowed to find their own equilibrium."

The ITU-T standardizes the methods and procedures for conducting conversation tests and evaluating conversational quality in ITU-T Recommendation P.805 (2007). For each test, two participants sit in separate, sound-proof rooms. They are connected with each other through a transmission chain over a telephone network simulation and are asked to hold a conversation. Afterwards, they give their opinion of the quality on different rating scales. Optionally, a recording of the conversation can be made for later analysis. A typical test setup is illustrated in Figure 2.1.

The test conditions may be introduced by the simulated telephone network between the two test participants or by the environment (e.g., a loudspeaker in the test room playing environmental noise). The participants may be untrained "naïve", experienced, or experts in the field.

2.3.1 Standardized Conversation Tests

The ITU-T Recommendation P.805 (2007) provides instructions on what types of conversations the participants of a conversation test should have. For untrained subjects, the tasks presented should be of cooperative nature and result in semi-structured conversations. They should allow for interruptions by the subjects, they should be easily learned, provide intrinsic motivation, and there should exist a sufficient number of equivalent

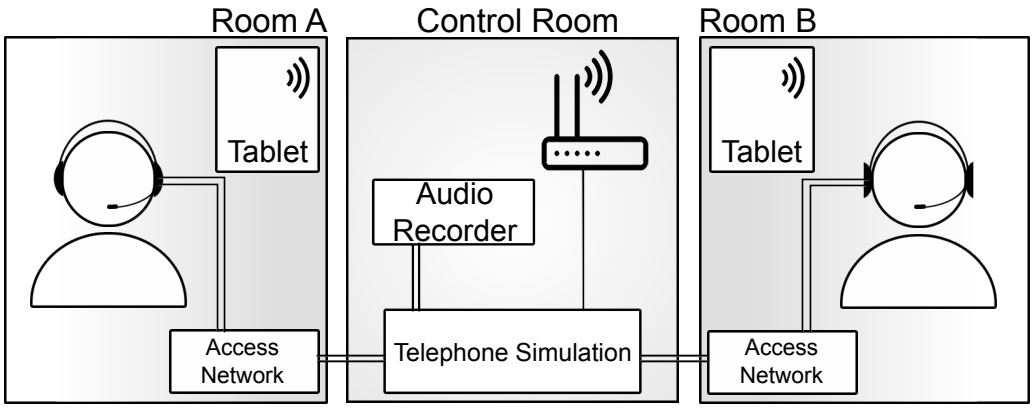

Fig. 2.1 Setup of a conversation experiment as described in ITU-T Recommendation P.805 (2007). Two participants are conversing in two different rooms over a telephone simulation. The audio is usually recorded for later analysis and quality ratings are recorded digitally or with paper.

versions of the task. Generally, the tasks provide two different sets of information to each participant with the goal of exchanging the information in a conversation. Already standardized conversational tasks are referenced in the recommendation. They are split into tasks that meet the requirements for a realistic, rich two-way conversation and less realistic, more competitive tasks that focus on being very interactive.

While there are many conversational tasks referenced in the recommendation, this thesis focuses on two tests in particular: the Short Conversation Test (SCT) and the Random Number Verification (RNV) task.

Short Conversation Test

The Short Conversation Test (SCT) is a set of conversation tasks in which participants perform typical telephone conversations in a role-play type of fashion (Möller, 2000; ITU-T Contribution SG12-C35-E, 1997). The scenarios include booking hotel rooms, ordering at a pizza delivery service, railway inquiries, and rental of cars and apartments. In every scenario, one participant is the caller with a certain request, and the other person is the callee (usually playing the role of an institution or company). The caller has a specific goal (e.g., buying plane tickets) with some additional restrictions on that request (e.g., the destination and date of the flight). Both participants have fields that need to be filled with information by the end of the conversation and certain information that the interlocutor might request. Lastly, the caller is given an improvisational question (e.g., asking for a special deal) to which the callee does not have a specific answer. This incites some improvisational dialogue that is not pre-determined by the conversation scenario.

Each conversation takes approximately 2–4 minutes, and the scenarios are balanced so that both participants have roughly the same speaking duration. Also, the information exchanged in the conversations (i.e., names, places, numbers, etc.) has been designed to incite meta communication. For example, names are often spelled in an unusual manner, providing the possibility for clarification requests. Two examples of SCT scenarios are given in Appendix A.

Random Number Verification Task

The Random Number Verification (RNV) task was first described in Kitawaki and Itoh (1991). In this task, participants receive six rows of six numbers. While the numbers are mostly the same for the two interlocutors, there are one or two numbers in each row that differ between the two versions. The participants are instructed to take turns reading the rows to each other as fast as possible while trying to correct any instances where the numbers do not match.

While this type of scenario is rather unrealistic in nature, it produces highly interactive conversations with many speaker alternations in a short amount of time. Each RNV conversation takes approximately 30 seconds up to 2 minutes. An example of an RNV scenario is given in Appendix B.

2.3.2 Multidimensional Conversation Quality

Based on the perceptional dimension analysis of speech quality in listening scenarios done by Wältermann (2012), the concept of the multidimensional quality space was extended to conversation scenarios. According to Köster and Möller (2014) and based on Guéguin et al. (2008), the conversational quality can be divided into three distinct *phases*: the listening phase, the talking phase, and the interaction phase. This separation is based on the different states of the conversation, where each interlocutor is either listening, speaking, or engaging in a speaker alternation, which produces an interaction between the conversation partners.

The quality of the listening phase can be determined in a subjective listening quality test, as described in Section 2.2. As described by Wältermann (2012) and Côté et al. (2007), the four perceptual dimensions of the listening phase "coloration", "noisiness", "discontinuity", and "sub-optimum loudness" can be assessed. The talking phase is targeted to impairments that degrade the own voice. Especially in systems with talker echo, where the feedback of one's own voice is audible, the talking quality can be degraded. For this phase, the two dimensions, "Impact of one's own voice" and "Degradation of one's own voice", have been identified (Köster and Möller, 2015). The interaction phase is described as the alternation between talking and listening (Möller et al., 2017). Köster and Möller (2015) identified one perceptual dimension "interactivity" for this phase. An overview of the phases, their perceptual dimensions, and possible impairments for each dimension is shown in Table 2.3.

ITU-T Recommendation P.804 (2017) standardizes a test method for conversational quality that includes all three phases and their dimensions. In this test procedure, two participants converse with each other over a simulated telephone line as described in ITU-T Recommendation P.805 (2007). However, for each condition, the participants rate three sessions. In the first session, the participants perform a short conversation test, after which they rate all seven perceptual dimensions. In the second session, each of the two participants reads a sentence to the other one, after which they rate the quality of the speaking and listening phases, respectively. This way, each participant is rating

Table 2.3 Overview of the seven perceptual quality dimension for a conversational situation from ITU-T Recommendation P.804 (2017)

Phase	Perceptual dimension	Description	Possible Source
Listening phase	Noisiness	Background noise, circuit noise, coding noise	Coding, circuit, or background noise
	Discontinuity	Isolated and non-stationary distortions	Packet loss
	Coloration	Frequency response distortions	Bandwidth limitations
	Loudness	Important for the overall quality and intelligibility	Attenuation
Speaking phase	Impact of one's own voice	How is the back-coupling of one's own voice perceived	Sidetone and echo
	Degradation of one's own voice	How is the back-coupling of one's own voice degraded	Frequency distortions of the sidetone and echo
Interaction phase	Interactivity	Delay and disrupted interaction	Delay

the speaking phase as well as the listening phase in this session. In the third session, the participants perform a RNV task and rate the quality dimension of the interactivity phase.

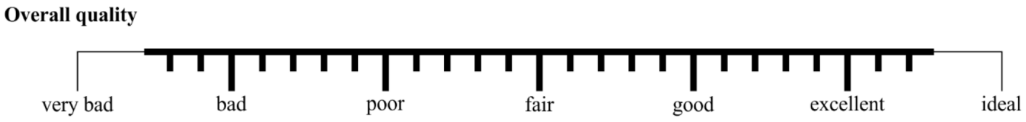

Fig. 2.2 Extended Continuous Scale of the overall conversational quality as shown in ITU-T Recommendation P.804 (2017)

Instead of the ACR scale, this test method uses the Extended Continuous Scale (ECS) (as shown in Figure 2.2), which is developed by Bodden and Jekosch (1996). In comparison to the discrete ACR scale, the ECS provides the possibility to rate in between the given categories. This makes the scale more sensitive and conveys to the user that the categories are equidistant. Also, the scale adds the "overflow areas" *very bad* and *ideal* to each end of the scale. These extreme rating options can encourage participants to use the rest of the scale more extensively, reducing scale-end effects.

A comparison of the ACR and ECSwas published by Köster et al. (2015), that used the two scales for speech quality rating. For better comparison between the two scales and to make the ECS more widely used, a transformation function is given to map extended continuous MOS ratings into estimates of a 5-point ACR MOS:

$$\widehat{MOS}_{ACR} = -0.0262 \cdot MOS_{EC}^3 + 0.2368 \cdot MOS_{EC}^2 + 0.1907 \cdot MOS_{EC} + 1 \qquad (2.1)$$

where $\widehat{MOS}_{ACR}$ is the estimation of the ACR MOS based on the extended continuous ratings and MOS_{EC} is the MOS as acquired by the Extended Continuous Scale as shown in Figure 2.2.

2.3.3 Delay and Interactivity

Transmission delay has long been a focus in speech quality research, as pure delay is not an audible impairment and only affects the interactivity of a conversation. ITU-T Recommendation G.114 (2003) defines acceptable values for one-way end-to-end transmission delay. Generally, delay levels between $0\,ms$ and $150\,ms$ are considered acceptable for most user applications. Delay levels of over $200\,ms$ are only considered acceptable for inter-regional calls, and a one-way transmission delay of over $400\,ms$ is considered not acceptable. While for circuit-switched networks, these limits can be reasonably met, current mobile connections and OTT speech services have limited control over the IP network used to rout the speech traffic, and extensive signal processing on the users' devices further increases delay levels. More recent work points toward an even higher acceptability threshold for transmission delay in everyday conversations, stating that transmission delay of over $400\,ms$ has no dramatic effect on the conversational quality MOS (Egger et al., 2010; Raake et al., 2013). The degradation of perceived quality due to transmission delay increases the time it takes the speech signal to arrive at the other end and thus creates different *conversation realities* at both ends (Hammer, 2006). Events that happen at one end of the conversation (interruptions, pauses, speaker changes) are not necessarily happening at the other end as well. One example is illustrated in Figure 2.3.

It has been shown that the impact of end-to-end transmission delay on the conversation does not only depend on the transmission time but also on the interactivity of the conversation (Kitawaki and Itoh, 1991; Hammer et al., 2004). For high delay levels (above about $800\,ms$), a significant difference in conversational quality MOS is observed, depending on whether the participants carry out conversations with low interactivity (e.g., SCT) or high interactivity (e.g., RNV) (Egger et al., 2010).

In order to analyze the interactivity of a conversation and its impact on conversational quality, quantitative measurements have to be defined to objectively compare conversation scenarios and special events in the conversation (i.e., interruptions, speaker alternations, etc.). These conversational parameters are based on the work of Brady (1968), in which conversations of two persons A and B are split up into four states:

- **Mutual Silence (MS)**: Moments in which neither person A nor person B is talking
- **Speaker A (SA)**: Moments in which only person A is talking and person B is silent
- **Speaker B (SB)**: Moments in which only person B is talking and person A is silent
- **Double Talk (DT)**: Moments in which both person A and person B are talking

Generally, a conversation resides mostly in the states SA and SB, with a sizeable amount of silence (MS). Situations in which both speaker A and speaker B are talking (DT) occur mostly between speaker changes, where a turn is handed over with short overlaps.

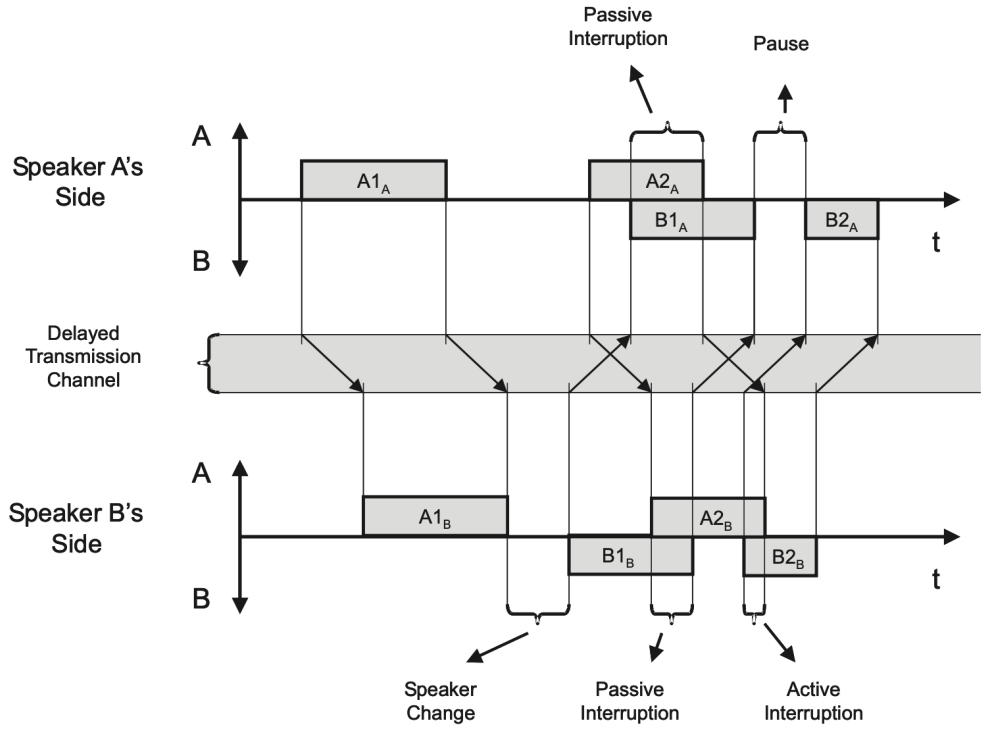

Fig. 2.3 An illustration of a delayed conversation with different events happening at each end (Hammer, 2006)

2.3.4 Parametric Conversation Analysis

Based on the analysis method from Brady (1968), and Lee and Un (1986), a multitude of conversational parameters have been defined that try to capture some aspects of the interactivity of a conversation. These parameters have been formalized in the Parametric Conversation Analysis (P-CA) framework by Hammer (2006). The P-CA is based on the four conversational states SA, SB, MS, and DT, and parameters are mostly described by a series of transitions between these states. For that, a Markov model with these states is used, defining transition between the SA and SB states and the MS and DT states (visualized in Figure 2.4). The reasoning for these transitions is that a speaker change always involves either an interruption (i.e., two speakers speaking at the same time) or a period of mutual silence.

In practice, a P-CA can be executed semi-automatically if the conversation is recorded with two the speakers on separate audio channels. Then, an automated Voice Activity Detection (VAD) can be performed to determine the conversational states for each point in the conversation. Based on these states, parameters and metrics can be derived that are used to analyze the interactivity of conversation and the impact of transmission delay.

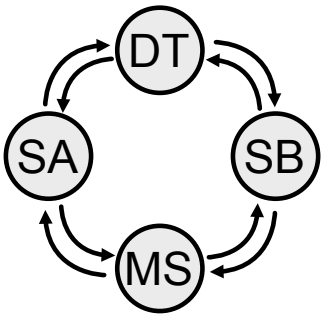

Fig. 2.4 Markov model of a conversation based on the four conversation states. There are no transitions defined between the states SA and SB - all speaker changes either go through the states MS and DT.

The following sections will give an overview of conversational metrics and parameters that are the most useful for analyzing interactivity in telephone conversations, as well as simulated conversations. Some of the parameters are specifically intended for use in conversations with transmission delay. An overview of the conversational parameters is given in Table 2.4.

Table 2.4 Overview of the conversational parameters extracted by the Parametric Conversation Analysis, split by general parameters and parameters that are specific for conversations with transmission delay.

Parameter Name	**Description**	**Unit**
length	Length of the conversation	seconds
State probability MS	Ratio of conversation in state MS	probability
State probability SA	Ratio of conversation in state SA	probability
State probability SB	Ratio of conversation in state SB	probability
State probability DT	Ratio of conversation in state DT	probability
Sojourn time MS	Time conversation sojourns in state MS	seconds
Sojourn time SA	Time conversation sojourns in state SA	seconds
Sojourn time SB	Time conversation sojourns in state SB	seconds
Sojourn time DT	Time conversation sojourns in state DT	seconds
Speaker Alternation Rate	Number of alternations each minute	alternations / minute
Interruption Rate	Number of SI each minute	interruptions / minute
Active Interruption Rate	Number of active SI each minute	interruptions / minute
Passive Interruption Rate	Number of passive SI each minute	interruptions / minute
Double Talk Rate	Number of NI each minute	double talk / minute
Pause Rate	Number of pauses each minute	pauses / minute
Turn time	Average duration of a turn	seconds
Turn count	Number of turns in the conversation	turns
Conversational Temperature	How heated a conversation is	degrees
SAR_C	SAR corrected for decrease due to delay	alternations / minute
Unintended Interruption Rate	Interruptions not intended by the speaker	interruptions / minute
Intended Interruption Rate	Interruptions intended by the speaker	interruptions / minute

State Probabilities and Sojourn Times

The state probabilities and sojourn times are the most basic ones of the conversational parameter (Brady, 1968). The state probabilities are defined as the ratio of the conversation that can be categorized in a certain state. At the same time, it is also the unconditional probability at which the state will occur at any time in the conversation.

The sojourn times are defined as the average time in seconds the conversation will sojourn in one of the four states. The sojourn time of SA and SB then represent the average length of the utterances of the respective speaker, while the sojourn time of DT and MS yields basic information about how interactive the conversation is. The sojourn times of an average conversation are used in ITU-T Recommendation P.59 (1993) for the definition of artificial conversational speech.

For delayed conversations, changes in the state probabilities and sojourn times can be observed (Egger et al., 2010). Especially the state probabilities for MS and DT increase for higher one-way transmission delays. However, the initial level and the rate of the increase differ between lowly and highly interactive conversations. Thus, these state probabilities alone cannot be used to determine the general interactivity level and the impact of transmission delay.

Speaker Alternation Rate

The Speaker Alternation Rate (SAR) is one of the main conversational parameters used to determine the interactivity of a conversation (Egger et al., 2010; Raake et al., 2013; Hammer et al., 2004). It directly corresponds to the interactivity of a conversation and can be easily calculated. It is defined by Hammer (2006) as the number of transitions between the states SA and SB (i.e., SA-MS-SB, SB-MS-SA, SA-DT-SB, and SB-DT-SA) divided by the length of the conversation in minutes (see Equation 2.2). Short occurrences of double talk of a continued utterance of a speaker (SA-DT-SA or SB-DT-SB, e.g., during a backchannel) are not considered a speaker alternation.

$$SAR = \frac{\#SA\text{-}MS\text{-}SB + \#SB\text{-}MS\text{-}SA + \#SA\text{-}DT\text{-}SB + \#SB\text{-}DT\text{-}SA}{DUR} \tag{2.2}$$

Because the SAR directly reflects the interactivity of a conversation, it can be used as an indicator of how much the conversation is impacted by delay (Hammer et al., 2004). However, the SAR of conversations with high interactivity (i.e., with high SAR values at $0\,ms$ delay) starts to be impacted at lower delay levels and more strongly. Inversely, the SAR of conversations with low interactivity starts to drop at higher delay levels, and the impact is not as pronounced. For this reason, a direct relationship between the perceived quality and the SAR of a conversation cannot be drawn.

In Egger et al. (2012), the Corrected Speaker Alternation Rate (SARc), a delay-based extension of the speaker alternation rate, is proposed. Other than the SAR, it takes into account the added transmission time due to the delay and thus captures the interactivity of a conversation independently of the transmission delay. As described in Section 2.3.3, transmission delay only affects the arrival of the speech of one's interlocutor and thus results in different *conversation realities*. Speaker alternations that happen at one end of the telephone conversation do not necessarily happen at the other end. The definition

of the SARc is thus dependent on the side of the conversation on which the interactivity is measured. In Schoenenberg (2015), the SARc for person A in the conversation is defined as:

$$SAR_C^A = \frac{(\#SA\text{-}MS\text{-}SB^A + \#SB\text{-}MS\text{-}SA^A + \#SA\text{-}DT\text{-}SB^A + \#SB\text{-}DT\text{-}SA^A)}{DUR - (\#SA\text{-}MS\text{-}SB^A \cdot 2 \cdot Ta)} \quad (2.3)$$

Here, SAR_C^A is the SARc from the perspective of speaker A and $\#SA\text{-}MS\text{-}SB^A$ denotes the number of transitions from the state SA to MS to SB (i.e., a speaker alternation with mutual silence as the transition state) from the perspective from speaker A. The denominator counts all speaker alternation occurrences from the viewpoint of speaker A. In the divisor, the length of the conversation (DUR) in minutes is reduced by the number of speaker changes from person A to B with silence in between, multiplied by two times the one-way transmission delay Ta in minutes. The reasoning of Schoenenberg (2015) is that for person B's response to return to person A, the speech is delayed by twice the amount of one-way transmission delay. Thus, the duration of the conversation is reduced by the full two-way transmission delay overhead of all speaker changes from person A to person B, increasing the calculated speaker alternation rate. The calculation of the SAR_C^B from the perspective of speaker B is analogous to Equation 2.3, but subtracts the transitions $\#SB\text{-}MS\text{-}SA^B$ from the duration.

The SARc results in a relatively stable parameter measuring a conversation's interactivity, independent of the transmission delay. However, the definition assumes that the interlocutors of a conversation with transmission delay do not alter their turn-taking behavior.

Interruptions and Double Talk

In Schoenenberg (2015), an interruption can be classified into a Successful Interruption (SI), where a speaker is interrupted by the interlocutor resulting in a speaker alternation, and into Non-successful Interruption (NI), where the interruption does not end the turn of the current speaker (e.g., during a backchannel). The Interruption Rate (IR) is defined as the number of SI per minute, while the Double Talk Rate (DTR) is defined as the number of NI per minute.

Hammer (2006) additionally defines *active* and *passive* interruptions. In an active interruption, a participant interrupts the currently active speaker, while in a passive interruption, the speaker under study is interrupted. For telephone conversations without transmission delay, an active disruption on one side of the conversation results in a passive interruption on the other end. For conversations with transmission delay, the number of active and passive interruptions differ between the two sides. Based on the active and passive interruption, the corresponding rates Active Interruption Rate (AIR) and Passive Interruption Rate (PIR) are defined.

$$AIR^A = \frac{\#SB\text{-}DT\text{-}SA^A}{DUR} \quad (2.4)$$

$$PIR^A = \frac{\#SA\text{-}DT\text{-}SB^A}{DUR} \tag{2.5}$$

Equations 2.4 and 2.5 define the active and passive interruptions from the perspective of speaker A. The active and passive interruptions from the perspective of speaker B require the opposite state transitions (SA-DT-SB for the AIR and SB-DT-SA for the PIR).

In VoIP conversations with transmission delay, not every interruption is intended by the interrupting person. This is due to the fact that a delayed transmission of an utterance may arrive after the interlocutor has already started their turn. Egger et al. (2010) defines an Unintended Interruption Rate (UIR) and an Intended Interruption Rate (IIR) that reflect the unintended nature of an interruption that may arise due to transmission delay:

> "The UIR is based on the rate of passive interruptions that interlocutors experience during a conversation. However, it counts only those passive interruptions which were actually caused by delay, thereby excluding all occurrences of active interruptions that were deliberately caused by a speaker."

For the calculation of the UIR and IIR, access to the conversation on both ends is needed. For each passing interruption occurring on one side, the situation of the same interrupting utterance has to be examined from the other perspective of the conversation. If the interruption was present from the perspective of the actively interrupting person, it is counted as an intended interruption. If the other perspective shows no interruption, it is classified as *unintended*, because on their end of the conversation there was no interruption.

Pauses

Hammer (2006) defines a pause as silence between the speaking states of the same speaker. This corresponds to the state transitions SA-MS-SA and SB-MS-SB. The Pause Rate (PR) is thus defined as:

$$PR = \frac{\#SA\text{-}MS\text{-}SA + \#SB\text{-}MS\text{-}SB}{DUR} \tag{2.6}$$

where $\#SA\text{-}MS\text{-}SA$ are the number of pauses by speaker A, $\#SB\text{-}MS\text{-}SB$ are the number of pauses by speaker B, and DUR is the duration of the conversation in minutes.

Conversational Temperature

The conversational temperature is a metric of conversational interactivity defined by Reichl and Hammer (2004) and further elaborated in Hammer et al. (2005). The conversational temperature is calculated based on the sojourn times of the four conversational states, and it is implicitly defined by three axioms. The axiom of "limiting behavior" limits the range of the temperature between 0 and infinity, making the most non-interactive conversation have a temperature of 0°. The second axiom of "normalization" scales

the temperature of the conversation based on the sojourn times of an abstract "average conversation". The conversational temperature of this average conversation was scaled to "room temperature" (21.5°). The third axiom of "monotonicity/first-order behavior" describes that decreasing sojourn times of any of the conversational states lead to an increase in the conversational temperature.

Given these three axioms, the conversational temperature τ can be estimated with a least squares estimation with the following equation:

$$\hat{\tau} = \arg\min_{\tau} \sum_{I} (t_I^{Ref} \cdot exp(\frac{\tau^{Ref}}{\tau} - 1) - t_I)^2 \tag{2.7}$$

Here, I represents a conversational state (SA, SB, MS, or DT), t_I is the sojourn time of that state, t_I^{Ref} is the reference sojourn time of that state, and τ^{Ref} is the reference value of an average conversation set to 21.5°.

2.4 Parametric Quality Prediction

Parametric quality prediction models use characteristics of the transmission system to estimate the expected overall quality. Network providers may use these models to plan a new transmission network, or they may tune the parameters of their already existing network based on the parametric prediction models. Parametric models like the Bellcore TR model (Cavanaugh et al., 1976) and the OPINE model (Osaka and Kakehi, 1986) were the first models to be used to predict the quality of POTS networks. In ITU-T Recommendation P.564 (2007), a network monitoring model is standardized that predicts the one-way listening-only quality based on speech packet-level information.

2.4.1 E-model

The most widely used network planning model and the only parametric model standardized by the ITU-T that predicts conversational quality (MOS-CQE) is the E-model (ITU-T Recommendation G.107, 2015). The model was a result of merging different opinion models and was initially standardized by the European Telecommunication Standard Institute (ETSI) (Johannesson, 1997). The E-model covers the effects of attenuation, circuit and ambient noise, non-optimum sidetone, talker and listener echo, pure delay, as well as digital coding at different bitrates. While the E-model predicts speech quality only for Narrowband (NB) communication, it was extended in ITU-T Recommendation G.107.1 (2015) for Wideband (WB) communication, and in ITU-T Recommendation G.107.2 (2019) for Superwideband (SWB) and Fullband (FB) communication scenarios. Unlike many other speech quality models, the E-model predicts the quality of a conversation and not the listening quality. Thus, it is able to account for delay and the interactivity of a conversation.

The main output of the E-model is the transmission rating R which describes the overall quality experienced by a communication partner during conversations over a telephone channel, which shows the characteristics as defined by the parameters of the

model. The E-model assumes that impairments are independent of each other and can be quantified in terms of impairment factors on the transmission rating scale R. This can be done by subtracting the impairment factors from a maximal transmission rating, which is given by the basic signal-to-noise ratio of the connection.

The different versions of the E-model have different maximal transmission ratings and formulae for the impairment factors. Each version will be described in the following sections, and only formulae important for the impairments considered in this thesis will be explained in detail.

Narrowband E-model

The transmission rating R for the narrowband E-model can be calculated with the following formula[1]:

$$R = Ro - Is - Id - Ie,eff + A \tag{2.8}$$

The Ro term represents the basic signal-to-noise ratio of the connection in the transmission rating scale, with 100 being the maximum value achievable. Included in this term are noise sources at the sending side, receiving side, circuit noise, and the noise floor at the receiving side. The impairment factor Is is the sum of all impairments that occur simultaneously with the voice transmission, like non-optimum sidetone and quantizing distortions. Id is the impairment factor that represents impairments caused by the delay of voice signals. The factor Ie,eff is the *effective* equipment impairment factor that considers impairments due to codecs, as well as bursty packet loss. A is an advantage factor that increases the transmission rating due to some advantage, like mobility or access to hard-to-reach locations (e.g., via satellite connections).

The overall transmission rating R can be converted into a rating on the 5-point overall conversational quality scale using an S-shaped function using $Rx = R$:

For Rx < 0:

$$MOS_{CQE} = 1$$

For 0 < Rx < 100:

$$MOS_{CQE} = 1 + 0.035Rx + Rx(Rx - 60)(100 - Rx) \cdot 7 \cdot 10^{-6}$$

For Rx > 100:

$$MOS_{CQE} = 4.5 \tag{2.9}$$

This calculation considers that in a typical subjective experiment, the maximum average rating commonly being observed is around 4.5 on the MOS_{CQE} scale ranging from 1 to 5. Setting the values for Rx to different factors of R, the conversion of the wideband and fullband E-model can be performed.

[1] To be in line with ITU-T Recommendation G.107 (2015), ITU-T Recommendation G.107.1 (2015), and ITU-T Recommendation G.107.2 (2019), no subscript will be used for E-model formulas.

The Id impairment factor consists of the impairments that occur due to the delayed transmission of the voice signal. Specifically, it is the sum of impairments that occur due to talker echo ($Idte$), listener echo ($Idle$), and absolute transmission delay (Idd):

$$Id = Idte + Idle + Idd \tag{2.10}$$

The Idd impairment factor is calculated based on the absolute one-way transmission delay Ta (given in milliseconds) and two interactivity parameters sT and mT:

For $Ta \leq mT$:

$$Idd = 0$$

For $Ta > mT$:

$$Idd = 25\{(1 + X^{6 \cdot sT})^{\frac{1}{6 \cdot sT}} - 3(1 + [\frac{X}{3}]^{6 \cdot sT})^{\frac{1}{6 \cdot sT}} + 2\} \tag{2.11}$$

with:

$$X = \frac{\log \frac{Ta}{mT}}{\log 2} \tag{2.12}$$

Here, mT denotes the minimal perceivable delay in milliseconds, and sT describes the delay sensitivity of the users of the system. The minimal perceivable delay parameter mT shifts the Idd impairment curve to the right and sets the impairment to 0 if the absolute delay Ta is smaller than or equal to mT. The delay sensitivity parameter sT changes the slope of the logistic Idd function, with higher sT values resulting in a steeper increase of Idd. Table 2.5 shows the three delay sensitivity classes recommended by ITU-T Recommendation G.107 (2015).

Table 2.5 Delay sensitivity classes for different use cases as recommended by ITU-T Recommendation G.107 (2015).

Class	sT	mT (ms)	Use case
Default	1	100	Used when conversations are very interactive or targeted delay requirements are unknown.
Low	0.55	120	Applicable in cases where users have a low sensitivity to delay.
Very low	0.4	150	Applicable in cases where users have a very low sensitivity to delay.

While not being part of the recommendation, Raake et al. (2013) provide two mapping functions, calculating the mT and sT values based on the SARc (defined in Equation 2.3):

$$mT = 436.02 - 71.56 \cdot \log\left(16.76 + SAR_C\right) \tag{2.13}$$

$$sT = 0.246 + 0.02 \cdot \exp\left(0.053 \cdot SAR_C\right) \tag{2.14}$$

To obtain these equations, the SARc of individual conversations were averaged over the different delay settings into a scenario-test-specific mean SARc. For each of these values, the sT and mT values have been plotted, and the respective functions have been fitted using a least-squares curve fitting.

The effective equipment impairment factor Ie, eff combines the effects of coding (Ie) with impairments due to packet loss:

$$Ie, eff = Ie + (95 - Ie) \cdot \frac{Ppl}{\frac{Ppl}{BurstR} + Bpl} \tag{2.15}$$

where Ie is the equipment impairment factor that is specific for the codec and bitrate used, Ppl is the packet loss percentage (between 0 and 20 %), and Bpl is the packet loss *robustness* factor, that is also specific for the codec used. Thus, codecs that employ a form of Packet Loss Concealment (PLC) can be taken into account. Values for Ie and Bpl are listed in Appendix I of the ITU-T Recommendation G.113 (2007). The $BurstR$ is the burst ratio that defines how bursty the packet loss is to be expected. It is defined as the average length of observed bursts in an arrival sequence divided by the average length of bursts expected for the network under "random" loss. A burst ratio of 1 then denotes randomly distributed loss, while a burst ratio of 2 means that packet loss bursts are, on average, twice as long as with the same Ppl under random loss.

The $BurstR$ can be modeled as a 2-state Markov model with the transition probability p from the *found* state to the *loss* state and probability q from the *loss* state to the *found* state:

$$BurstR = \frac{1}{p+q} = \frac{\frac{Ppl}{100}}{p} = \frac{\frac{1-Ppl}{100}}{q} \tag{2.16}$$

The ITU-T Recommendation G.107 (2015) defines, that values of $BurstR > 2$ are only valid for Ppl values smaller than 2 %.

Wideband E-model

The E-model was extended in Möller et al. (2006) and Raake et al. (2010) to be able to capture wideband telephony and was standardized in ITU-T Recommendation G.107.1 (2015). Similar to the narrowband version of the E-model, the wideband transmission rating R can be defined as:

$$R = Ro, WB - Is, WB - Id, WB - Ie, eff, WB + A \tag{2.17}$$

Because of the greater bandwidth available, the maximum signal-to-noise ratio Ro, WB is extended to be 129. A conversation of the wideband transmission rating scale to a 5-point ACR scale can still be performed with Equation 2.9, setting $Rx = {}^{R}/_{1.29}$.

Unlike in the narrowband version of the E-model, the wideband E-model does not include the interactivity parameters mT and sT into the calculation of the delay impairment factor Idd:

For $Ta \leq 100\,ms$:

$$Idd = 0$$

For $Ta > 100\,ms$:

$$Idd = 25\{(1 + X^6)^{\frac{1}{6}} - 3(1 + [\frac{X}{3}]^6)^{\frac{1}{6}} + 2\} \tag{2.18}$$

with:

$$X = \frac{\log(\frac{Ta}{100})}{\log 2} \tag{2.19}$$

Here, the minimal perceivable delay is intrinsically set to $100\,ms$ and the delay sensitivity to 1. Thus, unlike the narrowband version, the wideband E-model cannot predict the differences in MOS between conversations of varying interactivity.

The effective equipment impairment factor Ie, eff, WB also differs from the narrowband version, as it does not include the burst ratio:

$$Ie, eff, WB = Ie, WB + (95 - Ie, WB) \cdot \frac{Ppl}{Ppl + Bpl} \tag{2.20}$$

Here, the burst ratio is implicitly set to 1. Thus, the wideband E-model cannot predict the effects of bursty packet loss. Values for Ie, WB and Bpl are listed in Appendix IV of the ITU-T Recommendation G.113 Amendment 1 (2009).

Fullband E-model

The E-model was further extended in Mittag et al. (2018) and Möller et al. (2019), which resulted in the standardization of the fullband E-model in ITU-T Recommendation G.107.2 (2019). As there could be no audible difference found between the ratings of super-wideband and fullband coded clean speech, this version of the E-model can be used for both super-wideband and fullband scenarios. Analogous to the narrowband and wideband E-model, the basic formula is re-written as:

$$R = Ro, FB - IS, FB - Id, FB - Ie, eff, FB + A \tag{2.21}$$

The even greater bandwidth has increased the maximum R-value to 148. Because there is, as of now, no noise considered in the basic signal-to-noise ratio Ro, FB it is set to 148. Again, the conversation of the fullband transmission rating scale to a 5-point ACR scale can be performed with Equation 2.9, setting $Rx = {}^{R}/_{1.48}$.

For the delay impairment factor Id, FB, only the effects of pure delay have been defined:

For $Ta \leq 100\,ms$:

$$Idd = 0$$

For $Ta > 100\,ms$:

$$Idd = 1.48 \cdot 25\{(1 + X^6)^{\frac{1}{6}} - 3(1 + [\frac{X}{3}]^6)^{\frac{1}{6}} + 2\} \tag{2.22}$$

with:

$$X = \frac{\log(\frac{Ta}{100})}{\log 2} \tag{2.23}$$

No mT and sT parameters to reflect the conversational interactivity are included in the Idd formula of the fullband E-model as of now. Notably, in the formula in Equation 2.22, a factor of 1.48 was added to account for the larger range of the transmission rating scale in the fullband version. However, the wideband E-model does not include a similar factor (see Equation 2.18).

The Ie, eff, FB value is calculated analogous to the wideband E-model:

$$Ie, eff, FB = Ie, FB + (132 - Ie, FB) \cdot \frac{Ppl}{Ppl + Bpl} \tag{2.24}$$

As with the wideband E-model, there is no burstiness calculation included, setting the burst ratio implicitly to 1 and allowing only predictions for randomly distributed packet loss. Values for Ie, FB and Bpl are listed in Appendix V of the ITU-T Recommendation G.113 Amendment 2 (2019).

2.5 Signal-based Quality Prediction

Signal-based quality prediction models use speech signals that are either transmitted over a speech transmission system or degraded by a speech processing pipeline to estimate the perceived speech quality of these systems. Generally, they can be split into two categories: full-reference models and reference-free models.

Full-reference (sometimes called *intrusive* or *double-ended*) models make use of the speech that was degraded by the speech transmission system, as well as a reference signal that does not include any of the degradations added by the system. They are considered *intrusive*, as they require the recording of speech on both ends of the speech transmission system to be able to estimate the speech quality. These types of models time-align the two speech signals, and, based on the difference between the two signals, they produce a speech quality estimate. ITU-T defines the narrowband full-reference model PESQ (Perceptual Evaluation of Speech Quality) in ITU-T Recommendation P.862 (2001) and the current state-of-the-art super-wideband speech quality prediction model POLQA (Perceptual Objective Listening Quality Assessment) is standardized in ITU-T Recommendation P.863 (2014). While these models predict only the overall listening-quality MOS, there exist models that predict the listening dimensions as well. DIAL (Diagnostic Instrumental Assessment of Listening quality) is a double-ended model that

predicts the listening dimension of narrowband and wideband speech (Scholz, 2008; Huo, 2015). Most recently, a full-reference machine-learning-based model similar to Non-instrusive Speech Quality Assessment (NISQA) was developed, that predicts the four listening-dimensions (Mittag and Möller, 2020).

Reference-free (sometimes call *non-instrusive* or *single-ended*) models only use the degraded signal from the speech transmission network. They form a speech quality estimate from the degraded signal and thus do not need a reference. ITU-T standardizes a narrowband reference-free speech quality model in ITU-T Recommendation P.563 (2004). The state-of-the-art model NISQA is able to predict the quality dimensions in a listening situation without a reference (Mittag et al., 2021; Mittag, 2022).

These signal-based models focus mainly on the prediction of speech quality in a listening situation and the four listening-quality dimensions. However, there have been hybrid models that predict the conversational quality based on a signal that was degraded by a speech transmission system as well as parameters of the transmission.

2.6 Hybrid Quality Prediction Models

Hybrid quality prediction models utilize the methods of both parametric and signal-based prediction. This is especially useful for predicting conversational quality, as the quality in listening situations can be predicted well with the degraded speech signal, and the degradation of the interaction can be modelled with parameters of the transmission (i.e., delay) and of the type of conversation. In the following, two hybrid models are described that predict conversational quality.

2.6.1 Objective Conversational Speech Quality Model

A hybrid, full-reference, narrowband model was developed by Guéguin et al. (2006) and extended by Guéguin et al. (2008). While it does not have an official name, in this thesis, it will be referred to as the Objective Conversational Speech Quality (OCSQ) model, as it was described in Guéguin et al. (2006). Similar to the multidimensional analysis of conversational telephony by Köster (2018), the conversational quality is split into three *contexts*: the listening quality, the speaking quality, and the interaction quality. However, the goal of this model is not to predict the perceptional dimensions but rather to predict the overall conversational quality.

For predicting the *listening quality*, OCSQ makes use of the predictions of PESQ, which is a narrowband, full-reference, signal-based, listening quality model standardized in ITU-T Recommendation P.862 (2001).

The *speaking quality* is predicted with the use of the Perceptual Echo and Sidetone Quality Measure (PESQM) (Appel and Beerends, 2002). This full-reference model uses a clean reference speech signal and a special degraded signal. The degraded signal combines the reference signal with a degraded signal that was transmitted through the system under study. This way, the model is able to detect and account for the echo that would otherwise be missed due to time alignment.

The *interaction quality* is predicted using the one-way transmission delay as an additional parameter. Because the effects of echo and sidetone are already captured in the speaking quality, this part of the model focuses on the effects of pure delay. While the final model includes a delay threshold parameter that can be changed depending on the amount of interactivity, this threshold is recommended to be a constant value.

The three different quality ratings of the listening, speaking, and interaction contexts are assumed to independently contribute to the conversational quality and are combined using a regression equation:

$$\widehat{MOS}_{conv} = \alpha \cdot MOS_{talk} + \beta \cdot MOS_{list} + \gamma \cdot \max\left(0, \text{delay} - \text{delay}_{thr}\right) + \delta \qquad (2.25)$$

where MOS_{talk} is the objective talking quality (MOS-TQO) as predicted by PESQM, MOS_{list} is the objective listening quality (MOS-LQO) as predicted by PESQ, delay is the one-way absolute transmission delay in milliseconds and delay_{thr} is a delay threshold (similar to the minimal perceivable delay mT of the narrowband E-model, see Section 2.4.1) and is set to a constant value of $400\,ms$. The coefficients are set in Guéguin et al. (2008) to $\alpha = 0.4059$, $\beta = 0.5519$, $\gamma = -1.7376$, and $\delta = 0.171$.

While the OCSQ model performs very well on the training data, the used dataset did not contain combinations of impairments. Also, the participants on the conversation tests only used the SCT (see Section 2.3.1), which resulted in low interactivity conversations. This explains the high delay threshold delay_{thr} of $400\,ms$ suggested in Guéguin et al. (2008).

2.6.2 Instrumental Diagnostic Conversational Quality

In Köster (2018), a hybrid, super-wideband, diagnostic conversational quality model is described that predicts the conversational quality, the three conversational phases, as well as the perceptual dimensions of each phase (see Section 2.3.2) to allow for diagnostic insights. Although this model has no official name, it will be referred to as the Instrumental Diagnostic Conversational Quality (IDCQ) model in this thesis. Like the OCSQ model, the IDCQ model is structured hierarchically. First, for each conversational phase, the quality of every perceptual dimension of that phase is predicted. Then, based on the estimations of the dimensions, a quality for the conversation phase is predicted. Finally, the conversational quality is predicted from the estimated quality of each conversation phase.

Because this model has been created from a limited dataset, it should be viewed as the first approach toward such a diagnostic conversational quality model.

Listening Quality

The listening quality is predicted based on the four perceptual dimensions *Noisiness*, *Discontinuity*, *Coloration*, and *Loudness*. These are predicted using the DIAL model in its super-wideband mode. The predicted perceptual dimensions are then combined into a MOS for the listening phase (MOS-LQO):

$$\widehat{MOS}_{LI} = -1.955 + 0.436 \cdot \widehat{MOS}_{Noi} + 0.516 \cdot \widehat{MOS}_{Dis} + 0.117 \cdot \widehat{MOS}_{Col} + 0.305 \cdot \widehat{MOS}_{Lou} \quad (2.26)$$

where $\widehat{MOS}_{Noi}$, $\widehat{MOS}_{Dis}$, $\widehat{MOS}_{Col}$, and $\widehat{MOS}_{Lou}$ are the DIAL predictions of the dimension *Noisiness*, *Discontinuity*, *Coloration*, and *Loudness* respectively.

Speaking Quality

The speaking quality consists of the two perceptual dimensions *Impact of one's own voice* and *Degradation of one's own voice*. Both of these dimensions are predicted by two parameters of the speech signal: the attenuation (ATT) of the speech signal in regards to the reference signal and the back coupling delay (T_B), which describes the shift between the reference signal and the degraded signal.

The quality dimension *Impact of ones own voice* (MOS_{Ios}) is estimated as:

$$\widehat{MOS}_{Ios} = 3.842 - 0.394 \cdot ATT - 0.01 \cdot T_B \quad (2.27)$$

The quality dimension *Degradation of one's own voice* (MOS_{Dos}) is estimated as:

$$\widehat{MOS}_{Dos} = 3.742 - 0.282 \cdot ATT - 0.009 \cdot T_B \quad (2.28)$$

The overall speaking quality MOS_{SP} is then estimated using the predictions form Equations 2.27 and 2.28:

$$\widehat{MOS}_{SP} = 0.144 + 0.026 \cdot \widehat{MOS}_{Ios} + 0.819 \cdot \widehat{MOS}_{Dos} \quad (2.29)$$

Because of the limited amount of parameters ATT and T_B, which do not take into account any degradation that might occur on the back-coupled speech signal, the accuracy of the speaking quality prediction is rather low.

Interaction Quality

The interaction quality consists only of one quality dimension: the *interactivity*. This dimension is predicted with on the overall one-way transmission delay T_O:

$$\widehat{MOS}_{Int} = 3.554 - 0.001 * T_O \quad (2.30)$$

where $\widehat{MOS}_{Int}$ is the estimation of the perceptual dimension *interactivity*. Again because of the limited dataset, the transmission delay T_0 is only contributing very slightly to the MOS estimate. With the predicted interactivity, the overall interaction quality (MOS_{IN}) is estimated:

$$\widehat{MOS}_{IN} = -0.299 = 0.942 * \widehat{MOS}_{Int} \quad (2.31)$$

Similar to the prediction for the speaking phase of the conversation, this model achieves a rather low accuracy.

Conversational Quality

From the estimation of the listening phase (Equation 2.26), the speaking phase (Equation 2.29), and the interaction phase (Equation 2.31), the overall conversational quality is predicted:

$$\widehat{MOS}_{CO} = -0.393 + 0.188 \cdot \widehat{MOS}_{LI} + 0.354 \cdot \widehat{MOS}_{SP} + 0.477 \cdot \widehat{MOS}_{IN} \tag{2.32}$$

While the overall conversational quality prediction results in a good prediction when using the quality ratings of the perceptual dimension directly, it has a low correlation with the predicted dimensions from the data points available in Köster (2018). While the listening phase estimation provides an acceptable accuracy, the speaking and interaction phase prediction have low accuracy, resulting in a consistency of $\rho = 0.44$. A validation with more data points and a prediction model with more parameters for the speaking and interaction phases would be needed to improve a prediction based on this approach.

2.7 Packet Loss and Understandability

When considering the effects of bursty packet loss, the listening quality alone is not sufficient to describe the impact on the conversation. The effects of packet loss can be defined by the percentage of packets lost over a given time frame, the length of speech contained in a single packet, the burstiness of the loss, and the codec that is used (ITU-T Recommendation G.107, 2015). As described in Equation 2.16, the burstiness can be described by a two-state Markov model. Depending on the severity of burstiness of the packet loss, the speech signal might be affected locally but very heavily. This may result in the speech completely dropping out, even with a PLC algorithm engaged (Raake, 2006).

The Speech Intelligibility Index (SII) and its predecessor, the Articulation Index (AI) are standardized measures that have a high correlation with the intelligibility of speech in a listening situation (American National Standards Institute, 1997). The SII itself is not a measure of how likely a spoken sentence is understood, but instead of how many audio cues are usable in a given setting (Hornsby, 2004). The SII uses frequency-specific information on the speech levels, the "noise" levels, and their auditory thresholds, which are weighted by the importance of each frequency band in regards to speech understanding. The resulting index can be transformed into speech understanding scores with the help of transfer functions. These functions are specific to the material that is listened to, as unknown random syllables and previously known full sentences have different chances of being understood.

Intelligibility models are also used in speech synthesis, where there is a need to assess the intelligibility of the synthetic voice. The Semantically Unpredictable Sentences (SUS) test consists of automatically generated grammatically correct but semantically unpredictable sentences (Benoît et al., 1996). Because the participants trying to understand these sentences are not able to gain information from context, only pure au-

dible information is available to be able to judge the intelligibility. However, in realistic conversation scenarios, there is almost always context given and conversation partners will use this implicit information for the continuation of the conversation.

2.8 Turn-Taking

Turn-taking is the set of practices speakers use to organize the conversation and allocate speaking turns. The analysis of turn-taking in the conversation has its roots in the "simplest systematics" defined by Sacks et al. (1974). There, general rules for turn-taking have been laid out that describe the process of selecting the next speaker (either through the current speaker or by self-selection), as well as the continuation of a turn by the current speaker. Turn-taking incorporates many auditory, visual, and contextual cues (Ford and Thompson, 1996). Especially in telephone conversations, where no visual cues are present, people rely on the immediacy of signals in prosody and content to perform smooth and uninterrupted turn-taking.

Recent work in the area of turn-taking focuses on the analysis of turn-taking behavior in conversations (Lunsford et al., 2016; Niebuhr et al., 2013), end-of-turn prediction (Liu et al., 2017; Skantze, 2017), and rule-based turn-taking models for the intended use in SDS (Selfridge and Heeman, 2012; Baumann, 2008).

In Lunsford et al. (2016), the duration of turns and their timings are analyzed. Turns are analyzed based on the offset from the end of the previous utterance. Thus, turn-continuations (the current speaker keeps the turn) and turn-transitions (the speaker changes) are analyzed as equal alternatives to what could have occurred. For this analysis, gaps and overlaps between turn-transitions are recorded, forming a probability distribution on the offset in relation to the end of the previous turn. Overlaps in the speech during the turn-transition are counted as a negative offset, while gaps between the turns are counted as a positive offset. Turn-continuations are counted separately with positive offsets (as it is not possible to interrupt oneself). Measuring the turn-transitions and turn-continuations based on the offset results in two alternative models on how turns are allocated between the speakers. However, the timing of turn-keeping and turn-yielding is also dependent on the dialogue context (Heeman and Lunsford, 2017).

During the analysis of turn-taking, generally, short utterances produced by short backchannels (e.g., "yes" or "okay") need to be taken into account so as not to distort the overall length of turns. In Heldner et al. (2011), a minimum utterance length of 200 ms is proposed to remove any very short utterances. Also, in Lunsford et al. (2016), preprocessing steps are applied to filter out backchannels and other short interruptions that do not make up a valid turn.

For the field of Spoken Dialogue Systems, modeling turn-taking increases the interactivity and thus the naturalness of spoken dialogue systems. Anticipating the end of the user's turn can make an interaction more fluid and human-like (Liu et al., 2017). One approach for the implementation of realistic turn-taking in dialogue systems is the use of an end-of-turn prediction model that tries to anticipate the end of the user's turn in real-time before they have finished their utterance. This approach can lead to more realistic gaps and overlaps, or at least to a reduction in the silence between requests

from the user and the answer from the system. Depending on the time-sensitivity, end-of-turn prediction models include just the prosodic information (Ferrer et al., 2002), or also lexical information (Liu et al., 2017). More recent approaches include the use of recurrent neural network (Skantze, 2017), as well as transformer-based models (Ekstedt and Skantze, 2020). Besides modeling and predicting the end of a turn, also models for the prediction of backchannels have been build byKawahara et al. (2016).

2.9 Simulation of Dialogue

The simulation of dialogue has been used in the field of dialogue systems to model user behavior (Eckert et al., 1997), to evaluate the usability of such systems (Hillmann, 2017; Pietquin and Hastie, 2013; Engelbrecht et al., 2009), or to train the dialogue manager (Schatzmann et al., 2006).

For the user simulation used in the automatic usability evaluation of spoken dialogue systems, the simulated user is designed in a way to mimic the interaction of real users with the systems (Möller, 2004). Approaches to simulating a user are closely related to those of modeling a dialogue system itself. User simulations are trained using agenda-based dialogue management (Schatzmann et al., 2007), Hidden-Markov-models (Cuayáhuitl et al., 2005; Schatzmann and Young, 2009), as well as machine learning methods (Janarthanam and Lemon, 2009; Schatzmann et al., 2006). To realistically model the strategies and behaviors of users, these simulations have been adapted to model changes in interaction style and overall conversation goals (Hillmann and Engelbrecht, 2015). Also, misunderstandings by the user and errors in the speech recognition and language understanding of the system are modelled (Engelbrecht et al., 2009). These simulations can be evaluated in order to monitor and improve performance. This evaluation can be done in the form of performance metrics of the conversation (e.g., number of turns, task success), or a simulated dialogue may be compared to human dialogue. One way to measure the similarity of two dialogues is to compare the probability distributions of dialogue acts and measure their distance with the KL (Kullback-Leibler) divergence and dissimilarity (Pietquin and Hastie, 2013). Kullback and Leibler (1951) defines the KL divergence between two distributions P and Q as:

$$D_{KL}(P||Q) = \sum_{i \in X} p_i \log(\frac{p_i}{q_i}) \tag{2.33}$$

With respect to two distributions of dialogue acts that are compared, p_i and q_i are the frequency of dialogue acts in the histogram of the distributions P and Q, respectively, while X is the probability space that is shared between the two sets of dialogue acts. Because this divergence is not symmetric, it may result in different values for $D_{KL}P||Q$ and $D_{KL}Q||P$. Thus Pietquin and Hastie (2013) introduce a dissimilarity formula that can be used as a distance measure:

$$DS(P||Q) = \frac{D_{KL}(P||Q) + D_{KL}(Q||P)}{2} \tag{2.34}$$

The generated dialogue between the dialogue system and the user simulation can be analyzed to improve the dialogue manager during training, or to predict its quality in usability tests. Based on measurable parameters like task success or number of turns, models are built to predict the quality of the system under study (Möller and Skowronek, 2004; Hillmann, 2017).

Dialogue simulations are often executed on a dialogue-act level, where a dialogue system and a simulated user exchange semantic representations of what is being said. Notable exceptions include the simulation of spoken dialogue in Baumann (2008), where the system under study and the user simulation exchange speech signals to realistically simulate turn-taking. However, because of the focus on turn-taking, there is no content being exchanged between the simulated interlocutors. Another dialogue simulation that focuses on turn-taking is described in Padilha (2006), where turn-taking in group conversations is modeled. Here the simulation is performed on a textual level, and simulated speakers try to organize and handle turn-taking. In the domain of user simulation for the prediction of the quality of a dialogue system, Scheffler et al. (2009) describes an approach to simulate the dialogue on the speech level automatically, thus interacting with the dialogue system as a black box.

A simulation of conversations in the context of speech quality is standardized as Artificial Conversational Speech in ITU-T Recommendation P.59 (1993), where the on-off patterns of conversational speech are simulated based on the average sojourn times of recorded conversations. However, this simulation also does not concern with exchanging information, as the simulated interlocutors exchange artificial speech-like sound (ITU-T Recommendation P.50, 1999).

2.10 Incremental Dialogue Systems

Classical architectures of spoken dialogue systems follow a pipeline approach: each logical module of the dialogue system processes incoming data, extracts useful information, and forwards the produced data in a new representation to the next module (Jurafsky and Martin, 2009). Generally, speech from the user is processed by an Automatic Speech Recognitino (ASR) module that relies on Hidden-Markov-models, Dynamic Time Warping, or Neural Networks to transcribe the speech of the user into machine-readable text. This transcription is then processed by a Natural Language Understanding (NLU) module that extracts the intent of the user and named entities into a semantic representation called Dialogue Act (DA). These are given to the Dialogue Manager (DM) that decides (often with the use of the dialogue history and external data sources) which response to generate. The abstract representation of the response is turned into text by an Natural Language Generation (NLG) module. However, this module is sometimes skipped, and the DM directly produces the response in text form. Finally, a Text-To-Speech (TTS) module synthesizes the text into speech (Jokinen and McTear, 2009).

While this type of architecture is very robust and used widely, it has some limitations. One drawback is that dialogue systems designed in a pipeline architecture process the data sequentially, based on a complete utterance of a user. As a result, the different

modules of the system are idle during the utterance and start processing only when the previous module produces output. This may lead to slower reactions from the systems and worse predictions of the individual modules.

The concept of *incremental* dialogue systems presented in Schlangen and Skantze (2009, 2011) is a general model for incrementalizing the processing in dialogue systems. While dialogues are generally defined incrementally (i.e., each dialogue is made up of smaller utterances/turns), the incremental dialogue systems model describes the incrementality on the level of the utterance. With every minimal processable amount of an utterance, each component will be triggered into activity. The creation of this processing paradigm is motivated by an increase in the reactivity of the system, a better quality of prediction in each processing step (e.g., through forming of early hypotheses), a more natural interaction (e.g., through backchannels like "uh-huh" or "yeah"), and by adding realism to interactions with dialogue systems.

Generally, the incremental processing model has two main interfaces that are present in each incremental dialogue system: An incremental module processes incoming increments, forms hypotheses, and forwards them to other incremental modules. An Incremental Unit (IU) is the basic unit of information transmitted between the incremental modules. These units contain the data that makes up the increment, as well as auxiliary information about the underlying data and hypotheses based on it.

Incremental modules contain a *Left Buffer* that contains IUs that should be processed by the incremental module. When the incremental module produces hypotheses based on new input, it places them in the form of IUs into the *Right Buffer* where they are forwarded to connected incremental modules. The incremental modules keep an internal record of the data that is currently being processed, as new incremental units are placed in the Left Buffer. Each incremental unit has to be viewed as a working hypothesis (as not all the information is available), and thus, they can be revised later on. For this purpose, IUs can be updated (e.g., an ASR module might update a word or a part of a word with new speech incoming) and also be revoked (e.g., the NLU module might revoke a concept if an update in the ASR module changed a recognized word). Generally, not every new incremental unit that is arriving at the left buffer of an incremental module will produce an output IU.

Incremental Units are the basic units that transmit the incremental update as a *payload*. Additionally, they provide information about their hypothesis, about previous IUs that were generated by the same incremental module (horizontal relationship), as well as information about IUs that its hypothesis is based on (vertical relationship). Each IU has a *successor* relationship, which creates a chain of IUs from the same line of hypotheses. The vertical connection of incremental units is the *grounded* link. With this link, every IU can reference the unit that the current hypothesis was based on. This unit is usually of the type one layer above in the abstraction chain. Thus, an IU coming from an NLU module containing a hypothesis about a concept named in the utterance of the user might be grounded in an IU from an ASR module, referencing the word (or words) containing the text of the concept, which in turn might be grounded in an IU from a microphone module, containing the speech signal of the word. Each IU has a *committed* flag, that indicates whether it is no longer used for the current forming of a hypothesis and will not be revised. This can help improve the accuracy when creating hypothesis based on these IUs.

While the incremental processing model describes the method in a general, abstract way, there are many models and approaches in the domain of spoken dialogue systems that make use of the concept. Speech recognition models are widely used in an incremental way (Selfridge and Heeman, 2012), end-of-turn prediction's main use is the employment in incremental SDSs (Skantze, 2017), and state-of-the-art NLU modules have been incrementalized as well (Rafla and Kennington, 2019). Especially recent advances of spoken dialogue in human-robot-interaction are based on the incremental processing paradigm (Kennington et al., 2020).

InproTK (Incremental Processing Toolkit) is a framework for building incremental spoken dialogue systems presented in Baumann et al. (2010) and Baumann and Schlangen (2012b). This framework is based on the programming language Java and implements the main concept laid out in Schlangen and Skantze (2011). InproTK has a range of example systems that showcase the concepts of incremental processing and the features provided by the toolkit. It was extended for the use in situated dialogue in Kennington et al. (2014), but the development of the toolkit has since been discontinued[2]. Based on this system, components like an incremental speech synthesis module have been created (Baumann and Schlangen, 2012a). Since 2019, a new, more modular version called "InproTK 2" is available online[3] and is used in human-robot-interaction systems (Fischer et al., 2021).

[2] Based on the activity of the git-repository `https://bitbucket.org/inpro/inprotk/src/master/`, last accessed March 11th, 2022.

[3] `https://github.com/timobaumann/inprotk`, last accessed March 11th, 2022.

Chapter 3
Simulation Architecture

A conversation simulation that should accurately reflect the effects of delay and packet loss has specific requirements. To be able to simulate conversation scenarios with distinct interactivity and to model misunderstandings due to bursty packet loss, the simulated interlocutors need to process information on a symbolic level. In order to model the changes in turn-taking due to transmission delay, the same simulation also needs to include realistic turn-taking with a focus on timing. For these reasons, the main requirement for the conversation simulation is an architecture that can reproduce a conversation on every layer of abstraction: from the speech signal to the textual information, as well as the dialogue act layer. Also, the architecture needs to process the information incrementally in order for realistic turn-taking to be modeled.

This chapter describes the simulation architecture used for the conversation simulation. First, a new incremental processing programming framework, "retico", is introduced that is able to model time-sensitive interactions like turn-taking and the repairing dialogue caused by misunderstandings due to packet loss. Then, the datasets are described that are used to model, train, and evaluate the simulation, as well as quality models. With the underlying data, the simulation architecture used for the conversation simulation is described in its incremental parts. The simulation is constructed to simulate SCT and RNV conversations. The simulation is initially constructed without a turn-taking model to act as a baseline. Finally, the conversations produced by the baseline simulation are evaluated semantically on the dialogue act layer, as well as on the interaction level, by analyzing the conversational interactivity parameters. Finally, the performance of the system architecture is briefly discussed.

The incremental processing framework retico and an overview of the simulation architecture have been partially published in Michael and Möller (2018), Michael and Möller (2019), and Michael (2020). The datasets used for the training and evaluation of the simulation have been described in Michael and Möller (2020a) and Uhrig et al. (2018).

3.1 Retico Incremental Processing Framework

During the time of the creation of the conversation simulation architecture, there was no actively maintained and up-to-date programming framework or toolkit available that could be used as an incremental platform on which to base the simulation. With the

requirements for the simulation and the abstract model of incremental processing, a new framework was developed by me to be used for the conversation simulation. In this section, the basic implementation mechanics of the framework are outlined, that are all based on the concepts of incremental processing described in Chapter 2.

The retico (an initialism for *real time conversation*) incremental processing programming framework is a Python library that implements the basic concepts of incremental processing as described in Schlangen and Skantze (2011). It is available as an open-source project on github[1]. A more modular version is currently in development to serve as a general-purpose platform for incremental processing[2]. The core of the framework defines interfaces and functionality for incremental modules and incremental units, and it also contains basic sound input and output processing capabilities. It also includes incremental modules from various fields of a spoken dialogue system. A user interface provides a way to create new modules and connect them together into an incremental network.

The basic definition of Incremental Units (IUs) provides access to the `grounded_in` and `previous_IU` references. Every incremental unit tracks its creation time as well as its age. With this timing information IUs from different sources can be synchronized. Also, IUs track the state of their `committed` and `revoked` status. Incremental units are defined with the inheritance of object-oriented programming in mind so that functionality can be abstracted and shared over different environments. For example, a synthesis module might require a `TextIU`, where the specific implementation might be a `SpeechRecognitionIU` generated by an ASR module or a `GeneratedTextIU` generated by an NLG module, each with their own set of parameters and functionalities.

The abstract definition of the incremental module provides access to the main processing loop in which each module receives IUs from the left buffer and might return one or more IUs to the right buffer. The connection between incremental modules and the transmission of IUs is handled by the framework. For this, a type checking is implemented so that only modules with compatible input and output IU types can be connected. Incremental modules have a unified interface to start and stop the processing, as well as a separate routine for setting up auxiliary functionality. That way, an incremental module can set up required services before the processing inside the module is started. Based on this, additional basic module types are defined: the `Consuming Module` that does not produce any output IUs (e.g., for access to a speaker or other output peripherals) and the `Producing Module` that does not take any IUs as an input (e.g., for access to microphones and sensors that are not based on the incremental system).

Retico contains basic audio processing functionality. A `MicrophoneModule` (producing module) is available to capture input from an internal microphone, external microphone, or a line input, and the frame size of the resulting IUs can be set. A `SpeakerModule` (consuming) module takes `AudioIU`s and outputs them on a selected speaker, while a `StreamingSpeakerModule` contains a jitter buffer to output the audio in a smooth and continuous manner. Besides the audio processing modules, retico also includes online and offline speech recognition modules (CMUSPhinx, Google ASR), natural language understanding (rasa NLU), dialogue management (agenda-based, rasa RNN-based, and n-gram-based modules), speech synthesis (Mary TTS, Google TTS),

[1] `https://github.com/thilomichael/retico`

[2] Repositories are available at `https://github.com/retico-team`

as well as a translation module (Google Translate). Several demonstration applications like a spoken translation service or a restaurant information system built with retico are available.

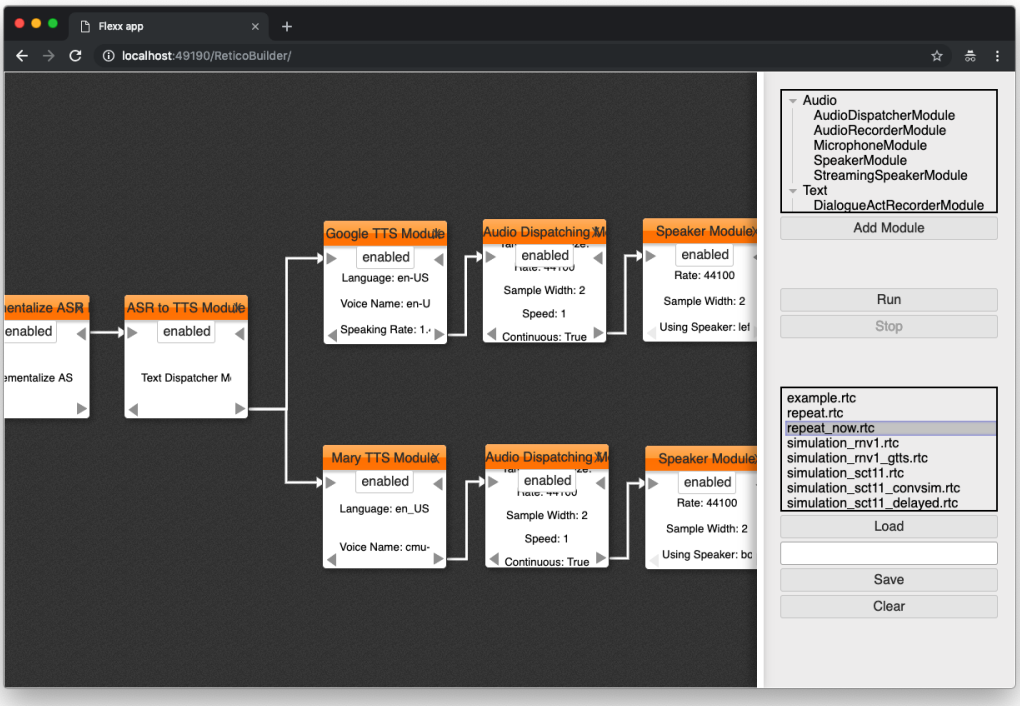

Fig. 3.1 Screenshot of the graphical user interface of retico. Incremental modules are showing details about their configuration. Connections between modules are visualized as arrows. A menu allows for the instantiation of new modules, the starting and stopping of network execution, as well as saving and loading of saved networks.

A graphical user interface (called "retico builder", see Figure 3.1) gives quick access to instantiate modules and connect them to form incremental networks. During instantiation of modules, parameters can be provided (e.g., the language of an ASR module can be selected in the user interface), and modules are represented visually on a canvas. With visual connection terminals, two modules with compatible input and output IU-type can be connected, which is visually represented as an arrow. Once multiple modules are connected, the network can be executed. During execution, each module displays information about the current state or about the most recent incremental units that are being processed. Networks can be saved to a file and loaded in later sessions. Also, a network saved to a file can be loaded in Python code directly. That way, complex networks (like a conversation simulation) can be created and connected visually, saved to a file, and later executed on a dedicated simulation server without the graphical user interface.

While retico defines general incremental processing interfaces, provides a programming interface for new modules, and contains basic audio processing modules, some functionality is deliberately not provided. The approach to incremental processing pro-

vided by retico assumes that each incremental module is capable of processing IUs faster or as fast as they arrive in its left buffer. Also, the time alignment of IUs from different sources is not handled by the framework. If an incremental module needs to align information from different sources, this may be achieved with the help of the time information attached to each IU. While incremental modules in retico may have multiple types of input IU, each module may only have one type of output IU. While this restriction helps with a consistent definition of where IUs are routed in the network, it limits the types of networks that can be created. This can be addressed by either implementing incremental modules that only concern one specific type of information or by creating IUs that contain a mixture of data. For example, a dialogue manager may produce an IU that contains a dialogue act from which text should be generated, but it also contains information about how to synthesize it.

3.2 Simulation Datasets

In order to model and later on evaluate a simulation that reproduces characteristics of a real conversation, appropriate datasets are needed. For this, two existing conversation datasets are used, and one dataset is specifically created for the simulation objective (see Table 3.1). All conversations were recorded as part of conversation experiments modeled according to ITU-T Recommendation P.805 (2007). In all three experiments, participants rated the overall conversational quality of each conversation on a 7-point extended continuous scale. The ratings were later transformed into ACR values with the Equation 2.1 provided in Köster et al. (2015).

Table 3.1 Overview over the three datasets used in this thesis.

Dataset Name	SMISS	CONVSIM	UWS
Scenarios	SCT #11, RNV #1	SCT, RNV	SCT
Delay (ms)	-	0, 800, 1600	0, 800, 1600
Packet-Loss	-	0 %, 15 %, 30 %	-
BurstRatio	-	4	-
Participants	40	58	20
Conversations	60	580	130
Coding	16-bit linear PCM @ 32 kHz	16-bit linear PCM @ 44.1 kHz	16-bit linear PCM @ 44.1 kHz

3.2.1 SMISS Dataset

The SMISS dataset was created as part of the research on the multidimensional analysis of conversational speech at the Quality and Usability Lab, Technische Unviersität Berlin. The experiment followed the subjective diagnostic test method for conversational speech quality analysis (ITU-T Recommendation P.804, 2017) and the participants performed SCT as well as RNV conversations. The test was carried out with a SWB conversational

system, which allowed the simulation of various impairments. Subjects were placed in separate, soundproofed booths and communicated over stereo headsets to minimize the possibility of acoustic echo. 40 naïve, German-speaking participants took part in the experiment, and each pair performed 11 conversations. The recorded conversations are stored with 16-bit linear Pulse Code Modulation (PCM) at 32 kHz.

The SMISS dataset was used to model the dialogue in the simulation. For this, one concrete SCT scenario and one RNV task were selected to be simulated. Because the baseline data for those two conversations should not include any degradations, the SCT scenario 11 (ordering a pizza, see Appendix A) and the RNV task no. 1 (see Appendix B) was selected from the available conversations. 30 of these SCT scenario 11 conversations and 30 of the RNV scenario 1 conversations were extracted from the SMISS dataset, resulting in 60 clean conversations without any degradations.

The conversations were transcribed and annotated with dialogue acts and concepts, as well as with end-of-turn information. The transcriptions were automatically generated with Google ASR, and errors were manually corrected. The conversations were then annotated with dialogue acts and concepts. The dialogue acts were selected to be general and thus usable for every standardized conversation type referenced in ITU-T Recommendation P.805 (2007), while the concepts reflect the specifics of the scenario.

Table 3.2 Annotation schema of the SCT scenario 11 and RNV task 1 conversations.

Dialogue Act Types	**Description**	**Example**
`greeting`	Greeting	"*Hello, this is pizzeria Roma.*"
`goodbye`	Farewell	"*Goodbye.*"
`provide_info`	Providing information	"*I want a vegetarian pizza.*"
`provide_partial`	Providing parts of an information	"*My phone number is 0 3 0 ...*"
`request_info`	Requesting information	"*What's your address?*"
`offer_info`	Offering information	"*Should I give you my adress?*"
`stalling`	Stalling the conversation	"*Uhm...*"
`request_confirm`	Request a confirmation of information	"*Main street 46?*"
`confirm`	Confirming that information	"*Yes, 46.*"
`misunderstanding`	Something was not understood	"*I did not understand that.*"
`thanks`	Giving thanks	"*Thank you.*"
`welcome`	Receiving thanks	"*You're welcome.*"

Table 3.2 shows an overview of all dialogue act types that were defined. Generally, all dialogue acts may occur with concepts. For example, when a person says "*Hello, this is pizzeria Roma*" this is annotated as the dialogue act `greeting` with the addition of the concept `callee_name`. The `provide_partial` dialogue act is used for the many occurrences where information is split over multiple turns. Examples for this might be an address that gets split by street name, postal code, and city, or it can be a block of numbers in the RNV task, where each number gets transmitted with a `provide_partial`. A `confirm` dialogue act might be a very general "*yes.*" or "*correct.*" when occurring without a concept. It also can be used in very specific confirmations when it is used with a concept (e.g., "*Yes, a vegetarian pizza.*" would be annotated as `confirm:pizza_type`). Also, the concept named in the conversation were annotated.

In contrast to the dialogue acts, which can be used for many different conversation tests and scenarios, the concepts are specific for each scenario. The concepts used for scenario 11 of the SCT and scenario 1 of the RNV task are shown in Appendix C.

In addition to the dialogue act annotation and the transcription, the beginning and end of each turn (defined by an annotated dialogue act) have been annotated with beginning-of-turn and end-of-turn markers. These turn annotations were used to extract the gaps, overlaps, and pauses (as described in Section 2.8).

3.2.2 CONVSIM Dataset

The CONVSIM dataset was created specifically for use in the conversation simulation approach and thus includes delay as well as packet loss impairments (Michael and Möller, 2020a). The experimental procedure followed ITU-T Recommendation P.805 (2007) and participants rated the conversational quality, as well as the listening dimensions and the interactivity dimension on the 7-point ECS scale. 58 German-speaking participants (age 18 – 71, 28 of them female) without hearing impairments were located in separate soundproof booths and communicated through monaural headsets. They were connected with a fullband telephone simulation (see Table 3.1) that transmits speech with 16-bit linear PCM at 44.1 kHz. Each pair of participants performed ten SCT and ten RNV conversations, of which the first two had no impairments and were used to familiarize the subjects with the test protocol. The following 18 SCT and RNV conversations were degraded with one of $0\,ms$, $800\,ms$, and $1600\,ms$ one-way end-to-end transmission delay levels, combined with 0 %, 15 % and 30 % zero-insertion packet loss with a burst ratio of 4.0. The high delay values were chosen to incite extreme cases of interruptions and double talk, as strong impacts on the conversations give a good baseline for the simulation. The high packet loss probability values and the burst ratio of 4.0 were chosen so that large chunks of speech were cut out of the utterances, which incites misunderstandings between the interlocutors. This results in 580 conversations, each with two ratings of the conversational quality and the recorded quality dimensions.

As each conversation partner was recorded on a separate channel, the turns of each speaker were segmented based on the automatic VAD of each audio channel. Based on this turn segmentation, a P-CA (see Section 2.3.4) was performed, and the gaps, overlaps, and pauses between the turns were extracted (see Section 2.8). For the conversations with transmission delay, also the delay-dependent conversational parameters (IIR, UIR, SARc) were calculated. Based on the annotation from the automated turn segmentation, the individual turns of the conversations with 0 %, 15 %, and 30 % packet-loss were transcribed, to later aid in the analysis and modeling of disruptions in the conversation due to packet loss.

An overview over the overall conversational quality (MOS-CQS) of the CONVSIM dataset can be seen in Figure 3.2 for SCT conversations and in Figure 3.3 for RNV conversations, with ACR MOS calculated according to Equation 2.1. As already shown in similar experiments (Raake et al., 2013; Egger et al., 2010), the perceived quality of RNV conversations is much more strongly impacted by the transmission delay than the quality of SCT conversations. This is due to the higher conversational interactivity of RNV conversations. Regarding the conversational quality of packet-loss-affected

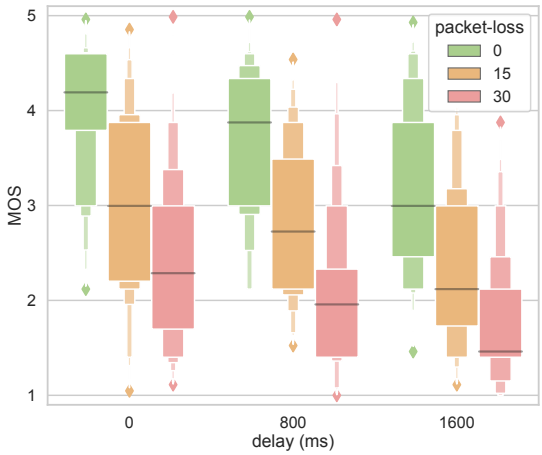

Fig. 3.2 Mean opinion score for the SCT conversations of the CONVSIM dataset at 0 ms, 800 ms, and 1600 ms delay as well as 0 %, 15 %, and 30 % packet loss.

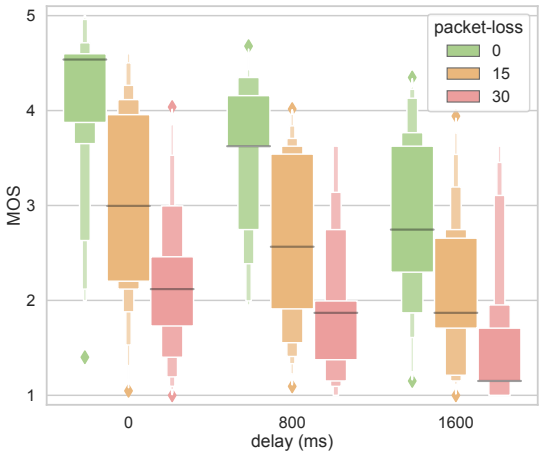

Fig. 3.3 Mean opinion score for the RNV conversations of the CONVSIM dataset at 0 ms, 800 ms, and 1600 ms delay as well as 0 %, 15 %, and 30 % packet loss.

conversations, the SCT and RNV conversations seem to be impacted the same. However, in the combination of delay and packet loss, the quality of the RNV conversations are again more strongly affected.

3.2.3 UWS Dataset

The UWS dataset was recorded at the University of Western Sydney, and the results are published in Uhrig et al. (2018). In contrast to the SMISS and CONVSIM experiments, this experiment was done as part of an electroencephalogram (EEG) study. Participants were located in separate soundproof cabins and communicated over a fullband communication network that transmits the speech with 16-bit linear PCM at 44.1 kHz (see Table 3.1). During the conversation, the EEG activity of the participants was recorded, which caused some constraints on the subjective assessment as described by ITU-T Recommendation P.805 (2007). Participants had only limited movement capabilities during the study so as to not disturb the recording of the EEG signal. The overall conversational quality was assessed on the 7-point ECS.

Ten pairs of English-speaking participants conducted 13 SCT conversations each, and for every conversation, either 0 ms, 800 ms, or 1600 ms of one-way end-to-end transmission delay was inserted. While this dataset is limited in the number of impairment conditions, as well as in the number of conversation scenarios, it is the only dataset containing conversations in English. For this, the SCT scenarios given in ITU-T Recommendation P.805 (2007) were adapted for Australian participants, as can be seen in Appendix A.

3.3 Incremental Simulation Network

The network of incremental modules comprising the simulation architecture is created with the retico framework. The general simulation approach consists of two spoken dialogue systems, agent A and agent B, that communicate on the speech signal level. Each agent represents one person in the conversation. Agent A takes the role of the caller in each conversation scenario, and Agent B fulfills the role of the callee. The speech signal gets routed in packets (i.e., incremental units) through a simulated telephone network module that is able to delay the arrival of the packets, as well as to replace packets with silence to model zero-insertion packet loss. Logging modules at the non-degraded and degraded end of each agent saved the resulting speech signal to a file for later analysis. Additionally, the dialogue acts, and the transcription of the conversation is logged and saved to a file. An overview of the general layout of the simulation network can be seen in Figure 3.4.

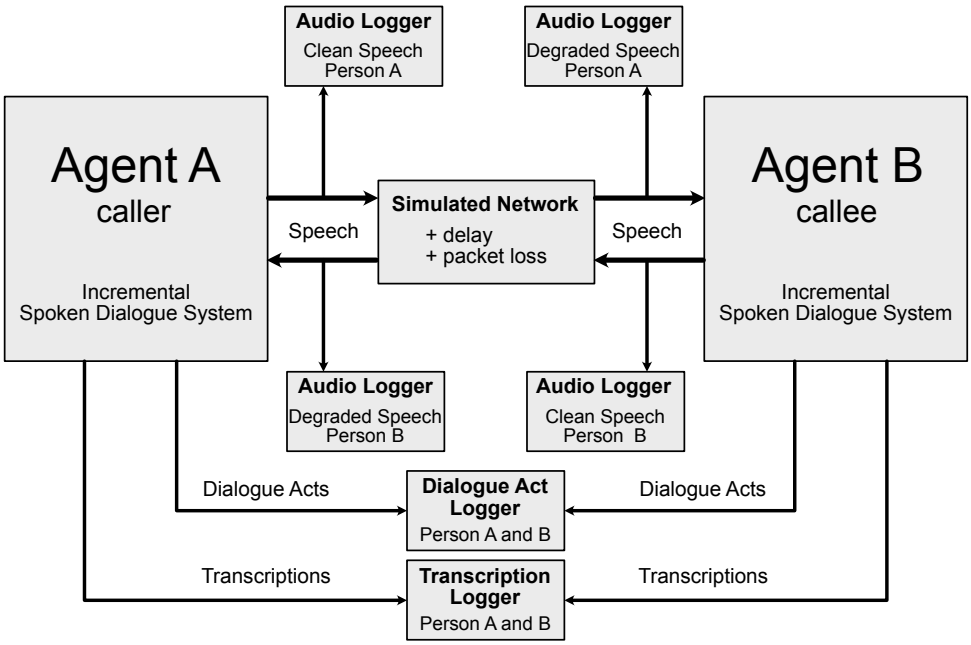

Fig. 3.4 Incremental network layout of the simulation. The two simulated conversation partners are abstracted as spoken dialogue systems, a simulated telephone network introduces impairments, and logging modules save the resulting data to disk.

Because of the incremental nature of this simulation, the agents act unsynchronized and independent of each other. A delay of the speech signal of one agent by the telephone network does not directly affect the mechanics of the other agent. Thus, the agents are only able to adapt their behavior based on the arrival of the delayed signals that are presented by the telephone network.

Current state-of-the-art models in many areas of spoken dialogue systems, like speech recognition, end-of-turn prediction, or natural language understanding, are not on par with human abilities in conversational scenarios. Thus, the simulated telephone network allows for a *side channel* to transmit information about the current state of the interlocutor

that might not be easily recoverable from the speech signal alone. The exact usage of this side channel is described in the sections detailing the incremental modules of the network.

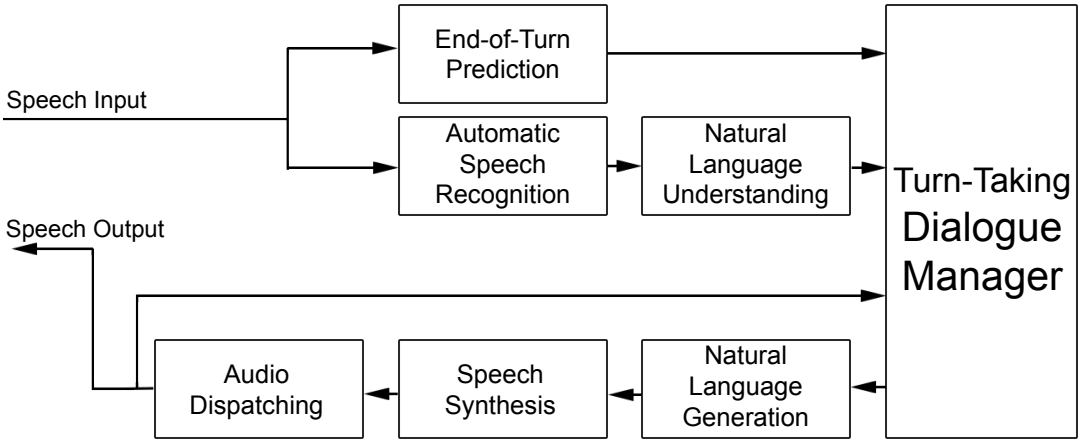

Fig. 3.5 Incremental network layout of one agent, including the End-of-Turn prediction, speech recognition, natural language understanding, dialogue management, language generation, speech synthesis, and speech dispatching modules.

The two spoken dialogue systems of agent A and agent B are constructed as incremental networks shown in Figure 3.5. The main component of the agents is a *Turn-Taking Dialogue Manager*, which orchestrates the taking of turns, dialogue management, and speech dispatching. The turn-taking information is provided by the end-of-turn prediction module that feeds directly from the speech input. The dialogue act and concepts are received by an incremental NLU module that receives the live transcripts from the incremental ASR module. Once the dialogue manager decides which dialogue acts and concepts should be returned, it provides this information to the NLG module. There, also a dispatching flag is provided so that the turn-taking dialogue manager can decide when the agent should output the speech and when it should stay silent. The generated text from the NLG module is then synthesized in the TTS module. Here, the dispatching flag is still included in the IUs that are being produced. Finally, the audio dispatching module buffers the synthesized speech. When the dialogue manager does not set the dispatching flag, the audio dispatching module produces silence, which is routed to the simulated network. Once the dispatching flag of the dialogue manager is received, it outputs the synthesized speech. The output of the audio dispatching module is also routed to the turn-taking dialogue manager itself. There it is used to monitor the progress of its audio output itself.

The following sections provide a more detailed description of the incremental architecture of the agents and the simulated network environment.

3.3.1 Speech Recognition, Natural Language Understanding

The speech recognition and natural language understanding modules of the simulation provide the transcript of the incoming speech and extract the intent and named entities in that transcription. As described in Section 3.1, the retico framework includes state-of-the-art ASR and NLU modules. However, current speech recognition models do not yield the same accuracy as a human would in the same conversation. For example, Siegert et al. (2020) reports Word Error Rate (WER) as high as 14 % for clean, device-directed, German speech with current state-of-the-art speech recognition systems. Thus, the speech recognition and natural language understanding modules used in the simulation do not actually employ a model, but rather rely on the information provided by the side channel of the simulated telephone network.

The speech signal IUs coming into the left buffer of the incremental ASR module are provided by the telephone network. Each IU has metadata attached to it that contains the transcription of the current packet of speech. This text is then extracted from the metadata and used as the transcription from the ASR module. In this baseline version of the ASR module, the speech data contained in the incoming IU is discarded and not used for the generation of the transcript.

The NLU module receives the transcripts of the speech recognition module. Because the simulated interlocutor has only a limited number of possible utterances that may be outputted (see Section 3.3.3), the module looks up the dialogue act and concept associated with the corresponding speech from the annotated SMISS dataset.

3.3.2 End-of-Turn Detection

Similar to the ASR and NLU modules, the end-of-turn detection module makes use of meta-information provided in the speech IUs provided by the network. For each packet of incoming speech, the module uses a voice activity detection based on a Gaussian mixture model (W3C Recommendation WebRTC, 2021) to classify if the interlocutor is speaking or not. If so, the meta-information of the IUs contain the information on how long the turn of the interlocutor will last in seconds. This information is updated for every packet that arrives from the incremental network. The information on whether the interlocutor is speaking and, if so, for how long they will continue to do so is forwarded to the turn-taking dialogue manager.

3.3.3 Language Generation and Speech Synthesis

The NLG module receives a dialogue act and optionally one or multiple concepts from the turn-taking dialogue manager to be turned into natural language text. Additionally, a flag is provided on whether the text that will be synthesized later in the pipeline should be dispatched (i.e., outputted to the interlocutor) or not. This flag is not used in the NLG module, but it is attached to each text IU that is generated by the module.

The NLG module contains a database of the dialogue acts and transcripts from the SMISS dataset. In order to locate an utterance that fits the dialogue act and concept provided by the dialogue manager, the module first locates all transcriptions that were annotated with the dialogue act provided. Then, if there are one or more transcriptions that fit all the provided concepts, a transcription is randomly chosen from the set, and together with the dispatching flag, it is sent to the TTS module. If there is no fitting dialogue act, a dialogue act without concepts is chosen. For example, if the dialogue act `confirm` is not available with the concept `pizza_type` (e.g., "*Yes, a vegetarian pizza*"), a transcription of the dialogue act `confirm` without any concepts is used ("*Yes.*" or "*Correct.*").

The speech synthesis (or TTS) module receives the transcripts from the NLG module and produces the corresponding speech. Each synthesized utterance is cached so that future requests for the same sentence will be faster. This is often useful when the dialogue manager "prepares" an utterance without the dispatching flag being set. Then, the natural language generation and speech synthesis is executed without the speech needing to be sent to the interlocutor. Once the speech should be dispatched, the dialogue manager sets the dispatching flag in its output IU and the speech synthesis is able to use the cached version of the synthesized speech. The production of the utterance itself can be done either by synthesizing the speech with a state-of-the-art synthesizer, like Google TTS or Mary TTS, or by using the utterances of the training database. For the use of the sound files in the SMISS dataset, a database is created, mapping the transcription to the timestamps in the recorded audio files via the dialogue act and turn annotations. Thus for each utterance that may be produced by the NLG module, the correct speech file can be loaded, and the position inside the conversation identified and copied to an output speech IU.

While the speech synthesis modules offer flexibility in the sentences that can be synthesized, it has drawbacks when performing turn-taking with them. Synthesized speech has a generally lower prosodic range, and thus the duration and stress of the synthesized speech often differ from the soundbites of the recorded conversations. Also, the addition of silences before and after the synthesis leads to unwanted changes in the way the agents overlap between turns. During turn-taking, precise overlaps and gaps (usually in the range of $200\,ms$) between the utterances of the two interlocutors need to be achieved. Thus, for the simulations described in this thesis, the TTS module is used in the mode where it uses soundbites from the SMISS dataset as speech output.

In order to aid the ASR module of the other agent, the TTS module adds the text that was synthesized as meta-data to the speech IU. This meta-data will be sent separately through a side channel when the speech is transmitted to the simulated telephone network.

3.3.4 Speech Dispatching

The agents in the simulation need to perform turn-taking, which requires precise timing of speaker overlaps and pauses between the turns. Thus, it is important for the turn-taking dialogue manager to not only monitor the progress of the interlocutor's turn but also to monitor and control its own speech output. The audio dispatching module fulfills both of these tasks.

The audio dispatching module receives the speech that should be spoken by the agent, together with a speech dispatching flag that is controlled by the turn-taking dialogue manager. The speech data is stored in a buffer, and depending on whether the dispatching flag of the incoming incremental module is set, the dialogue manager starts to dispatch either the buffered speech or silence at a predefined speed. That way, the audio dispatching module dispatches audio IUs at all times and alternates between dispatching silence when the dispatching flag is set to *off* by the turn-taking dialogue manager and dispatching the buffered speech when the flag is set to *on*.

The produced speech is sent in small increments to the simulated telephone network, where it is forwarded to the other agent. It is also sent to its own turn-taking dialogue manager, where it is used to monitor the current status of the speech output of the agent.

3.3.5 Turn-Taking Dialog Manager

The turn-taking dialogue manager fulfills the classical task of a dialogue manager by combining the current incoming dialogue act and concepts, the dialogue history, and an agenda of the dialogue to decide what the agent should say in its next turn (in the form of a dialogue act and concepts). However, for the simulation of turn-taking, the agents need to decide when to speak. For this, the module also monitors the progress of the interlocutor's speech to decide when to produce an utterance. For turn-taking, it is also essential for the turn-taking dialogue manager to monitor the progress of its own production of speech. A dialogue manager usually only produces output in the form of dialogue acts and concepts and thus does not have information about when it is producing speech and for how long (as this is usually the task of the speech synthesis module). In this incremental, turn-taking version of a dialogue system, the turn-taking dialogue manager also receives information about its own speech production from the audio dispatching module. This way, it is able to monitor the current speaking status of both sides (i.e., from the interlocutor and itself) to decide when to say what.

To be able to fulfill these tasks, the turn-taking dialogue manager receives three different types of input IUs: the end-of-turn prediction IUs are being received from the end-of-turn prediction module, and dialogue acts and concepts are coming from the NLU module. These two incremental information streams represent the current turn of the interlocutor. The turn-taking dialogue manager also receives the speech IUs of its speech dispatching module to track the progress of its utterances.

Dialogue Management

The dialogue management part of the module is implemented independently of the turn-taking mechanism. The dialogue act selection for every dialogue step is realized as an agenda-based dialogue manager based on Schatzmann et al. (2007), which uses the dialogue acts annotated in the SMISS dataset (see Table 3.2). The dialogue management itself is implemented in a two-tiered process. First, the agenda-based part of the dialogue manager uses the information about the agenda and current requests to select a candidate dialogue act with the corresponding concepts that represent the current action. In the second step, a dialogue act guiding system might modify the concepts referenced in the dialogue act or even the dialogue act itself to be aligned with dialogue acts and concepts already seen in the annotations of the SMISS dataset. This hybrid form of agenda-based and data-driven dialogue management is aimed at producing dialogues based on the structure of the SCT and RNV conversation scenarios while simultaneously modeling exchanges based on real-world data.

In a first step, a stack-based agenda is prepared based on the concepts that should be exchanged during the conversation. These concepts are loaded from an agenda file (see Appendix C) that is specific for each conversation scenario and type of agent (i.e., caller and callee). In the agenda file, the concepts that need to be requested from and given out to the interlocutor are structured in categories. In each category, the information needs to be transmitted before an agent is able to proceed with the next category (e.g., the transmission of the address to deliver the pizza to has to always come after the decision on which pizza to buy). With the concepts from the agenda, a stack is built up that requests or offers information based on the order defined in the agenda file. The dialogue acts `greeting` and `goodbye` are added on top and on the bottom of the stack. During the conversation, new dialogue acts (e.g., answers to requests for information) are put on top of the dialogue stack. In order to avoid unnecessary dialogue acts, the stack is cleaned after every step by removing dialogue acts that have been made obsolete.

Once a candidate dialogue act with optional concepts is selected, the dialogue act guiding system is revising it. A dialogue act with concepts that can be found in the training data will not be modified. When a dialogue act is not seen with the concepts provided, it is modified to either have less concepts (e.g., a `confirm` with the concept `pizza_type` might result in a `confirm` without a named concept) or it is modified to include more concepts (a dialogue act `provide_info` with the concept `toppings` might be paired with the concept `price`, if the dataset contains this annotation). There are dialogue acts that do not occur without a concept (e.g., `request_info` always requires a concept) in the training data. Thus, when the dialogue act and concept combination do not occur in the dataset, the dialogue act itself has to be changed. For example, when the agenda-based part of the dialogue management generates a `offer_info` dialogue act with the concept `pizza_name` (i.e., "*Should I tell you the name of the pizza?*"), it might be a valid combination of dialogue act and concept in theory. However, due to the structure of this particular conversation scenario, this request has never been posed by a participant in the training data. Thus, the dialogue act guiding system rejects that dialogue act, and the dialogue manager needs to fetch the next dialogue act from the stack. In this fashion, the dialogue act and concepts are broadened step by step until a dialogue act with concepts can be located that is contained in the annotated dataset.

Turn-Taking

The turn-taking mechanism is the same for both agents in the simulation and is based loosely on the rules described by Sacks et al. (1974). Grounded in the information from the end-of-turn prediction module, the agent has the information on whether its interlocutor is speaking, while based on its own feedback from the audio dispatching module, it has the information on whether it itself is currently speaking. With these four states, the main behavior of the agents is modeled with the following rules:

1. If the interlocutor is speaking, the agent is listening.
2. If the agent is speaking, it continues to speak until the current turn is finished.
3. If both the agent and its interlocutor speak at the same time, the agent stops speaking.
4. If neither the agent nor the interlocutor is speaking, the agent determines when to speak next.

While these rules describe a general interaction, two rules have to be modified in order for real turn-taking to take place. First, the third rule needs to be adapted in order for natural overlaps to occur during the changing of the active speaker. That is why the definition of "both the agent and its interlocutor speak at the same time" needs to be adapted. Because of possible overlaps, the agent will only stop speaking when the double talk is happening in the *middle* of its own turn. This is defined as being speech that is not in the first and last second of the utterance of the agent. Thus, small overlaps during turn-taking do not lead the agent to stop with its utterance.

For smooth turn-taking with gaps and overlaps in between the speaker changes, the fourth rule needs to be specified. At the end of every utterance, the agents need to independently decide if a turn-transition should occur or if the current speaker should keep the turn. For this, the concepts of Lunsford et al. (2016) are employed, in which turn-continuations and turn-transitions are considered equal alternatives. This means that for every turn, the two agents are filling competing roles. The current speaker is determining how long the pause between its current turn and its next turn should be, while the listening interlocutor determines when to take over the turn (either by a short overlap of the speaker's turn or by a gap after the current turn has ended). Depending on which agent speaks first, the turn-taking is negotiated without the need for a synchronization other than the speech itself. Both the turn-continuation and the turn-transition are modeled based on the seconds since the previous turn has ended. However, the turn-transition point might be negative to result in overlaps.

For the baseline simulation, the turn-continuation is kept statically at 2 seconds, while the turn-transition point is set to 1 second. This results in a simulation where only turn-transitions occur, as the pause that the current speaker makes is always longer than the transition timing of the interlocutor.

3.3.6 Data Logging

For the simulation to be evaluated, data needs to be extracted from every conversation to be analyzed. For this, data from the speech, text, and concept layer of the simulation are extracted.

An audio recorder module is placed at the outputs of the speech dispatching modules of each agent to capture the clean speech (i.e., speech that has not been altered by the simulated telephone network module) and at the outputs of the telephone network simulator modules to capture the degraded speech. The fullband audio is stored in wave files that can be used for parametric conversation analysis.

The transcriptions of the utterances of both agents are stored in a dialogue file. For this, a text recorder module is connected to the output of the NLG module of each agent. The text recorder module only records the uttered turns (turns where the dispatch flag was set) and adds the agent's type (caller or callee) and the timestamp to every transcription.

The simulated conversations are also recorded on the dialogue act level. For this, the produced dialogue acts of the turn-taking dialogue manager of each agent are collected by a dialogue act recorder module. Analogous to the text recorder module, the dialogue act recorder adds the agent's type and the timestamp.

3.3.7 Simulated Telephone Network

The simulated telephone network consists of two network modules that are unidirectional. Each network module may contain one or more degradations that are applied to the speech signal before it gets inserted into the right buffer. In addition to the degradations, the network maintains a side channel, where the transcriptions provided by the NLG module of the sending agent are being transmitted. This side-channel information is tied to the incoming speech IU so that a delayed transmission of incoming speech results in the same delay of the side channel.

The delay degradation component of the simulated telephone network contains a buffer that is filled with the incoming speech IUs. Then, once the buffer is filled to the size determined by the provided one-way transmission delay, the speech is taken from the buffer in order of first arrival. This results in the delay of the speech signal by the provided amount.

The packet loss degradation component requires the packet loss probability and the burst ratio as arguments. Then, following Equation 2.16 of the narrowband E-model (ITU-T Recommendation G.107, 2015), the p and q value for the two-state Markov model are calculated:

$$q = \frac{1 - \frac{Ppl}{100}}{BurstR} \tag{3.1}$$

$$p = \frac{\frac{Ppl}{100} \cdot q}{1 - \frac{Ppl}{100}} \tag{3.2}$$

where Ppl is the packet loss probability and $BurstR$ is the burst ratio. For each incoming IU the Markov model is used to determine whether the packet will be lost. The speech of the affected incoming IUs is set to zeroes, resulting in silence.

3.4 Evaluation of the Simulation Architecture

In order to evaluate the dialogue management and to test whether turn-taking is necessary to reproduce the interactivity of a conversation, 100 SCT scenario 11 conversations and 100 RNV scenario 1 conversations are simulated and compared to the SMISS dataset for analyzing the content of the dialogue and to the CONVSIM dataset to analyze the conversations on the signal level. While the turn-taking mechanisms have been implemented, but no turn-taking model has been inserted into the turn-taking dialogue manager, the simulation is operating in turn-steps (i.e., with a one-second pause in between each turn).

First, to compare the contents of the real conversations with the simulated dialogue, the distribution of dialogue acts is compared with the KL dissimilarity measure. Then, a P-CA is performed on the recorded speech of the CONVSIM dataset and the simulations. The comparison shows similarities and shortcomings of this baseline version of the simulation in turn steps.

3.4.1 Dialogue Act Evaluation

The dialogue acts of the simulation are compared to the annotations of the SMISS dataset. In this dataset, only SCT scenario 11 and RNV scenario 1 conversations are included, and thus, the simulation is expected to be similar in content. However, due to the structure of SCT and RNV conversation scenarios, the distribution of dialogue acts in other scenarios is expected to be comparable.

Figure 3.6 shows the average occurrences of each dialogue act in the empirical and simulated dialogue for the SCT scenario 11 conversations. Overall, the simulation matches the distribution of dialogue acts in the empirical data closely. However, the empirical data has more variance (as shown by the 95 % confidence interval in Figure 3.6). The dialogue act `misunderstanding`, while being very uncommon in the empirical data, is not present in the simulation. This is due to the misunderstandings not being implemented in the baseline simulation. The modeling of misunderstandings is described in Chapter 5. The dialogue acts `request_info` and `provide_info` have a higher occurrence rate in the simulations than in the SMISS dataset. Here, the reason is the lack of turn-taking in the simulation. Because the turn-taking model implements a fixed higher probability for a turn-transition, the active speaker has to change after each turn (i.e., after each uttered dialogue act). This leads to some unnatural requests for information that would have been answered if no speaker change had occurred.

As the greeting and farewell are always integrated into the stack of the agenda-based dialogue, these `greeting` dialogue act occurrences have no variance over the simulations (indicated with a missing error bar).

Figure 3.7 shows the dialogue act distribution in the RNV scenario 1 conversations of the SMISS dataset and the simulations. Compared to the SCT scenario 11, the conversation is much less diverse and very structured. Again, the simulated conversations are able to reproduce the distribution of the empirical data for almost all dialogue acts. As

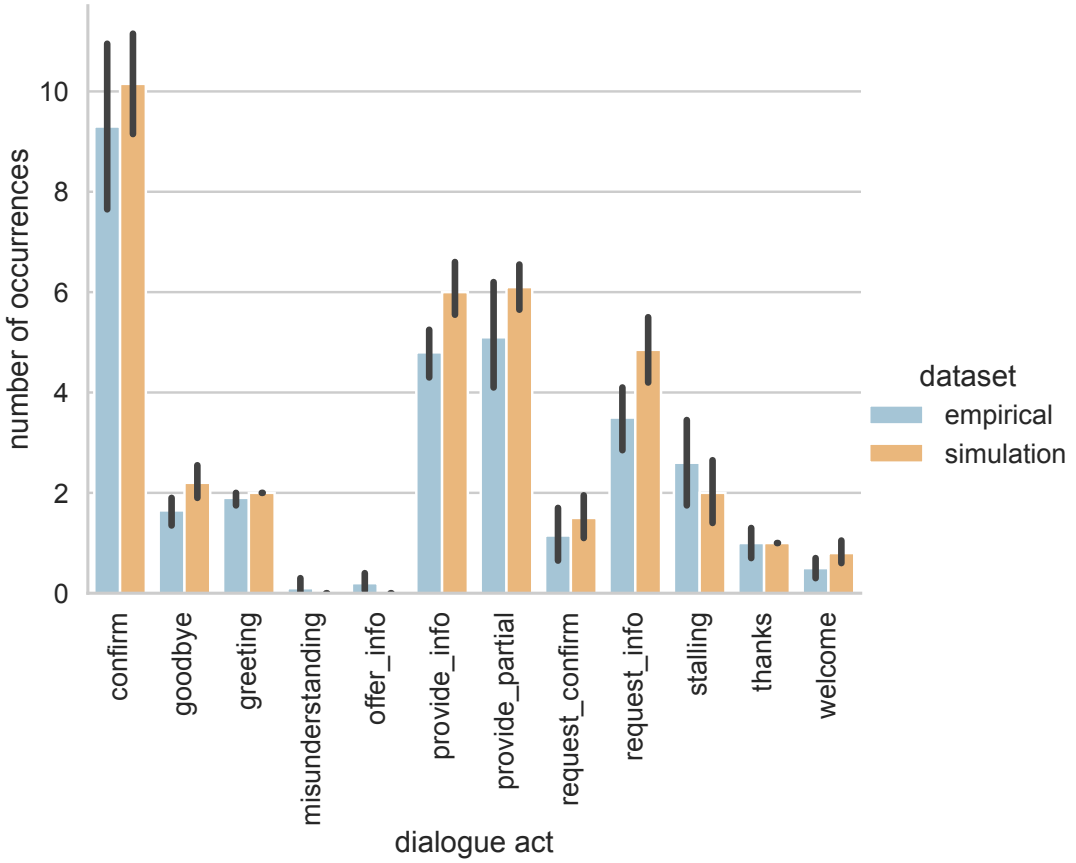

Fig. 3.6 Occurrences of dialogue acts in the SCT 11 conversation of the empirical data and the simulation. Error bars indicate the 95 % confidence interval.

with the SCT simulation, due to the nature of the agenda-based dialogue management, the `greeting` dialogue act is fixed at two occurrences per simulated conversation (one for each interlocutor).

The dialogue act `stalling` occurs more frequently in the simulated conversation than in the SMISS dataset. Here, as with the SCT simulations, the missing turn-taking results in the addition of `request_info` and `provide_info` dialogue acts. However, these are not present in the dataset, and thus, the dialogue act guidance system of the turn-taking dialogue manager rejects the dialogue act. Because the turn-taking dialogue manager needs to produce an utterance (because of the forced turn-transition in this baseline), the fallback implemented in the dialogue manager is the `stalling` dialogue act. Only then the interlocutor may continue with their agenda. This mismatch occurs between every block of numbers in the RNV scenario, as one participant confirms the last number of their interlocutor and then starts reading the next row of numbers (in a turn-continuation). This results in 5 additional `stalling` dialogue acts in each RNV conversation.

In order to evaluate not only the plain occurrences of the dialogue act but also the order in which they occur, sequences of dialogue acts were extracted from the simulations and the empirical data. For this, the number of occurrences of single dialogue acts (1-gram),

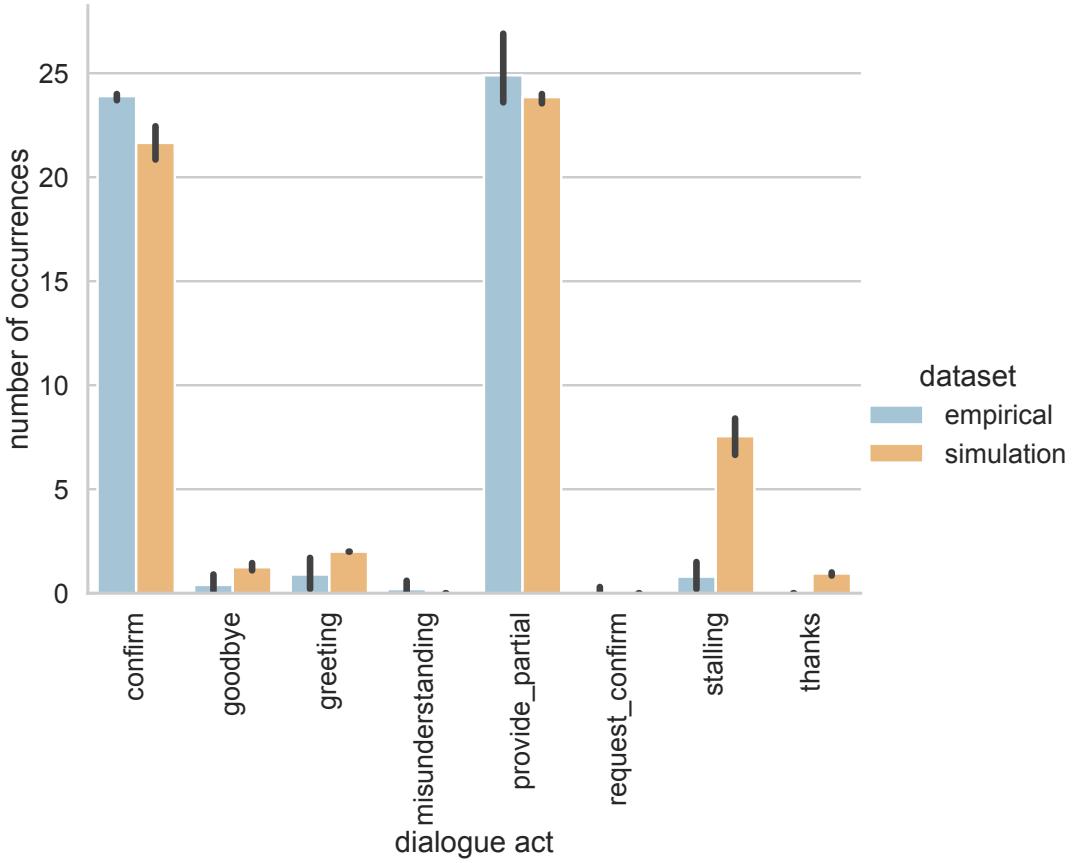

Fig. 3.7 Occurrences of dialogue acts in the RNV 1 conversation of the empirical data and the simulation. Error bars indicate the 95 % confidence interval.

question and answer pairs (2-gram), and full turn cycles (3-grams) was considered. These n-grams were constructed independently of which speaker uttered the dialogue act, as only the order in which they were spoken is relevant for this analysis.

Table 3.3 The Kullback-Leibler dissimilarity for n-gram probability distributions with n-grams of size 1, 2 and 3 of the natural conversations given the simulated conversations and the entropy of the natural conversations for each of the n-grams. Both split into SCT and RNV conversations.

	KL distance		Entropy	
n	**SCT**	**RNV**	**SCT**	**RNV**
1	0.0159	0.2191	2.9407	2.0948
2	0.4832	0.3763	5.0441	3.0191
3	0.5110	0.4384	6.3933	3.6678

The distribution of n-grams was then compared with the cross-entropy and KL dissimilarity metric (Equation 2.34). For the calculation, the binary logarithm was used, so the results represent the number of bits lost if the n-grams of the empirical conversations are approximated with the simulated ones. The distributions P and Q were modeled by the relative occurrences in the simulation and the empirical data. Table 3.3 shows the KL distance and cross-entropy for the SCT and RNV n-gram distributions.

The KL distance between the two distributions is very low for dialogue act occurrences themselves, which is in line with the distributions shown in Figure 3.6 and 3.7. The KL distance of the dialogue act 2-grams is higher with 0.48 for SCT and 0.37 for RNV conversations. This increase is expected due to the increase in the 2-gram possibility space and can be interpreted as half a bit of information missing when encoding one distribution with the other. For 3-grams, which encodes whole turn cycles, the KL distances increase slightly. This trend can also be observed in the cross-entropy, which represents the number of bits needed to encode the distribution of the empirical data with the n-grams of the simulation. Generally, the SCT conversations have a slightly higher distance and cross-entropy than the RNV conversations. This is an indicator of the more complex and diverse structure of the SCT task compared to an RNV conversation. While the KL dissimilarity and cross-entropy cannot be interpreted as absolute measurements, the interpretation of the underlying information theory shows a substantial similarity between the two n-gram sets.

The analysis of the distribution of dialogue acts and the distance between the n-gram distributions shows that the agenda-based approach, in combination with the data-driven dialogue act guiding system, is able to model the two types of conversations sufficiently. Small deviations in the occurrences of dialogue acts can be attributed to the limited turn-taking capabilities of this baseline implementation.

3.4.2 Interactivity Evaluation

The interactivity of the simulation is evaluated by comparing the conversational parameters obtained by a P-CA. In this evaluation, the simulations are compared to the CONVSIM dataset, which includes more conversation scenarios than the ones simulated. Generally, this will lead to more variance in the empirical data, as the simulation only reflects one conversation scenario for each conversation type. Because the simulation has a fixed length of silence in between turns, differences in interactivity between SCT and RNV conversations are to be expected.

Figure 3.8 shows the state probabilities for mutual silence, double talk, speaker A, and speaker B for the simulated and empirical SCT and RNV conversations. The state probability MS is higher in the simulations, which can be explained by the static, long silence in between the turns of the simulation. As in the empirical data, the simulated RNV conversations have more silence than SCT conversations. However, this effect is very pronounced in the RNV simulations, leading to over 70 % of the conversation in this state. The state probability DT is not present for the simulation, as, with the fixed turn duration of one second, no overlap can occur. The empirical data shows that both SCT and RNV conversations have high variances in the overlaps, but generally, only 3–4 % of the conversation consist of these states. The state probabilities for SA and SB are lower for the simulations, as the state probability for MS is taking up most of the conversation. This effect is more pronounced for the RNV conversations, as the long turn-taking pauses make up a larger proportion of this task. The empirical data shows that both speaker A and speaker B share a similar proportion of the conversation with 30 %.

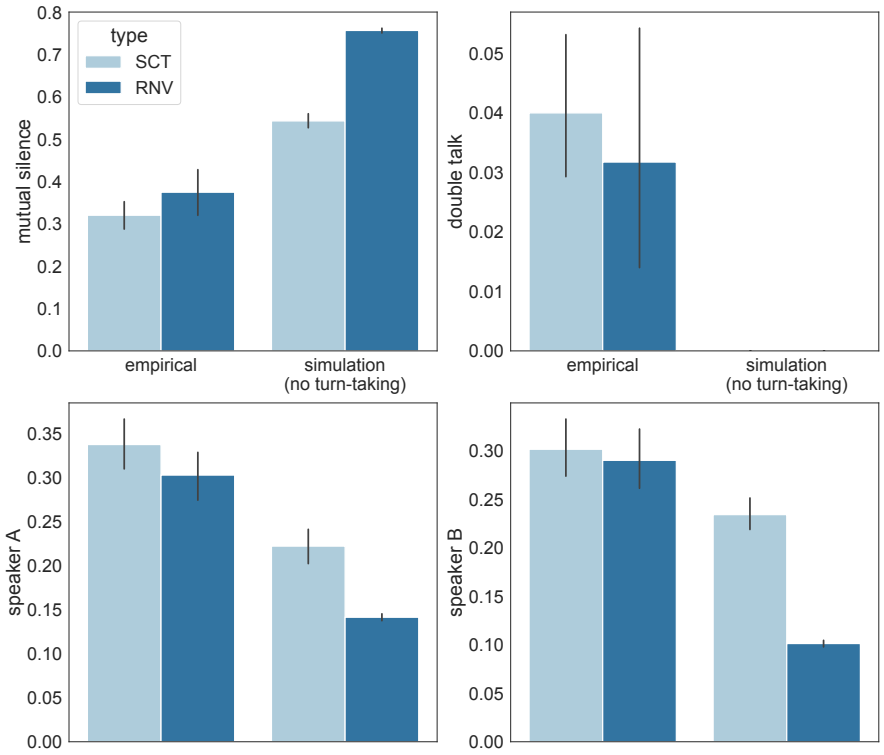

Fig. 3.8 State probabilities for mutual silence, double talk, speaker A, and speaker B for the empirical data and the simulations with no turn taking, split by SCT and RNV conversation types. Error bars indicate the 95 % confidence interval.

The sojourn times for the four states can be seen in Figure 3.9. For the sojourn time of MS, the fixed delay at one second for both simulated SCT and simulated RNV conversations can be seen. The slight deviations in the mutual silence can be explained by small areas of silence at the beginning and end of the utterances. The sojourn times of DT are non-existent for the simulations, as there is no overlap between speakers. The empirical data, however, shows shorter overlaps for RNV conversations, suggesting higher interactivity of this type of conversation. The sojourn times for speaker A and speaker B are very similar for both simulated and empirical conversations. For SCT conversations, the average length of SA and SB is around one second, while for RNV conversations, it is roughly half a second. Due to the simulation of the contents of these two types of conversations, this difference in utterance length is reflected in the simulations. Together with the state probabilities, it can be reasoned that the mismatch between the empirical and simulated data in the probabilities of SA and SB can be explained by the large pauses between turns alone, as the sojourn time of these two states indicates an accurate replication of the empirical data.

Figure 3.10 shows the Speaker Alternation Rate (SAR) for the simulated and empirical SCT and RNV conversations. Due to the mix of short utterance (i.e., sojourn times of SA and SB) and a smaller state probability of Mutual Silence, the SAR of the empirical RNV conversations is more than twice as high as for SCT conversations. While there is a difference in interactivity between the simulated SCT and RNV conversations, it is not as strong and can mostly be attributed to the shorter sojourn times of SA and

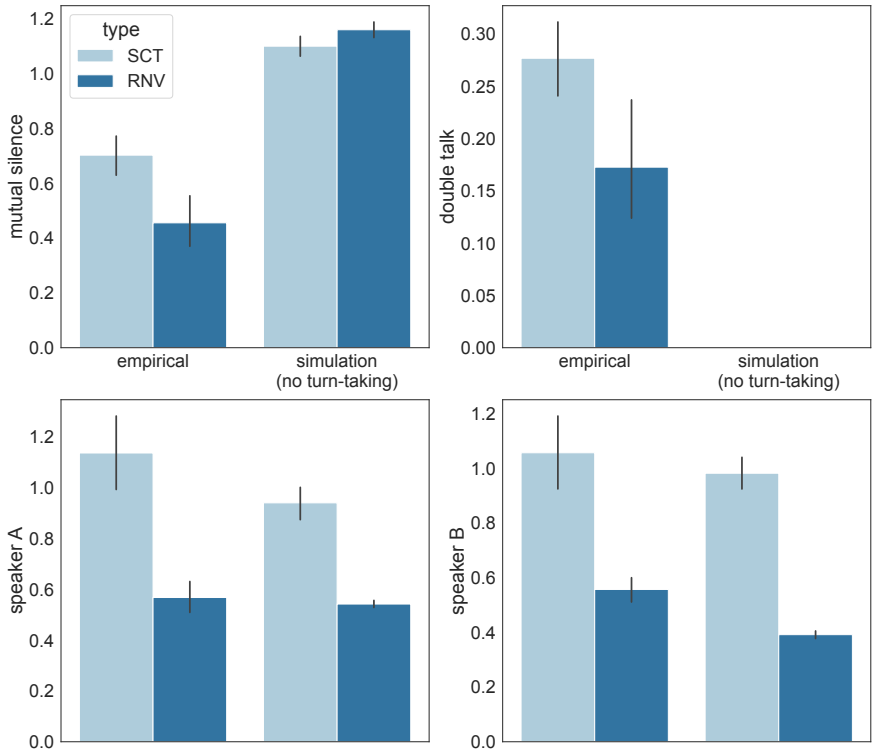

Fig. 3.9 Sojourn times for mutual silence, double talk, speaker A, and speaker B for the empirical data and the simulations with no turn taking, split by SCT and RNV conversation types. Error bars indicate the 95 % confidence interval.

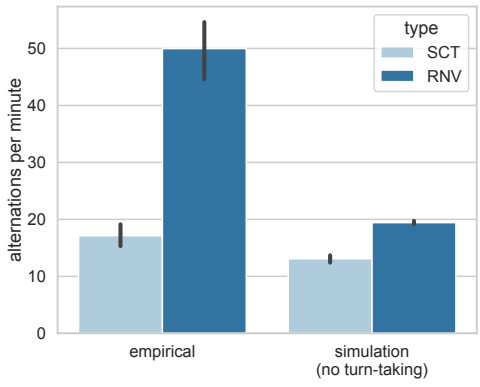

Fig. 3.10 Speaker alternations per minute for the empirical data and the simulation without turn taking, split by RNV and SCT conversation types. Error bars indicate the 95 % confidence interval.

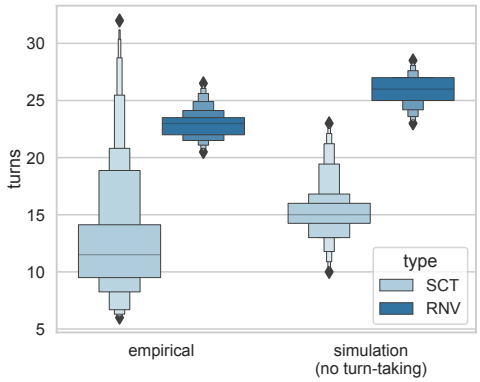

Fig. 3.11 Number of turns in a conversation for the empirical data and the simulation without turn taking, split by RNV and SCT conversation types.

SB. The simulation has less variance in the SAR compared to the empirical data, which can be partly explained by the large number of conversation scenarios included in the CONVSIM dataset.

Figure 3.11 shows the number of turns for the simulated and empirical SCT and RNV conversations. Due to the accurate modeling of the contents of the conversation, the overall turn count and the difference in turns between SCT and RNV conversations are matched well. However, the simulations have a slightly higher number of turns due to the extra requests required by the lack of turn-taking. Again, lower variance in the simulated data is present.

The interactivity analysis shows that the simulation is able to reproduce the general interactivity based on SAR, sojourn times, state probabilities, and the number of turns. A small but significant difference in the SAR can be seen between the simulated SCT and RNV conversations. However, without a proper turn-taking model, the simulation cannot match the difference in the interactivity of the two conversation types.

3.4.3 Simulation Performance

The incremental simulation architecture works based on real-time execution of the incremental dialogue systems and the connecting simulated telephone network. Thus, the run time of each simulated conversation is proportional to the length of that conversation. The incremental architecture decouples the processing of each incremental module, and the synchronization is achieved with the transmission of incremental units. Each incremental module runs in a separate computing thread, resulting in processing utilization of 15 % on average for a single simulation on a 2.6 GHz 6-core Intel Core i7. The overall memory consumption of a single simulation is about 500 MiB, as the speech files need to be kept in memory for faster processing.

The structure of independently working incremental modules connected with buffers between them adds additional processing overhead. While the modules in the simulation do not perform a computationally expensive operation, the transmittance of incremental units adds a delay of 5–10 ms per agent, which results in an overall processing delay of a maximum of 0.02 seconds at real-time processing speed. As the increase of the clock speed increases the processing of the individual modules, the effective processing delay increases linearly with the processing speedup. In practice, a simulation at twice the real-time speed still experiences a processing delay of 0.02 seconds for each transmitted utterance from one agent to the other, resulting in an effective delay of 0.04 seconds. Because of this constraint in the parallelization, the processing is set to real-time speed for all simulations described in this thesis.

3.5 Summary

In this chapter, an incremental conversation simulation is presented that is able to replicate the conversation scenarios SCT and RNV standardized by the ITU-T. The simulation is based on retico, a framework for incremental processing specifically built for this simulation that multiple universities have used in teaching and research since its creation. The simulation architecture is extendable and provides the necessary interfaces to model turn-taking, misunderstandings, and potentially other effects of transmission impairments on the conversation. The simulation is comprised of two virtual agents

that communicate with each other over a simulated VoIP transmission network. These agents are implemented with the same behavioral patterns and only differ in the dialogue management. As part of the agents, the system uses a new turn-taking dialogue manager approach, where one incremental module plans both what and when to speak. However, the current turn-taking mechanism implements an interaction in turn steps, validating the necessity of incremental turn-taking. The dialogue managing part of the module is based on a goal-oriented agenda-based dialogue manager that uses a stack of dialogue acts to plan the conversation and react to new input. The general dialogue manager is instructed to produce conversations of SCT and RNV type by providing conversation files that include all the concepts that should be given to and requested from the conversation partner. For the turn-taking part of the module, the turn-taking dialogue manager uses predictions from an end-of-turn module that predicts the time until the end of the utterance of its interlocutor. In addition, the turn-taking dialogue manager uses the information of its own speech dispatching module to know about the status of its current output. Together, these two information sources are accessible to form hypotheses about the current state of the turns in the dialogue. For natural language generation speech synthesis, the two virtual agents make use of training data recorded in real conversations. However, state-of-the-art speech synthesis algorithms may also be employed.

The evaluation of this simulation architecture and approach, in general, shows that the specific implementation is able to reproduce the contents of a conversation on the dialogue act level and the general differences in turn lengths on the interactivity level. The distribution of dialogue acts between simulated and empirical conversations shows very similar characteristics, and the n-gram KL distance between two dialogue act distributions is small. However, it has been shown that for accurate replication of a conversation's interactivity, the simulation needs to implement a form of timely turn-taking, as the simulation in turn-steps fails to replicate the distinct levels of interactivity. A final performance evaluation has shown that the simulation is able to run in real-time on standard hardware.

Chapter 4
Simulating Interactivity and Delay

The evaluation of the simulation architecture with the baseline turn-taking model has shown to be suitable for use in conversation simulation. However, the differences in interactivity between the SCT and RNV conversation cannot be reproduced without the implementation of a turn-continuation and a turn-transition model. Especially for transmission delay, it has been shown that conversations with different levels of conversational interactivity also degrade differently. Also, the delay itself is not audible but instead impacts the interactivity of a conversation. Thus, to simulate the difference in interactivity, as well as the effects of transmission delay on a conversation, it is necessary to reproduce the differences in turn-taking.

In this chapter, the key parameters to model and analyze the difference in turn-taking between SCT and RNV conversations are identified. Then, the turn-taking mechanism described in Chapter 3 is extended by a turn-continuation and a turn-transition model that is designed to replicate those differences. For this, the turn-taking behavior of the participants in the SMISS dataset is extracted and analyzed. After validation of the new turn-taking behavior, a one-way transmission delay is added to the simulated telephone network, and the performance of the simulation is evaluated on the CONVSIM dataset. After an adaption of the turn-taking model for conversation with transmission delay, a final evaluation of the turn-taking and interactivity properties of the simulation is performed.

The modeling of different interactivity levels in the simulation has been in part published in Michael and Möller (2020d) and the simulation of the changes in interactivity due to transmission delay has been described in Michael and Möller (2020c).

4.1 Simulating Turn-Taking in Conversations with Varying Interactivity

Research has shown that different conversation types have distinct levels of conversational interactivity (Raake et al., 2013; Egger et al., 2012). This overall interactivity can be measured with parameters like the SAR or the conversational temperature, which can be extracted by a P-CA. These parameters and thus the interactivity of a conversation is influenced by two factors: the length of each turn and the turn-taking behavior of the interlocutors. As the length of each turn in a conversation dictates how often speaker alternations and therefore active turn-taking can take place, this factor produces the

basic interactivity. For example, if a conversation is made up of very long utterances, quick turn-taking would not result in increasing the interactivity by much. However, for conversations like the RNV task, where speaker turns are short, a difference in turn-taking has a larger impact on the overall interactivity.

As shown in the interactivity evaluation in Section 3.4.2, the length of utterances of each speaker in the simulation accurately reflects the turn lengths in the empirical data. However, with no turn-taking in place, the interactivity of the simulated conversations is too low. Thus, the timing of the turn-taking in the SCT and RNV conversations has to be analyzed in order to model the differences in interactivity accurately.

4.1.1 Turn-Taking on a Conversation Level

To model turn-taking in the simulation, the turn-continuations and transitions between speakers in empirical conversations have to be examined. For this, the SCT and RNV conversations of the SMISS dataset have been analyzed with a P-CA. Similar to the turn-taking analysis of Lunsford et al. (2016), the extracted conversational states are used to determine the timings of turn-transitions and turn-continuations relative to the end of the previous utterances. Gaps are determined by the mutual silence in between a speaker transition and overlaps are determined by the double talk between a speaker transition (measured negative since the beginning of the new turn starts before the last turn has ended). Finally, pauses are determined by the mutual silence between the utterances of the same speaker. For this analysis, a turn-continuation is defined as a pause of at least 400 ms between the utterances of the same speaker, and the gaps, overlaps, and pauses have been averaged over each conversation. With this preprocessing, an analysis of the timing of turn-taking on the conversation level can be achieved.

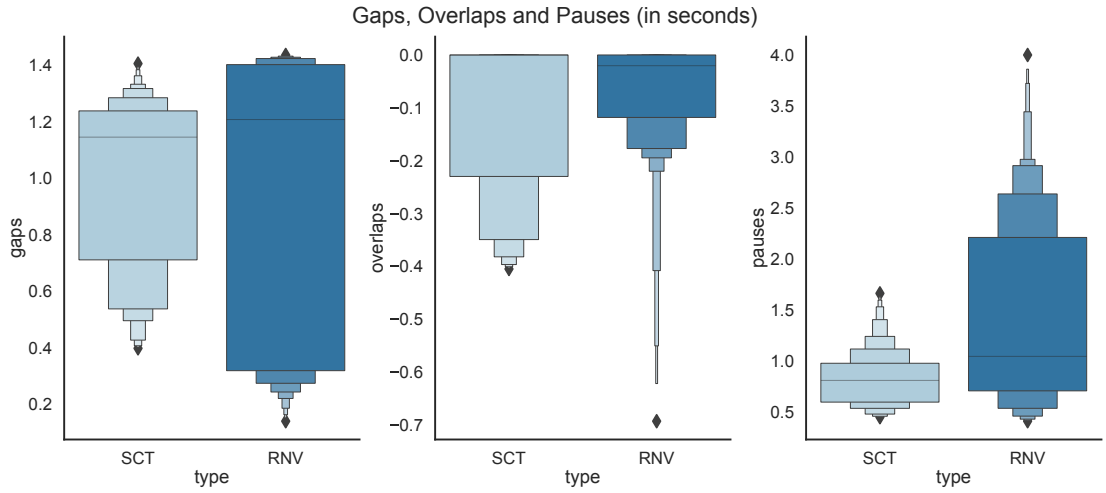

Fig. 4.1 Distribution of gaps, overlaps, and pauses averaged over SCT scenario 11 and RNV scenario 1 conversations in the SMISS dataset. Lengths are given in seconds relative to the end of the previous turn.

Figure 4.1 shows the distribution of gaps, overlaps, and pauses for the SCT and RNV conversations of the SMISS dataset. The median lengths of the gaps and overlaps of the RNV conversations are longer compared to the ones of the SCT conversations. The distribution of gaps show that for SCT conversations, they lie between 0.8 and 1.2 seconds, while for RNV conversations they center around 0.3 to 1.4 seconds. The distribution of the overlaps shows a greater variance for the RNV conversations, with the average overlap length in a conversation being 0.7 seconds before the end of the previous turn. The pauses also show a higher variance in the RNV conversations, with some conversations having very short pauses in between speaker turns and others having long pauses of up to 4 seconds on average. In contrast, the pauses in the SCT conversations lie around 0.5 to 1.0 seconds, indicating a more homogeneous pausing behavior in between turns.

Overall the RNV conversations have a higher variance in the turn-taking behavior. Gaps are shorter on average, although the median gap length is increased. These differences point toward a fundamental difference in turn-taking behavior between those two conversation scenarios that needs to be modeled.

4.1.2 Modeling Turn-Taking on the Interaction Level

In order to implement turn-taking in the simulation, the turn-transitions and -continuations need to be modeled on the level of individual speaker alternations and pauses. In the simulation architecture, an interface for two separate models was implemented. One model is active during and shortly after the turn of the simulated interlocutor to decide when a turn-transition should occur. The other model is active after the agent's own turn is completed to determine when a turn-continuation should occur. These two competing models then determine when, relative to the (predicted) end of the current turn, a transition or continuation should occur. Depending on which agent starts to speak, the other agent detects the beginning of a new turn, and the turn-taking models of each agent are restarted.

The gaps, overlaps, and pauses of the SMISS dataset were extracted for every turn-taking instance (i.e., transition and continuation), as opposed to the conversation level information for the analysis. In order to model the transitions and continuations as equal alternatives, the gaps and overlaps are combined in a *turn-transitions* metric that represents the instances of gaps and overlaps in one distribution, with overlaps being negative and gaps being positive.

Figure 4.2 shows the distributions of turn switches and pauses for the SCT scenario 11 conversations of the SMISS dataset. The distribution of turn switches is centered around 0.27 seconds, with a slight skew towards the overlaps. Due to the fact that pauses of a speaker longer than 0.4 seconds are considered turn-continuations, the distribution of pauses has its maximum at that value.

Figure 4.3 shows the same distributions for the RNV scenario 1 conversations. For the turn-transitions, the distribution of gaps and overlaps is much narrower compared to the turn switches of the SCT conversations. Here, the short overlaps of up to 0.5 secodns are more pronounced, but longer overlaps (over 1 s) do not occur. In the turn-switches, there is a slight increase of occurrences at one second. The distribution for

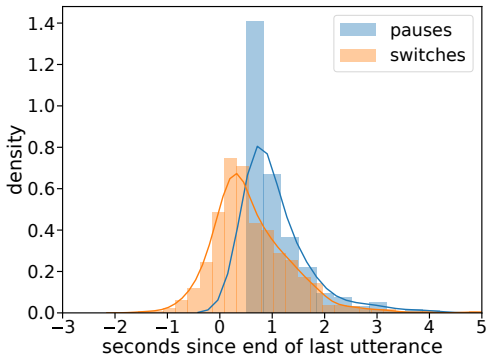

Fig. 4.2 Turn-transitions (labeled *switches*) and turn-continuations (labeled *pauses*) in conversations of the SCT scenario 11 relative to the end of the utterance in the previous turn.

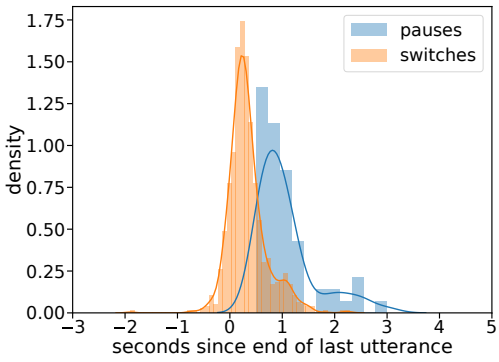

Fig. 4.3 Turn-transitions (labeled *switches*) and turn-continuations (labeled *pauses*) in conversations of the RNV scenario 1 relative to the end of the utterance in the previous turn.

turn-continuations is also narrower as compared to the SCT conversations. However, it is not as concentrated on the 0.4 second cutoff but more distributed. This may indicate the turn-continuations between blocks of numbers, where the last number is confirmed, and the next row begins with that same speaker.

The distribution of individual turn-continuations and -transitions shows that the distributions of both turn switches and pauses of the RNV conversations are narrower and shifted in comparison to the SCT scenario. Thus, the turn-taking model in the simulation needs to take the difference in the behavior into account.

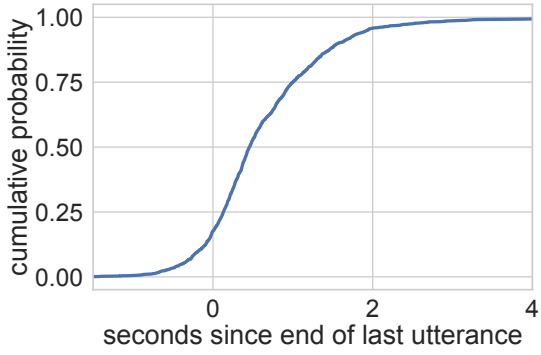

Fig. 4.4 Cumulative probability of the gaps and overlaps of the SCT conversations relative to the end of the utterance in the previous turn.

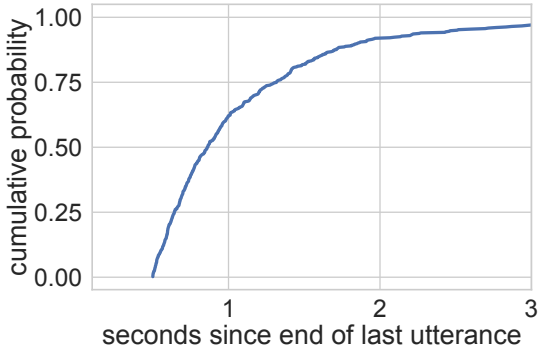

Fig. 4.5 Cumulative probability of the pauses of the SCT conversations relative to the end of the utterance in the previous turn.

In order to model the distributions of gaps, overlaps, and pauses of the two conversation scenarios, the cumulative probabilities of the distributions were calculated. The cumulative distribution of the gaps and overlaps in the SMISS conversations are shown in Figure 4.4 and the distribution of the pauses is shown in Figure 4.5. Both figures show the distribution for the SCT conversations – the cumulative distributions of the RNV conversations are analogous, however the distribution for gaps and overlaps are steeper. For points relative to the end of the previous utterance (x-axis), the distribution

shows the probability that a turn-transition (for Figure 4.4) or a turn-continuation (for Figure 4.5) has taken place for that value or a lower value. Thus, by randomly sampling from the probability values on the y-axis, the original distribution is reconstructed with the according values on the x-axis.

In order to fit the cumulative probabilities of turn-transitions (Figure 4.4), a logistic function of the form $f(x) = \frac{L}{1+e^{-k(x-x_0)}}$ is used to approximate the distribution. For the probabilities of turn-continuations (Figure 4.5), a quadratic function in the form of $f(x) = a \cdot x^{\frac{1}{2}} + b$ is used. With these functions, the cumulative turn-transition and turn-continuation distributions of the SCT and RNV conversations are fitted with a least-squares approach. The resulting functions are then inverted, in order for them to map from probabilities to seconds since the end of the previous utterance:

$$\hat{T}_{SCT} = -0.3226 \cdot \log(0.433 \cdot (-1 + \frac{1}{x})) \tag{4.1}$$

$$\hat{T}_{RNV} = -0.1598 \cdot \log(0.17 \cdot (-1 + \frac{1}{x})) \tag{4.2}$$

$$\hat{C}_{SCT} = 0.9251 \cdot (0.8432 + 2.9231 \cdot x^2) \tag{4.3}$$

$$\hat{C}_{RNV} = 1.3876 \cdot (0.3607 + 1.2007 \cdot x^2) \tag{4.4}$$

where $\hat{T}_{SCT}$ and $\hat{T}_{RNV}$ are the models of the turn-transitions for the SCT and RNV conversations respectively and $\hat{C}_{SCT}$ and $\hat{C}_{RNV}$ are the models of the turn-continuations for the SCT and RNV conversations respectively, with x being the cumulative probability between 0 and 1. By randomly sampling from these distributions, the original distributions of turn-transition and turn-continuation timings shown in Figure 4.2 and 4.3 are approximated.

Before these models can be used in the simulation to calculate the desired turn-transition and turn-continuation timing, the differently modeled behavior for SCT and RNV conversations need to be unified into one model. Based on the work of Heeman and Lunsford (2017), where turn-taking is analyzed depending on the context of the conversation, the different models for switches and pauses are conditionally selected by the dialogue act that preceded it. As the RNV conversations mostly consist of the `confirm` and `provide_partial` dialogue acts, turn-continuations and -transitions that occur after those dialogue acts are modeled with the RNV turn-taking estimators ($\hat{T}_{RNV}$ and $\hat{C}_{RNV}$). For all other dialogue acts, the turn-taking mechanism of the simulation is extended to use the SCT turn-taking models ($\hat{T}_{SCT}$ and $\hat{C}_{SCT}$).

With the new turn-taking mechanism, the turn-transition time of the currently listening agent and the turn-continuation time of the currently talking agent is calculated based on a random sample calculated from the respective model. With turn-transition and turn-continuations as equal possibilities, the turn-transition time of the random sample decides whether the active speaker is continuing their turn or if the listener is taking the next turn with either an overlap of speech or with a gap.

4.1.3 Evaluation of the Turn-Taking Model

In order to evaluate the improved turn-taking model, 100 SCT scenario 11 conversations and 100 RNV scenario 1 conversations are simulated and compared to the CONVSIM dataset. The simulated conversations are analyzed on the signal level by extracting the interactivity parameters with a P-CA. To highlight the changes due to the turn-taking model, results of the simulations without turn-taking are included. The empirical data in the CONVSIM dataset includes more scenarios for both SCT and RNV, and thus, this data has a higher variance than the simulated data, which consists of one scenario of each type.

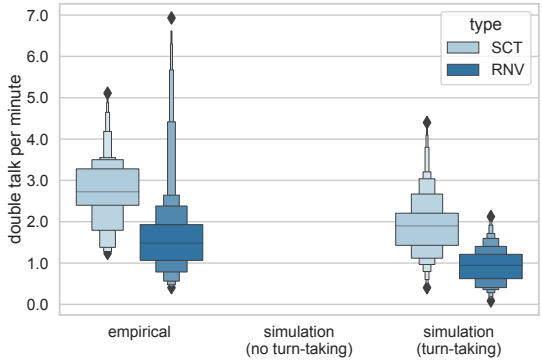

Fig. 4.6 Double Talk Rate (DTR) of the empirical and simulated SCT and RNV conversations, both with and without the improved turn-taking mechanism.

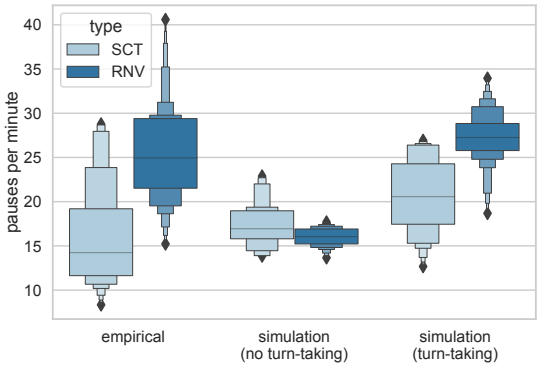

Fig. 4.7 Pause Rate (PR) of the empirical and simulated SCT and RNV conversations, both with and without the improved turn-taking mechanism.

Figure 4.6 shows the overlaps in the form of Double Talk Rate (DTR) of the CONVSIM dataset, the simulation without turn-taking, and the simulation with turn-taking. In contrast to the simulation in turn steps, the new turn-taking model is able to reproduce the difference in double talk. However, the occurrences of double talk per minute are slightly lower, with about one instance of double talk per minute less. The difference in DTR between the SCT and RNV scenarios visible in the empirical data is also visible in the simulated conversation.

Figure 4.7 shows the Pause Rate (PR) of the CONVSIM dataset and the simulations. While the PR of the simulation without turn-taking failed to reproduce the difference between RNV and SCT conversations, the simulation with turn-taking is able to model the higher PR in RNV conversations. In contrast to the DTR, the PR is too high in the turn-taking simulation, indicating that the implemented mechanism prefers turn-continuations over turn-transitions. However, with the turn-taking, the variance in the data is closer to the empirical conversations than the PR of the simulation in turn-steps is.

Figure 4.8 shows the lengths of the simulated conversations compared to the empirical dataset. Here, the addition of the turn-taking mechanism dramatically reduces the duration of the conversation while also increasing the variance. Due to the shorter utterances and faster turn-taking, the RNV conversations are more concise than the SCT

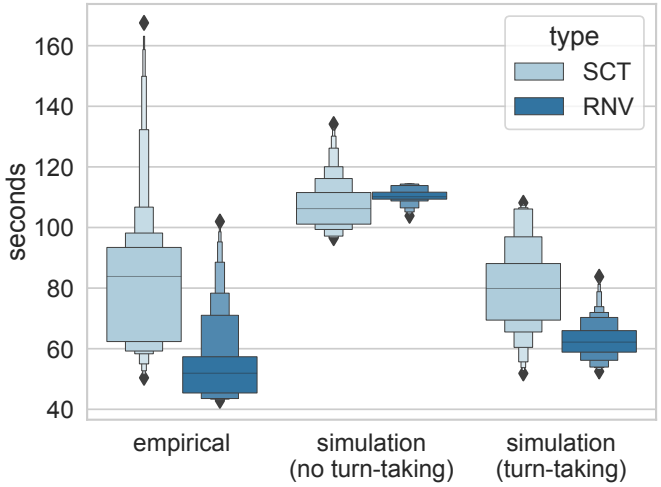

Fig. 4.8 Comparison of conversations length between the data of the CONVSIM dataset, the simulated conversations without turn-taking and the simulation with turn-taking.

conversations. The simulated RNV conversations are slightly longer than their empirical counterparts, which might indicate that the pauses and gaps in the RNV turn-taking model are too long.

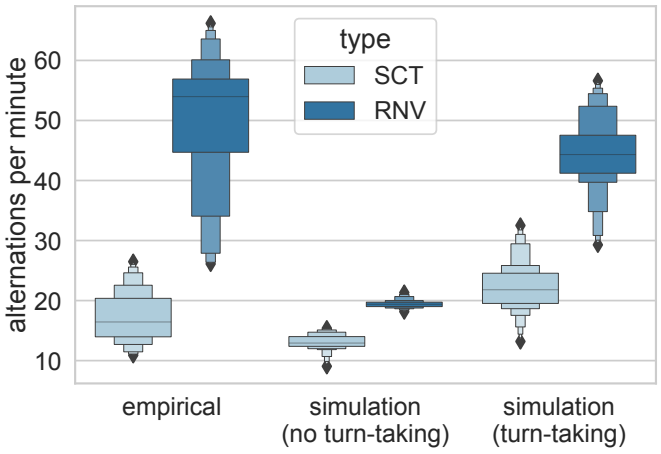

Fig. 4.9 Comparison of Speaker Alternation Rates between the data of the CONVSIM dataset, the simulated conversations without turn-taking and the simulation with turn-taking.

Finally, the Speaker Alternation Rate (SAR) shown in Figure 4.9 shows a clear difference in the overall interactivity of the two scenarios as well as an increase in variance. While the simulation without turn-taking already achieved a higher SAR for RNV conversations than for SCT conversations, the difference was not high enough, and overall, the SAR of these conversations were lower than the empirical conversations. The simulation with the turn-taking mechanism models the distribution of SAR better, with SAR conversations being slightly too interactive and RNV conversations having slightly fewer speaker alternations.

Overall, the improved simulation is able to model the differences in interactivity between SCT and RNV conversations on the level of individual gaps, overlaps, and pauses, as well as on a conversational level. The SAR as the primary indicator of the interactivity of a conversation shows the distinct interactivity levels of SCT and RNV conversations.

4.2 Turn-Taking in Conversations with Delay

In order to predict the impact of transmission delay on the conversational quality, the conversation simulation needs to model the changes in the conversation structure and, thus, the impact on the interactivity of a conversation. Because the turn-taking model implemented in the agents of the simulation is continuous and only depends on the incoming speech signal and the information provided by the side channel, the simulation is able to model interruptions and pauses due to delayed speech.

4.2.1 Impact of Delay on Conversations

One of the most notable changes in the conversation structure due to delay is the decrease in the interactivity of the conversation. Figure 4.10 shows the impact on the interactivity in the form of SAR for SCT and RNV conversations from the CONVSIM dataset.

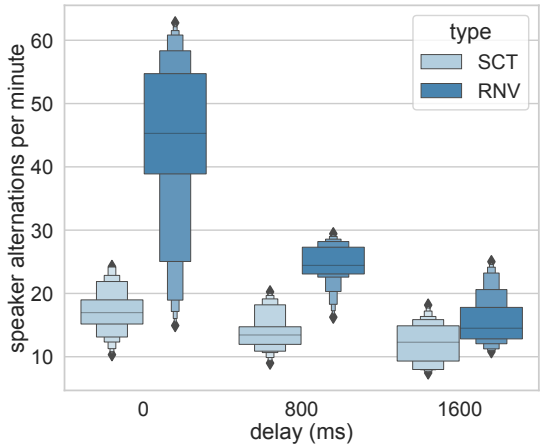

Fig. 4.10 Decrease of the Speaker Alternation Rate for SCT and RNV conversations at 0 ms, 800 ms, and 1600 ms delay for the CONVSIM dataset.

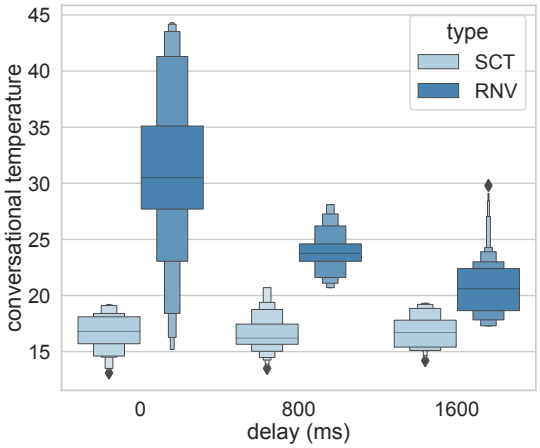

Fig. 4.11 Decrease of the conversational temperature for SCT and RNV conversations at 0 ms, 800 ms, and 1600 ms delay for the CONVSIM dataset.

While the SCT conversations have an overall lower average SAR at no delay (17.51), the SAR drops down to 13.59 at 800 ms and 12.45 at 1600 ms. In comparison, the RNV conversations have a much higher average speaker alternation rate at no delay (40.26) and drop much more severely to 24.97 alternations per minute at 800 ms and 16.09 alternations per minute at 1600 ms, which is even lower than the SAR of the SCT con-

versations at no delay. While the transmission delay impacts both conversation types, the RNV conversations experience a greater reduction in interactivity. The conversational temperature (shown in Figure 4.11) shows the decrease in conversational temperature for both conversation scenarios. The SCT conversations are consistently at below-average temperatures with 16.9° at 0 ms, 16.2° at 800 ms, and 16.5° at 1600 ms, indicating a low overall interactivity for all delay levels. In contrast, the RNV conversations can be considered *heated* with 29.6° at no delay and drop down to 24.4° at 800 ms and finally, *room temperature* with 20.8° at 1600 ms. This metric visualizes the differences in these conversation types, as the temperature of SCT conversations is not affected by the transmission delay, while the temperature of RNV conversations greatly reduces.

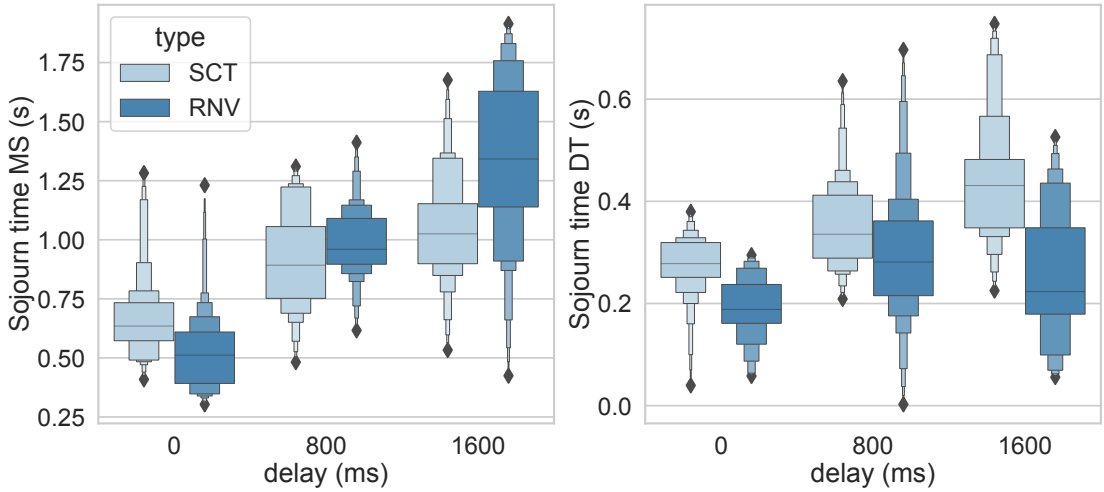

Fig. 4.12 Sojourn times of the states MS (left) and DT (right) for SCT and RNV conversations at 0 ms, 800 ms, and 1600 ms delay for the CONVSIM dataset.

Figure 4.12 shows the development of the sojourn times for the states Mutual Silence (MS) and Double Talk (DT). The average sojourn time MS for SCT conversations starts slightly higher at no delay, but with increasing delay, the sojourn time for RNV conversations is higher. For 800 ms delay the average sojourn time of state MS is 0.99 s for RNV and 0.92 s for SCT conversations, which is higher than the transmission delay itself. At 1600 ms delay the sojourn time increases to 1.35 s for RNV and 1.05 s for SCT conversations, which is significantly lower than the transmission delay. This indicates either that unintended interruptions occur more frequently or that the participants adapt their turn-taking to the higher delay levels. The average sojourn time of state DT shows a steady increase for SCT conversations, while for RNV conversation the sojourn time is highest at 800 ms with 0.3 s and reduces to 0.25 s at 1600 ms delay. Again, this indicates a change in turn-taking behavior, as the increase stagnates at higher delay levels.

Figure 4.13 shows the Intended Interruption Rate (IIR) and Figure 4.14 shows the Unintended Interruption Rate (UIR) for both SCT and RNV conversations at the three delay levels. The number of intended interruptions per minute drops significantly for both SCT and RNV conversations at 800 ms and 1600 ms delay. As the transmission delay increases, the speakers are no longer able to take turns with short overlaps, which make up the majority of intended interruptions. The UIR (Figure 4.14) is more than

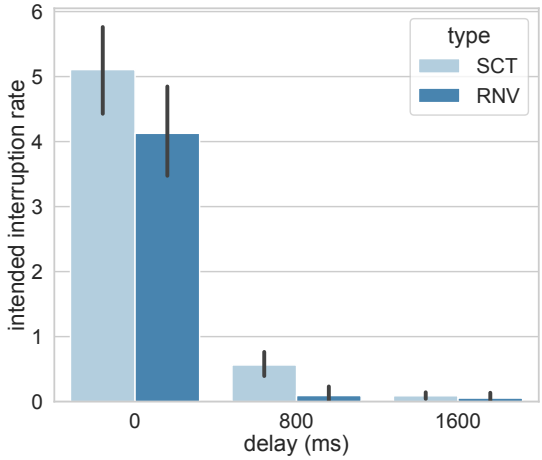

Fig. 4.13 Intended Interruption Rate (IIR) for SCT and RNV conversations at 0 ms, 800 ms, and 1600 ms delay for the CONVSIM dataset. Error bars indicate the 95 % confidence interval.

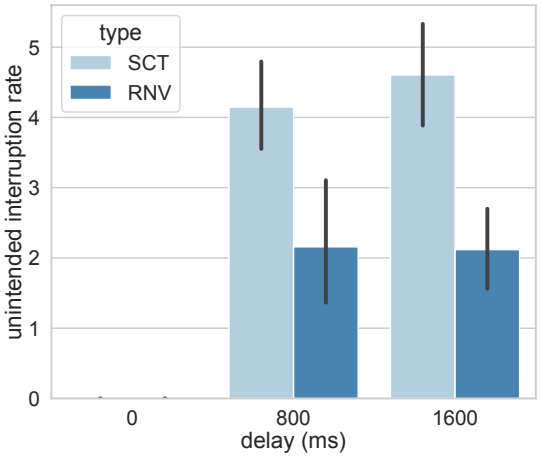

Fig. 4.14 Unintended Interruption Rate (UIR) for SCT and RNV conversations at 0 ms, 800 ms, and 1600 ms delay for the CONVSIM dataset. Error bars indicate the 95 % confidence interval.

twice as high for SCT conversations than for RNV conversations. This is consistent with the higher state probability and sojourn times for the mutual silence state in the RNV conversations at 800 ms delay. At 1600 ms delay, the number of unintended interruptions per minute stays consistent, which indicates that participants adapt their turn-taking to reduce unwanted interruptions.

4.2.2 Performance of the Turn-Taking Model

To evaluate the performance of the previously described turn-taking model when transmission delay is present, for each of the 2 conversation types (SCT and RNV) and each of the 3 delay levels (0 ms, 800 ms, and 1600 ms) 100 conversation were simulated, resulting in 600 simulated conversations. These simulations are compared to the conversations of the CONVSIM dataset.

Figure 4.15 shows the SAR for the simulated and empirical RNV conversations at the three delay levels. The decrease in SAR at 800 ms delay of around 25 alternations per minute is visible for both the simulated and CONVSIM data. The SAR at 1600 ms delay decreases further in the empirical data. However, it increases for the simulated conversations. A similar behavior can be observed in the SAR for the SCT conversation in Figure 4.16. While the overall number of speaker alternations is much lower, the increase in delay leads to a constant decrease in SAR for the CONVSIM data. For the simulations, the decrease again only appears for the 800 ms delay, and an increase in SAR occurs at 1600 ms delay. This saturation in the decrease of the SAR may occur in the simulations because the agents strictly follow the implemented turn-taking mechanism. There, an unintended interruption leads to silence from both interlocutors, after which the turn-taking model again decides which agent speaks in the next turn. This explains the initial decrease in SAR at 800 ms. For the higher delay level of 1600 ms, a substantial

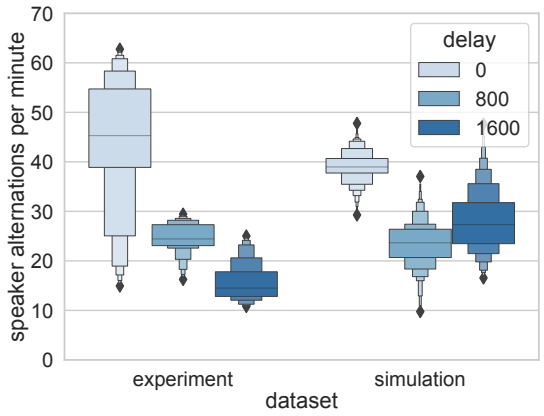

Fig. 4.15 Speaker Alternation Rate (SAR) of the empirical and simulated RNV conversations at 0 *ms*, 800 *ms*, and 1600 *ms* delay.

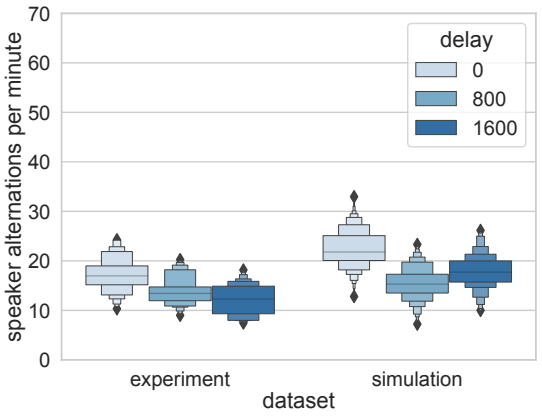

Fig. 4.16 Speaker Alternation Rate (SAR) of the empirical and simulated SCT conversations at 0 *ms*, 800 *ms*, and 1600 *ms* delay.

amount of each turn is transmitted before the agents notice an interruption. This can lead to short turns that are completely uttered before an interruption can occur. Based on the turn-taking model for turn-continuations (Equation 4.3 and 4.4), the agents continue with the next dialogue act while the turn from the interlocutor is still being transmitted. Because the turn-taking model does not adapt to these conditions, the agents proceed with the dialogue, resulting in more speaker alternations than for 800 *ms* delay.

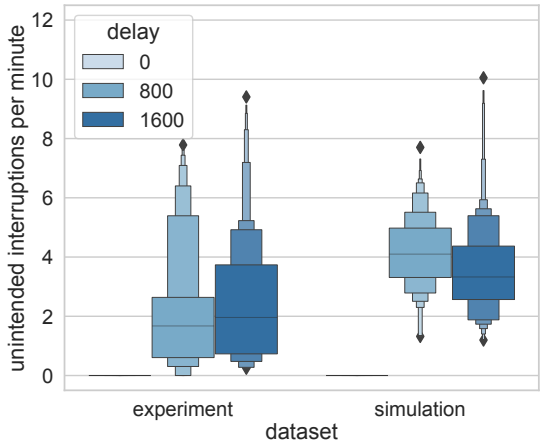

Fig. 4.17 Unintended Interruption Rate (UIR) of the empirical and simulated RNV conversations at 0 *ms*, 800 *ms*, and 1600 *ms* delay.

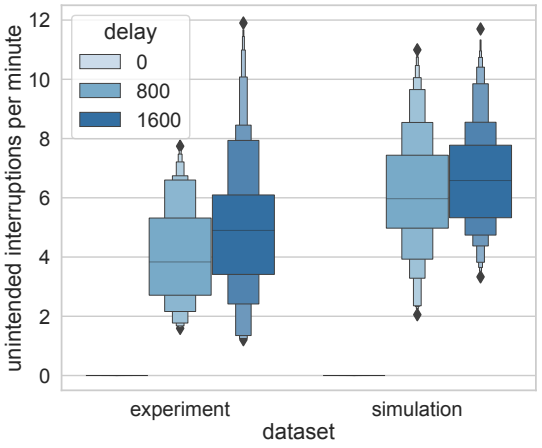

Fig. 4.18 Unintended Interruption Rate (UIR) of the empirical and simulated SCT conversations at 0 *ms*, 800 *ms*, and 1600 *ms* delay.

The Unintended Interruption Rate (UIR) of the RNV conversations (Figure 4.17) and SCT conversations (Figure 4.18) show that the simulation produces more unintended interruptions when compared to the conversations in the CONVSIM dataset. At 0 *ms* delay, the UIR for both the empirical and simulated data is 0, as all interruptions are intended by definition, if no delay is present. While for SCT conversations, the relative increase between 800 *ms* and 1600 *ms* matches the increase of the empirical data, the dampening observed in the SAR is visible in the UIR of the RNV conversations. There,

the rate of unintended interruptions decreases from 4.2 interruptions per minute at 800 *ms* delay to 3.6 interruptions per minute at 1600 *ms*. Again, the steady increase in interruptions as modeled by the simulation does not reflect the turn-taking behavior seen in the empirical data.

4.2.3 Adaptations of Turn-Taking for Delay

In order to adapt the turn-taking mechanism to conversations with transmission delay, the turn-transitions and turn-continuations of the empirical data have to be analyzed on the level of individual turns. For this analysis, the gaps, overlaps, and pauses are extracted from the conversations with 0 *ms*, 800 *ms*, and 1600 *ms* delay. For each turn, the *conversation reality* from the view of the turn-taking (or -keeping) participant was used. With this method, gaps, overlaps, and pauses are extracted as seen from the viewpoint of the participant that takes the next turn, and thus no delay is visible in these measurements. This highlights only the changes in the turn-taking behavior of the participants and no changes due to the pure transmission delay.

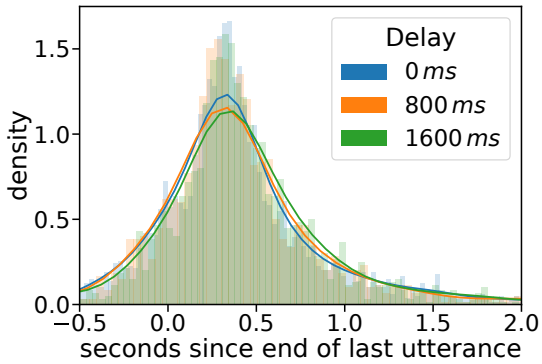

Fig. 4.19 Distribution and kernel density estimations of gaps and overlaps (turn-transitions) for the conversations of the CONVSIM dataset at 0 *ms*, 800 *ms*, and 1600 *ms* delay.

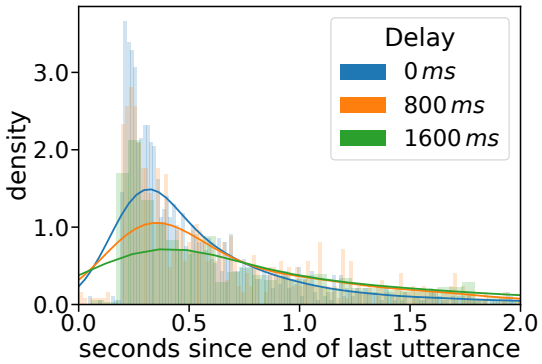

Fig. 4.20 Distribution and kernel density estimations of pauses (turn-continuatios) for the conversations of the CONVSIM dataset at 0 *ms*, 800 *ms*, and 1600 *ms* delay.

Figure 4.19 shows the distribution of gaps and overlaps in the CONVSIM dataset (of both SCT and RNV conversations), separated by the transmission delay. The kernel density lines of the three distributions show very similar behavior between the three delay levels, and a one-way ANOVA shows no significant difference between the distributions. This indicates that the turn-transition decisions of the participants (i.e., overlaps and pauses) do not change with an increase in transmission delay. Figure 4.20 shows the distribution of pauses in the CONVSIM dataset. Here, a difference between the distributions is visible in the kernel density estimations, and the one-way ANOVA is highly significant with $p \ll 0.01$. With increasing transmission delay, the distribution of pauses is flatter, resulting in an increase in the average time of a pause. In comparison, the standard deviation of pauses at no delay is 0.58 seconds, for 800 *ms* it increases to 0.69 and for 1600 *ms* to 1.21 seconds. This increase indicates that participants in conversations with noticeable transmission delays change their turn-taking behavior.

During turn-continuations, participants wait longer to allow for the delayed speech of their interlocutor to arrive. Their turn-transition behavior, however, does not change. Thus, in order to model turn-taking during delayed transmissions, it is sufficient to only model changes in the turn-continuation model.

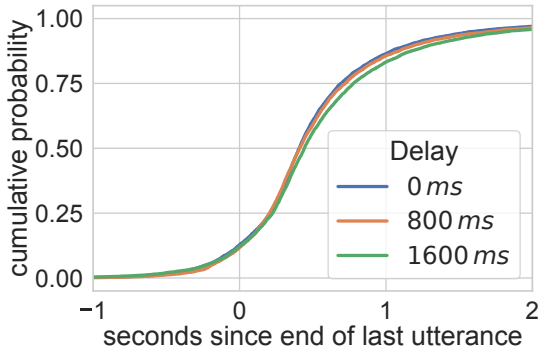

Fig. 4.21 Cumulative probabilities of gaps and overlaps for the empirical conversations at 0 ms, 800 ms, and 1600 ms delay.

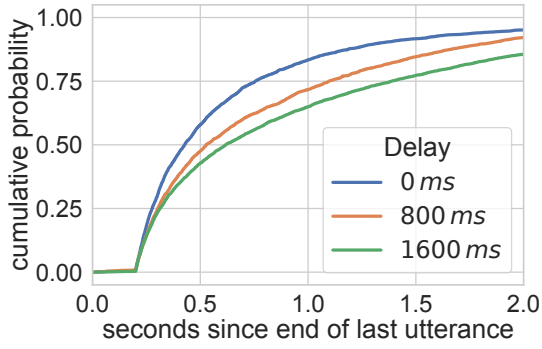

Fig. 4.22 Cumulative probabilities of pauses for the empirical conversations at 0 ms, 800 ms, and 1600 ms delay.

The differences in the distributions of both turn-transitions and turn-continuations are also visible in the cumulative distributions shown in Figure 4.21 and 4.22, respectively. While the cumulative probabilities of gaps and overlaps show no difference between the delay levels, the pauses have a lower cumulative probability for higher delay levels. At 0 ms delay 90 % of all pauses were shorter than 1.33 s, for 800 ms delay this value increased to 1.84 s and for 1600 ms delay, 90 % of all pauses were only shorter than 2.38 s. This increase in pause time can be interpreted as a change in the participant's behavior due to the transmission delay. After each utterance, the participants wait longer before continuing to talk, waiting for the delayed arrival of the interlocutor's turn.

In the empirical conversation, the conversation partners do not have knowledge about the delay level present in the transmission. Thus the adaption of their turn-taking behavior has to be based on a change in the interaction. In order to model this change in turn-continuation behavior without the information on the current delay level, the simulation needs to adapt its turn-taking mechanism based on parameters that are accessible to both agents in the conversation. As both agents can identify unwanted interruptions (i.e., interruptions that occur in the *middle* of an utterance, as defined in Section 3.3.5), the number of occurrences of these interruptions can be used to modify the turn-continuation behavior of the agents. Based on the number of unwanted interruptions, the turn-taking timing for the pauses is *dampened* by a constant factor of 0.2 s, as determined by the median difference of the cumulative distributions divided by the average number of unwanted interruptions in a simulation. As this dampening reduces the number of unwanted interruptions, the turn-taking speed finds an equilibrium. This results in the delay-adapted turn-continuation models for SCT and RNV conversations:

$$\widehat{CD}_{SCT} = 0.9251 \cdot (0.8432 + 2.9231 \cdot x^2) + (C_{UI} \cdot 0.2) \tag{4.5}$$

$$\widehat{CD}_{RNV} = 1.3876 \cdot (0.3607 + 1.2007 \cdot x^2) + (C_{UI} \cdot 0.2) \tag{4.6}$$

where $\widehat{CD}$ is the new turn-continuation model adapted for the overall one-way delay and C_{UI} is the number of unwanted interruptions the agent has experienced. Due to the delayed transmission of the speech signal, this unwanted interruption counter may be different between the agents, resulting in slightly different turn-taking behavior. With higher delay levels, the number of unwanted interruptions increases and, thus, the pause duration of each agent increases as well. This, in turn, decreases the likelihood of further unwanted interruptions occurring, which creates a stabilizing turn-taking adaption.

The final, delay-adapted turn-taking model uses Equations 4.1 and 4.2 for turn-transitioning and Equations 4.5 and 4.6 for turn-continuation, modeling the dampening of pauses in between turns as shown in Figure 4.22.

4.2.4 Evaluation of the Adapted Turn-Taking Model

In order to evaluate the performance of the delay-adapted turn-taking model, 30 SCT and 30 RNV conversations from 0 ms transmission delay up to 2000 ms transmission delay in 100 ms time steps were simulated, resulting in 1260 simulated conversations. These conversations are compared to the conversation with 0 ms, 800 ms, and 1600 ms transmission delay from the CONVSIM dataset. The interactivity and turn-taking are evaluated with P-CA parameters on the conversation level and on the turn-transition and -continuation level with the help of gaps, overlaps, and pauses.

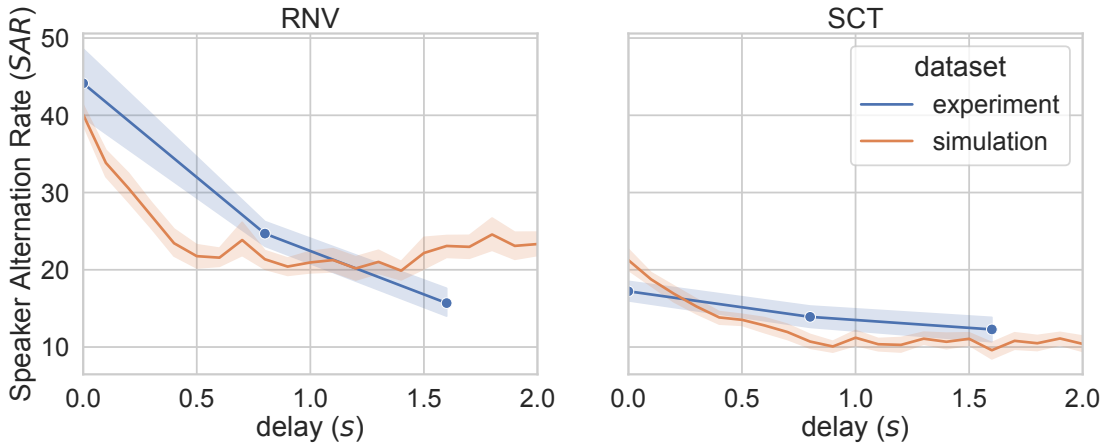

Fig. 4.23 Speaker Alternation Rate of empirical and simulated conversations with delay-adapted turn-taking for SCT and RNV conversations.

Figure 4.23 shows the SAR for both simulated and empirical conversations at various delay levels. The simulated RNV and SCT conversations with no delay match the experimental data, as the turn-taking algorithm is not impacted by the adaption at this stage. With an increase in delay, the simulated conversation's speaker alternation rate drops until it reaches saturation at around 800 ms delay. For RNV conversation, this results in the conversation having a higher SAR at 1600 ms than the empirical data, and for SCT conversation, it is slightly too low at that delay level. Overall, the simulation with the adapted turn-taking mechanism replicates the interactivity behavior of both SCT and RNV conversation of the CONVSIM dataset over the three delay levels.

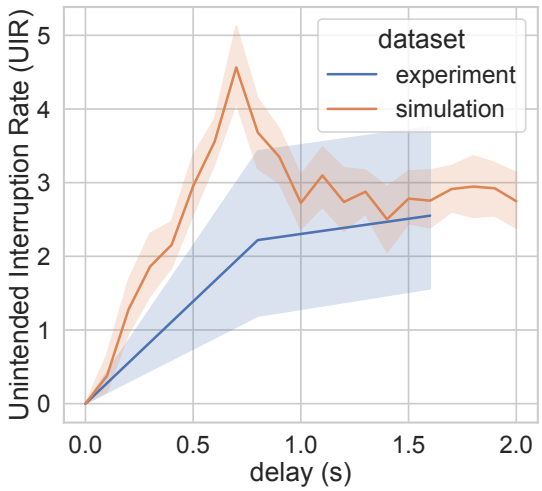

Fig. 4.24 Unintended Interruption Rate of empirical and simulated conversations with delay-adapted turn-taking for RNV conversations.

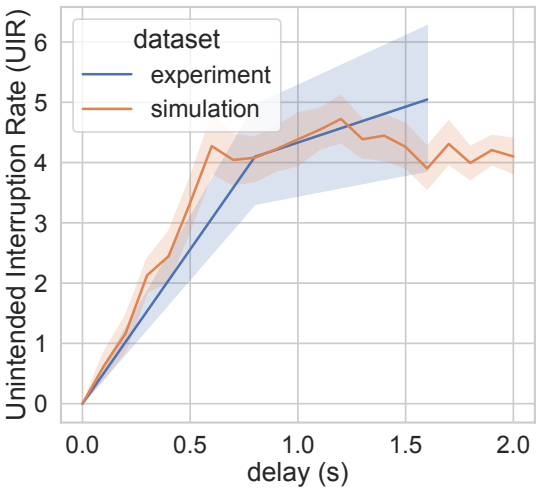

Fig. 4.25 Unintended Interruption Rate of empirical and simulated conversations with delay-adapted turn-taking for SCT conversations.

The Unintended Interruption Rate (UIR) of the simulated and empirical RNV conversation is shown in Figure 4.24. The simulated conversations have a higher UIR up to a peak at 700 ms with an average UIR of 4.5 unintended interruptions per minute. While this value is significantly higher than the UIR of the RNV conversations in the CONVSIM dataset, the dampening reduces the number of unintended interruptions for higher delay levels. At 1600 ms the UIR of the simulated RNV conversations matches the empirical data. The UIR of the simulated and empirical SCT conversations is shown in Figure 4.25. For this conversation type, the UIR of the empirical conversations is generally higher, which is accurately modeled by the simulation. In contrast to the RNV simulations, the overshooting of unintended interruptions at 700 ms is not visible here. The dampening results in a steady UIR of around 4.2 unintended interruptions per minute, which is in agreement with the empirical data. Overall, the variance in the empirical data is very high, which can be explained by the multitude of SCT and RNV scenarios included in the CONVSIM dataset, while the simulations only model one scenario of each conversation type.

Figure 4.26 shows the gaps and overlaps and 4.27 the pauses for the simulated conversations at the three delay levels (RNV and SCT conversations are combined). The simulation is able to reproduce the changes in turn-continuation (i.e., pauses) and the consistency of turn-transitions (i.e., gaps and overlaps) of the empirical data as shown in Figure 4.19 and 4.20. However, the distributions show small differences in gaps and overlaps, resulting in a significant difference between the three delay levels as determined by a one-way ANOVA. However, as measured with the results of SAR and UIR, the turn-taking modeling is able to distinguish between the delay levels sufficiently enough to model these conversation-level phenomena.

Overall the simulation is able to model the differences in turn-taking during delayed speech transmission without the need for a delay-based parameter of the turn-taking models. This results in an overall acceptable reproduction of the interactivity of the conversation types, as well as the changes in interactivity due to delay.

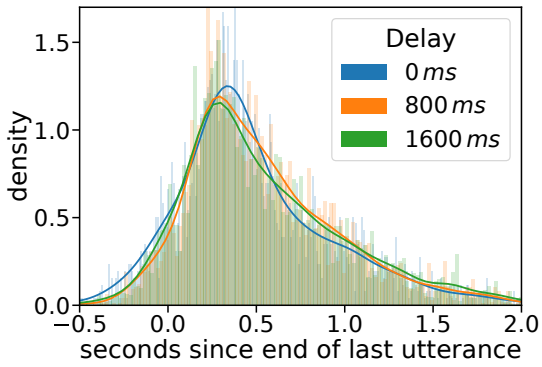

Fig. 4.26 Distribution and kernel density estimations of gaps and overlaps (turn-transitions) for the simulated conversations at 0 ms, 800 ms, and 1600 ms delay.

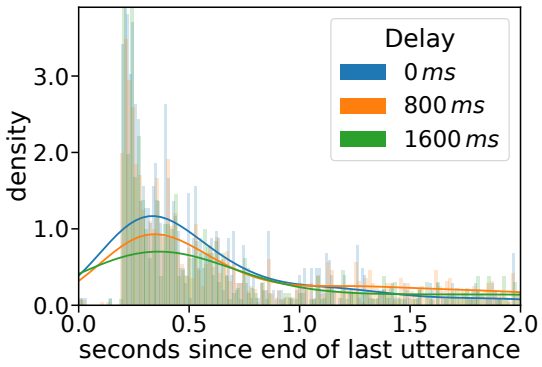

Fig. 4.27 Distribution and kernel density estimations of pauses (turn-continuations) for the simulated conversations at 0 ms, 800 ms, and 1600 ms delay.

4.3 Summary

In this chapter, the simulation was extended with a turn-taking mechanism that is able to reproduce the differences in the interactivity of SCT and RNV conversations, as well as the changes in interactivity due to transmission delay. In order to implement a model that enables the simulation to perform turn-taking, the differences in the interactivity of SCT and RNV conversations were analyzed on the conversation level. Then, a turn-taking model was proposed on the level of individual turns by modeling turn-continuations based on the pauses in turn-keeping and turn-transition based on the gaps and overlaps between speaker changes. The turn-taking mechanism of the simulation then uses the turn-transition and turn-continuation depending on the type of dialogue act that is currently uttered. An evaluation of this approach showed that the turn-taking model is able to replicate the differences in interactivity between the two simulated conversation types.

This turn-taking model was then used in a conversation with added transmission delay. An analysis of the resulting simulated conversations showed that the turn-taking model is able to replicate the changes in interactivity for low levels of delay. However, for higher delay levels, the interactivity of the simulated conversations starts to deviate from the empirical conversations. Thus, the gaps, overlaps, and pauses of conversations at different delay levels were analyzed and revealed that there is a significant change in turn-taking behavior during higher delay levels. Specifically, the duration of pauses in turn-continuations (i.e., pauses in between turns of the same speaker) increased with higher levels of transmission delay. This behavioral change was implemented into the turn-taking mechanism by dampening turn-continuations based on the number of unwanted interruptions an agent encounters in a conversation.

A final evaluation showed that the improved and delay-adapted turn-taking model of the simulation is able to both replicate the interactivity levels of SCT and RNV conversations, as well as the degradation of the interactivity due to delay.

Chapter 5
Simulating Conversation Disruptions and Packet Loss

The effects of packet loss on a conversation and its perceived quality can be widely different, depending on which and how many parts of the conversation are affected. For small numbers of lost packets that come in short bursts, the degradation mainly affects the audible perception of the transmitted speech. While PLC methods of modern codecs try to hide the loss of the speech by interpolating the signal, artifacts such as a robotic voice or even small section of silence can be perceived by the user. The effects of this type of short packet loss on the conversational quality have long been studied and can be modeled by the parameter of the transmission (e.g., the packet loss probability and burst ratio). However, with increasing packet loss probability or, more commonly, rare but long bursts of consecutively lost packets, the amount of affected speech starts to impact the understandability and, thus, the flow of the conversation. Depending on where the packet loss burst occurs, the transmission of information can be corrupted, resulting in the need to retransmit the information with repairing dialogue. This changes both content and structure of the conversation, which cannot be modeled by the parameters of the transmission alone. That is why the conversation simulation needs to model the impact of highly bursty packet loss on the conversation.

This chapter focuses on analyzing the structure of conversations affected by bursty packet loss and modeling the changes in the interaction. First, the interactivity of conversations with packet loss is investigated, and the concept of *conversation disruptions* is defined, which formalizes misunderstandings that occur due to lost packets. With the newly defined concept of conversation disruptions, the conversations of the CONVSIM dataset are analyzed, and differences in the interactivity are shown. Then, based on a turn-level analysis, conversation disruptions are modeled for the use in the simulation. Finally, the disruptions and the changes in interactivity are simulated and the resulting conversations compared to the empirical data.

The analysis of the impact of bursty packet loss on the conversational structure has been published in part in Michael (2021) and Michael and Ibrahim (2022).

5.1 Interactivity in Conversations with Packet Loss

In order to analyze the impact of bursty packet loss on the interactivity of a conversation, the conversations of the CONVSIM dataset with 0 %, 15 %, and 30 % packet loss and a burst ratio of 4.0 were used to perform a P-CA.

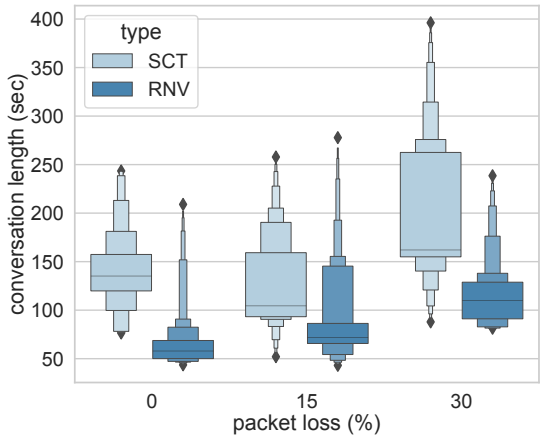

Fig. 5.1 Length of SCT and RNV conversations in the CONVSIM dataset at 0 %, 15 %, and 30 % bursty packet loss.

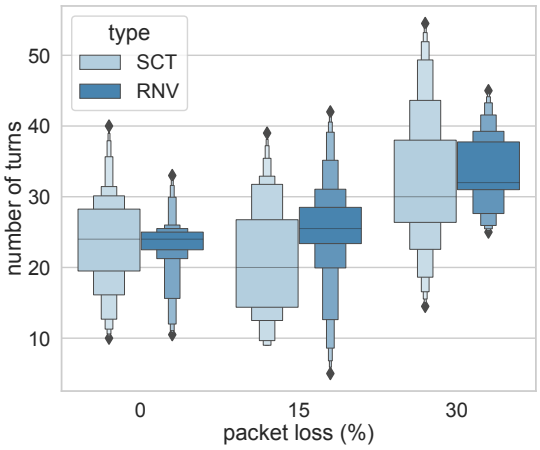

Fig. 5.2 Number of turns in SCT and RNV conversations of the CONVSIM dataset at 0 %, 15 %, and 30 % bursty packet loss.

Figure 5.1 shows the median and distribution of the conversation lengths of SCT and RNV conversations at 0 %, 15 %, and 30 % packet loss. While RNV conversations show an increase in length with higher packet loss probabilities, the SCT conversations become shorter at 15 % and then increase in length again at 30 % packet loss. Generally, RNV conversations are shorter than SCT conversations, even at high levels of packet loss. The increase in conversation length can also be seen in the number of turns (Figure 5.2). Both SCT and RNV conversations have a similar number of turns at 0 % packet loss. Again, at 15 %, the number of turns goes down for SCT conversations, while for RNV conversations, it increases. While the interactivity parameters alone are not able to explain this reduction in conversation length and turn count for the SCT conversations at 15 % packet loss, the transcriptions of the conversations indicate a reduction of small-talk. Participants tend to leave out more open exchanges when the conversation is affected by packet loss. At 15 % packet loss, this leads to a decrease in the conversation's length. At 30 % packet loss, the repairing dialogue due to the severe packet loss leads to a further increase in length and number of turns. Because the RNV task does not leave room for open conversations, this phenomenon can only be observed for SCT conversations.

The increase in conversation length and turn count for the higher packet loss levels indicates that misunderstanding speech causes repairing dialogue, which needs additional turns and thus prolongs the conversation. Table 5.1 shows the median values for length and turn count and also the increase relative to the 0 % packet loss condition. While the relative changes of the turn count seem to match the relative change in the conversation length, the median length of RNV conversations is almost twice as long (90 %) at the 30 % packet loss level. In contrast, the increase in turn count is only 23.08 %. Because the length of each turn is similar for each packet loss condition, the disproportionate increase in conversation length cannot be attributed solely to the increase in the length of the utterances.

Figure 5.3 shows that the increase in length stems in part from the fact that the RNV conversations are slowed down by the increasing packet loss, while the SCT conversation stays at the same level of SAR. The state probability of Mutual Silence in

Table 5.1 Median length and turn count and turn length for the SCT and RNV conversation at 0, 15 and 30 % packet loss and increase of length and turn count relative to the 0 % packet loss condition.

Scenario	Packet Loss	Median length (s)	increase in %	Median turn count	increase in %	Median turn length (s)
SCT	0 %	135.15	-	24	-	1.25
	15 %	104.50	−22.68	20	−16.67	1.17
	30 %	162.05	19.90	30	25.00	1.29
RNV	0 %	57.90	-	25	-	0.63
	15 %	71.82	24.04	26	4.00	0.61
	30 %	109.95	89.90	32	23.08	0.75

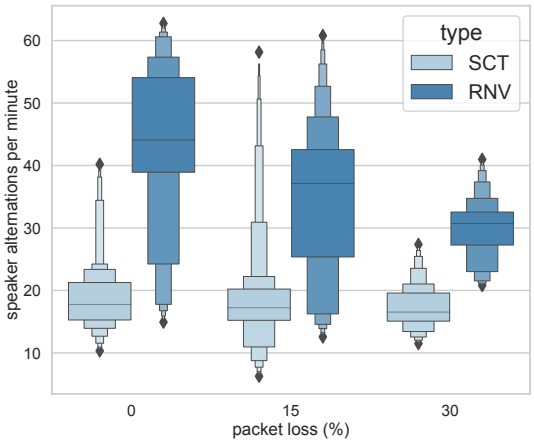

Fig. 5.3 SAR of SCT and RNV conversations in the CONVSIM dataset at 0 %, 15 %, and 30 % bursty packet loss.

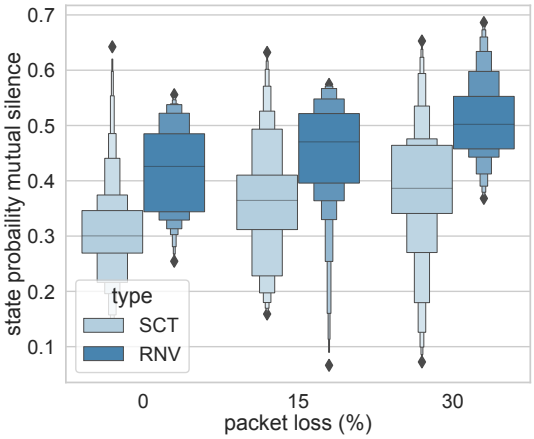

Fig. 5.4 State probability MS of SCT and RNV conversations in the CONVSIM dataset at 0 %, 15 %, and 30 % bursty packet loss.

Figure 5.4 shows that the lower speaker alternation rate is mainly caused by an increase in silence. This indicates that the high interactivity scenario (i.e., the rapid exchange of numbers) has an influence on how much packet loss impacts the conversation. One reason for the difference in the two scenarios might be the density of information in each utterance. While the sentences in SCT conversations tend to be longer, with relatively few words important for the understanding of the conversation, the utterances in RNV conversations mostly consist of only the information that needs to be transmitted in order to advance the conversation.

5.2 Disruptions in Conversations with Packet Loss

Longer bursts of packet loss result in the omission of information that might be important for continuing the conversation. Depending on the importance of the utterance that has been affected, the listener might need to ask for a retransmission of information. A participant in a conversation with packet loss intuitively considers the impact on the understandability and either continues with the dialogue with incomplete or interpolated

knowledge about the previous utterance or actively asks for retransmission of the information with repairing dialogue like "*I didn't understand that.*" or "*Could you repeat that?*". This decision is based on the context of the conversation, as well as the listener's own ability to reconstruct the meaning of the packet loss-affected utterance.

As a third party, however, the reason behind such a disruption of the conversation cannot be reconstructed from the dialogue alone. Assessing the decision process behind each packet loss burst is unpractical, as it would itself disrupt the conversation. That is why understandability in a conversation cannot be assessed without the information about the context available in that specific scenario. Thus, in the analysis for the conversation simulation, only *conversation disruptions* are considered.

The term *conversation disruption* is defined here as every turn in which a conversation partner has to ask for the retransmission of information. These information requests might be of general nature (e.g., "*Could you please repeat that?*") or might refer to a specific concept (e.g., "*Which pizza was that?*"), but always relates to the turn directly preceding it. A conversation disruption is not necessarily rooted in a misunderstanding due to packet loss but may also occur due to other circumstances. Also, not every part of an utterance that is misunderstood has to cause a conversation disruption. When the meaning of a turn can be extracted from the remaining speech, the interlocutors often continue with the dialogue without acknowledging the packet loss.

With the help of the transcriptions, each conversation disruption was annotated, and for each turn, the percentage of lost speech was calculated. The packet loss information, together with the transcriptions, were used to determine which words were affected by packet loss. For the analysis, a word with at least 50 % of lost phonemes was considered "lost" (i.e., not understandable). Figure 5.5 shows an exemplary part of a conversation with all available annotations.

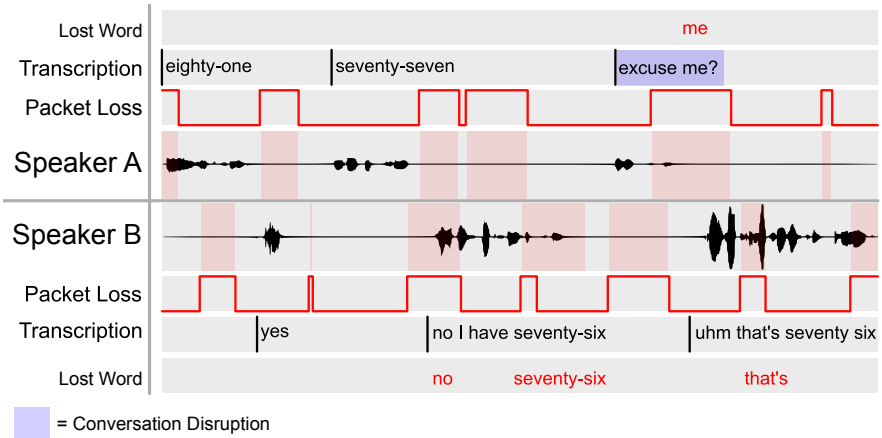

Fig. 5.5 Exemplary overview of an annotated conversation. The speech is recorded in separate channels, the packet loss pattern is shown in red, the conversation is transcribed and force aligned. Conversation Disruptions are marked in blue and words where more than 50 % of phonemes were affected by packet loss are annotated in red.

Figure 5.6 shows the total number of conversation disruptions for SCT and RNV conversations at the three packet loss levels. While there are almost no disruptions in conversations without packet loss for both SCT and RNV conversations, the average

number of disruptions increases to 1.5 at 15 % packet loss and to 5.28 at 30 % packet loss. This increase shows that, as expected, the higher packet loss probabilities incite more conversation disruptions. However, even though the average length and number of turns of SCT and RNV conversations are different, the number of conversation disruptions shows no significant difference.

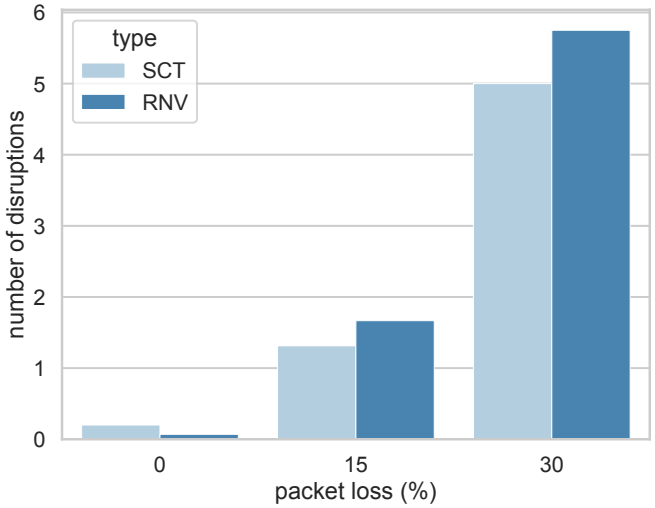

Fig. 5.6 Number of conversation disruptions for SCT and RNV conversations of the CONVSIM dataset with 0 %, 15 %, and 30 % bursty packet loss.

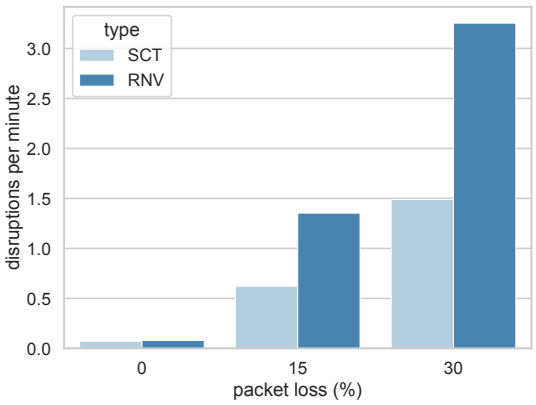

Fig. 5.7 Conversation disruption rate for conversations of the CONVSIM dataset with 0 %, 15 %, and 30 % bursty packet loss.

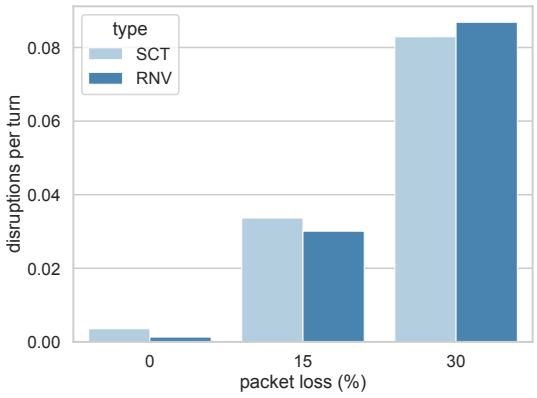

Fig. 5.8 Average disruptions per turn for conversations of the CONVSIM dataset with 0 %, 15 %, and 30 % bursty packet loss.

The Conversation Disruption Rate (CDR), defined as the number of conversation disruptions per minute, is shown in Figure 5.7. Here, a clear distinction between SCT and RNV conversations at the 15 % and 30 % packet loss level is visible. RNV conversations have more than twice the CDR than SCT conversations, which can be explained by the much shorter turns. The number of conversation disruptions per turn (shown in Figure 5.8) shows no significant difference between the conversation types, which is to

be expected, as the number of conversation disruptions as well as the number of turns in both SCT and RNV scenarios are similar. While the experience of the participants might focus on the number of disruptions relative to the amount of information transmitted (i.e., number of turns), the conversation structure and interactivity are dependent on the CDR.

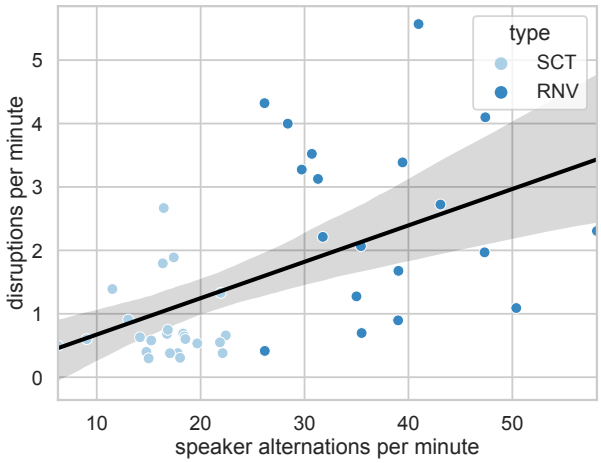

Fig. 5.9 Correlation between the Conversation Disruption Rate and the Speaker Alternation Rate for conversations of the CONVSIM dataset with 15 % and 30 % bursty packet loss, split by SCT and RNV conversations.

Figure 5.9 shows the moderate positive linear correlation of the SAR and the CDR with $r = 0.53$ and $p < 0.01$. For this correlation only conversations with 15 % and 30 % packet loss were considered. This indicates that a higher speaker alternation rate generally leads to more conversation disruptions when bursty packet loss is present.

5.3 Simulating Conversations with Bursty Packet Loss

The analysis of conversation with bursty packet loss has shown that this degradation affects a conversation in two ways. First, the bursts of lost speech lead to misunderstandings that require repairing dialogue for the conversation to continue. These conversation disruptions increase the length of the conversation and add additional repairing turns. Secondly, the impaired transmission of speech leads to a change in the turn-transition behavior of the interlocutors. The transition between speakers is performed with more silence in between the utterances (i.e., gaps). This behavioral change leads to more and longer stretches of mutual silence. The analysis has also shown that the difference in conversational interactivity cannot be explained by only one of these effects.

Thus, to simulate the effects of bursty packet loss on the conversation, both conversation disruptions and the change in turn-taking behavior have to be modeled. For this, the conversation disruptions need to be analyzed and reproduced on a turn level and the turn-continuation mechanism of the simulation has to be adapted.

5.3.1 Modeling Conversation Disruptions in a Simulation

While conversation disruptions affect the overall structure and interactivity of the conversation, they are rooted in a misunderstanding of parts of the utterance due to packet loss. Thus, to model these conversation disruptions and the utterances that caused them, the phenomenon has to be analyzed on the turn-level. That is why the relationship between the amount of lost speech (due to packet loss) in each utterance and the following occurrence of a conversation disruption is investigated.

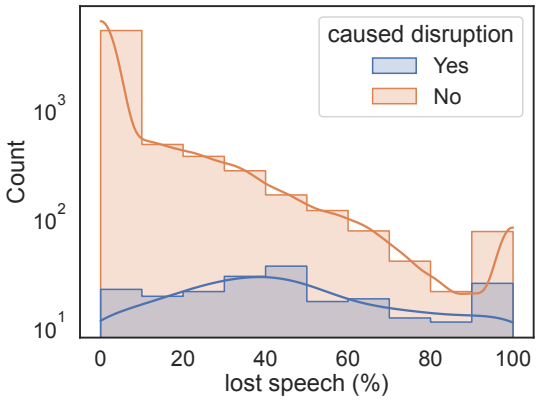

Fig. 5.10 Logarithmic histogram of lost speech for each utterance in the CONVSIM dataset, divided by whether they resulted in a conversation disruption or not.

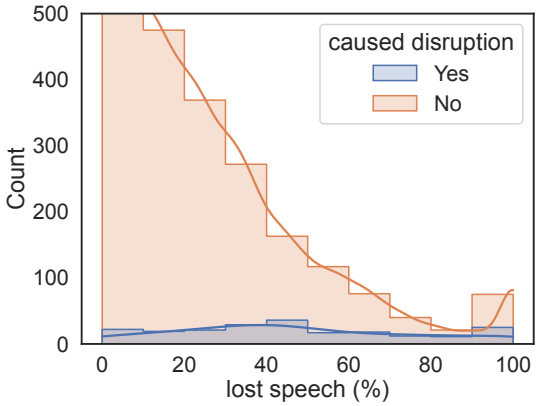

Fig. 5.11 Zoomed linear histogram of lost speech for each utterance in the CONVSIM dataset, divided by whether they resulted in a conversation disruption or not.

Figure 5.10 shows the logarithmic distribution and kernel density estimation of the utterances based on what percentage of speech was lost and separated by whether these utterances resulted in a conversation disruption in the next turn (i.e., repairing dialogue) or not. This analysis only considers the amount of lost speech in each utterance and is done independently of the packet loss probability of the transmission. Thus, utterances of conversations with 0 % 15 % and 30 % packet loss were combined in the analysis. As utterances without any lost speech make up the vast majority of data points, Figure 5.11 shows the same distribution in linear scale but zoomed in on the y-axis. Here the relative scale between the two categories of utterances is visualized. While the distribution decreases for both categories at over 50 % of speech lost, the relative amount of conversation disruption causing utterances increases for lost speech between 0 and 90 %. For utterances with over 90 % of speech lost, the relative percentage of utterances that caused a disruption decreases again. From the transcriptions and recorded audio of the conversations, it can be hypothesized that the amount of lost speech results in the listener not noticing that the interlocutor was speaking at all. Thus, relatively fewer conversation disruptions are observed after utterances with such high percentage of lost speech.

In order to model the occurrences of conversation disruptions based on the percentage of lost speech in the preceding utterance, the relative amount of understood utterances have to be estimated from the two distributions in Figure 5.11. For this, the samples of utterances that produced a conversation disruption and those that did not were sampled

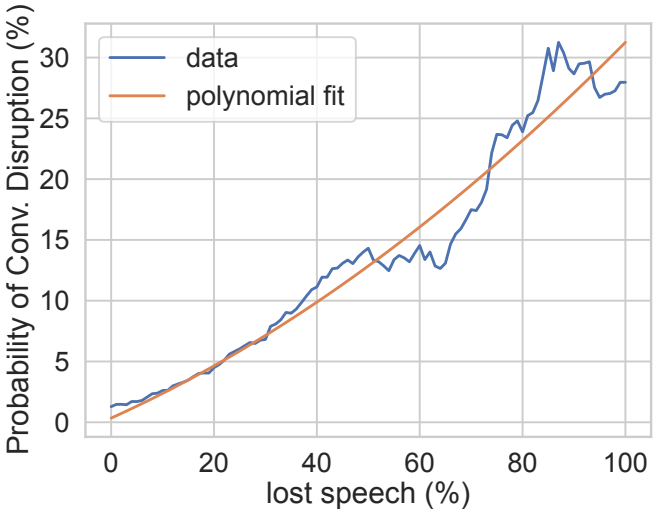

Fig. 5.12 Average of utterances that lead to a conversation disruption at levels of percent lost speech, as well as a polynomial fit for the resulting curve.

with a delta of 0.1 at every percentage of lost speech to determine the percentage of utterances that caused a conversation disruption at that level. This percentage is shown in Figure 5.12, where the relative occurrences of conversation disruption increase with higher amounts of lost speech. The resulting conversation disruption probability is then modeled based on the percentage of lost speech as a polynomial function with a least-squares fit. The resulting function has a Root-Mean-Square Error (RMSE) of 1.9 % and is used together to predict the probability of a disruption in the following turn:

$$\widehat{P_{CD}} = 0.1394 \cdot P_{LS}^2 + 0.1652 \cdot P_{LS} + 0.0035 \tag{5.1}$$

where P_{LS} is the ratio of lost speech in the utterance and $\widehat{P_{CD}}$ is the estimated probability of a conversation disruption in the following turn.

To integrate this model into the simulation, the information about the lost speech of each utterance has to be available to the agents. Thus, the simulated telephone network introduces the status of the packet loss into the side channel. Then, the turn-taking dialogue manager of each agent calculates from the relative amount of lost packets of each utterance the percentage of lost speech. With Equation 5.1 and a random number generator, it is decided whether the agent *misunderstands* the previous utterance. If this is the case, the dialogue manager inserts a `misunderstanding` dialogue act with the appropriate concepts on top of the stack. As with all other dialogue, the dialogue act guiding system distributes misunderstandings based on the occurrences in the training data. As the SMISS dataset does not contain bursty packet loss and thus, only a few conversation disruptions, 10 SCT and RNV conversations with 15 % packet loss and 10 SCT and RNV conversations with 30 % packet loss of the CONVSIM dataset are transcribed, annotated and added to the training set of the simulation.

5.3.2 Modeling Turn-Taking in a Simulation with Packet Loss

While the conversation disruptions change the structure of the conversation, they alone cannot explain the reduction in interactivity in conversations with bursty packet loss. To quantify and later model the changes in turn-taking on the level of turn-transitions and turn-continuations, the gaps, overlaps, and pauses of the conversations of the CONVSIM dataset at 0 %, 15 %, and 30 % bursty packet loss are extracted and analyzed.

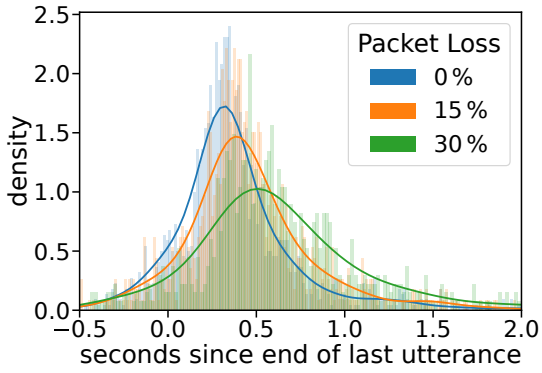

Fig. 5.13 Distribution and kernel density estimations of gaps and overlaps (turn-transitions) for the conversations of the CONVSIM dataset at 0 %, 15 %, and 30 % bursty packet loss.

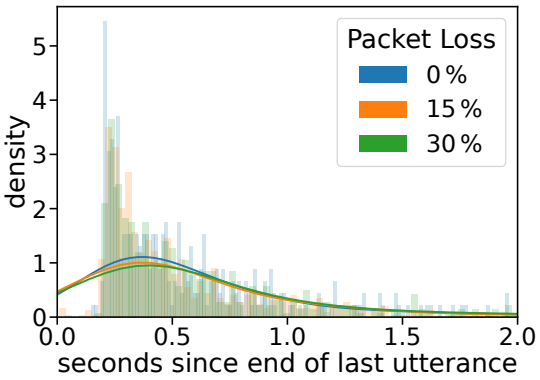

Fig. 5.14 Distribution and kernel density estimations of pauses (turn-continuations) for the conversations of the CONVSIM dataset at 0 %, 15 %, and 30 % bursty packet loss.

Figure 5.13 shows the timing of the turn-transitions in the form of gaps and overlaps in the conversations of the CONVSIM dataset at the three packet loss levels (RNV and SCT conversations are combined). A one-way ANOVA shows a significant difference between the turn-transition distributions. This indicates that the conversation partner waits longer when taking over the turn from their interlocutor. Thus interruptions become more sparse, and there is more silence between turns. The distribution of turn-continuations as shown in Figure 5.14 shows no discernible difference between the three packet loss levels, and the one-way ANOVA is not significant.

The cumulative probabilities of these distributions show the differences in turn-transition and turn-continuation probabilities. Figure 5.15 shows the cumulative distribution of the turn-transitions, which shows the significant difference in the probability relative to the seconds since the end of the previous utterance. Figure 5.16 shows the cumulative probability for turn-continuations, which shows no significant difference. One explanation for this behavior might be that participants are under more cognitive load when listening to the highly degraded speech of their interlocutor and thus, take more time taking the turn. For turn-continuations, this increase in load is not present, and turn-continuation times stay consistent between the levels of packet loss.

In order to model these changes in turn-transitioning behavior, the turn-transition behavior of the simulated agents has to be adapted. In degraded conversations, the participants do not know the exact amount of packet loss that is present in the conversation but only hear the amount of lost packet in the current turn of their interlocutor. Thus, modeling the changes in turn-transitions based on the packet loss probability alone

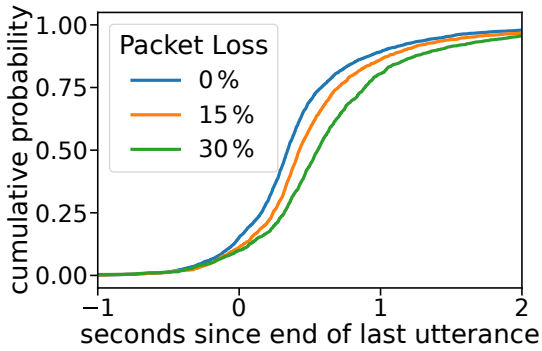

Fig. 5.15 Cumulative probability of gaps and overlaps (turn-transitions) for the conversations of the CONVSIM dataset at 0 %, 15 %, and 30 % bursty packet loss.

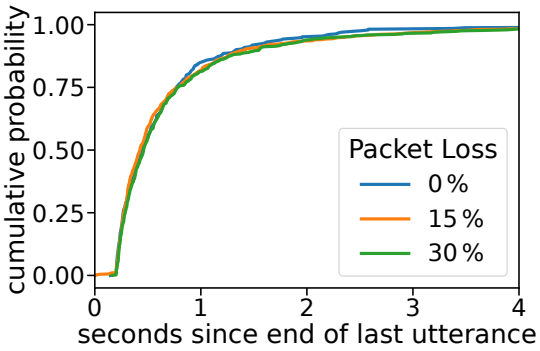

Fig. 5.16 Cumulative probability of pauses (turn-continuations) for the conversations of the CONVSIM dataset at 0 %, 15 %, and 30 % bursty packet loss.

would not be sufficient. In order to achieve a *dampening* effect of turn-transitions that is consistent throughout the conversation, the changes in the model are based on the number of conversation disruptions that have occurred (and thus implicitly on the amount of lost speech in incoming turns). Based on the number of conversation disruptions, the turn-transitioning is dampened by a constant factor of 0.055 seconds, determined by the median difference of cumulative distributions divided by the average number of conversation disruptions. This results in the packet-loss-adapted turn-transition model for SCT and RNV conversations:

$$\widehat{TPL_{SCT}} = -0.3226 \cdot \log(0.433 \cdot (-1 + \frac{1}{x}) + (C_{CD} \cdot 0.055) \tag{5.2}$$

$$\widehat{TPL_{RNV}} = -0.1598 \cdot \log(0.17 \cdot (-1 + \frac{1}{x}) + (C_{CD} \cdot 0.055) \tag{5.3}$$

where $\widehat{TPL}$ is the new turn-transition model adapted for packet loss and C_{CD} is the number of conversation disruptions the agents have experienced. Each simulated agent counts only the numbers of their own conversation disruptions, and thus, these counts might be different between the agents. With an increase in bursty packet loss, conversation disruptions also increase, resulting in slower turn-transitioning time.

The resulting delay- and packet-loss-adapted turn-taking model uses Equations 5.2 and 5.3 for the determination of turn-transitions and the Equations 4.5 and 4.6 for the determination of turn-continuations for the dialogue acts dominant in the SCT and RNV conversation, respectively.

5.4 Evaluation of Simulations with Disruptions and Packet Loss

In order to evaluate the performance of the two modifications of the simulation regarding bursty packet loss, 30 SCT and 30 RNV conversations from 0 % packet-loss up to 30 % packet-loss in 5 percent point increases and a burst ratio of 4.0 were simulated, resulting in 420 simulated conversations. They are compared to the conversations with 0 %,

15 %, and 30 % packet loss with a burst ratio of 4.0 of the CONVSIM dataset. First, the occurrences of conversation disruptions in the simulation are compared to the empirical data. Then, the changes in interactivity parameters of the conversation are analyzed by performing a P-CA on both the simulated and empirical conversations. Finally, the packet-loss-adapted turn-taking model is evaluated by comparing the gaps, overlaps, and pauses of the simulations with the distributions of turn-taking in the CONVSIM dataset.

Figure 5.17 shows the number of conversation disruptions per minute for simulated and empirical RNV conversations and Figure 5.18 for SCT conversations. For both simulated and empirical conversations, the number of disruptions per minute increases with higher packet loss probabilities. However, while the CDR increases to 3.25 at 30 % packet loss for the empirical RNV conversations, the empirical SCT conversations only has 1.49 conversation disruptions per minute at the same packet loss probability. This difference is replicated by the simulation for both conversation types, with SCT conversations having a slightly too high CDR at 30 % packet loss, and RNV conversations having a slightly too high CDR at 15 % packet loss. In order to assess the conversation disruptions not only on a time basis but also on a turn basis, the disruptions per turn have to be analyzed.

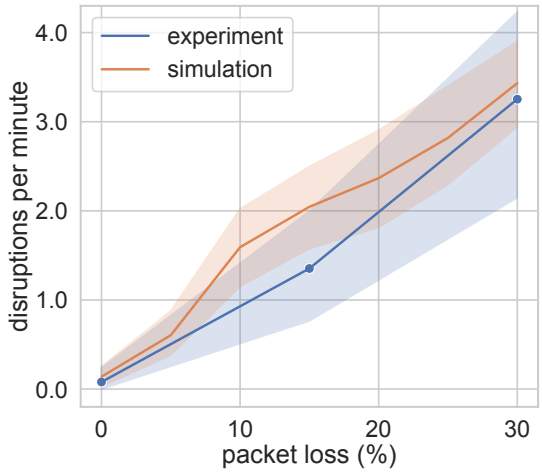

Fig. 5.17 Conversation Disruption Rate in RNV conversations of the CONVSIM dataset (experiment) and simulations at various packet loss probabilities with a burst ratio of 4.0.

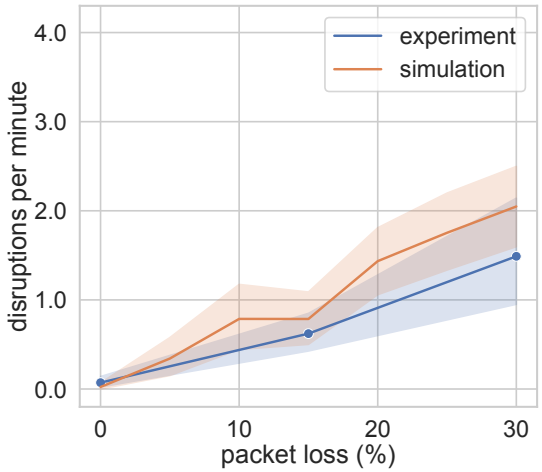

Fig. 5.18 CDR in SCT conversations of the CONVSIM dataset (experiment) and simulations at various packet loss probabilities with a burst ratio of 4.0.

Figure 5.19 shows the conversation disruptions per turn for empirical and simulated RNV conversations and Figure 5.20 for SCT conversations. Here, the empirical SCT and RNV conversations show no significant differences. Again, the simulation is able to replicate the disruptions per turn for both SCT and RNV conversations sufficiently while being slightly too high for both conversation types at both 15 % and 30 % packet loss. The different behavior of the conversation types in CDR and the absence of a difference in disruptions per turn are modeled by the simulated conversational data and confirm the conversation disruption model implemented to be suitable to replicate this phenomenon.

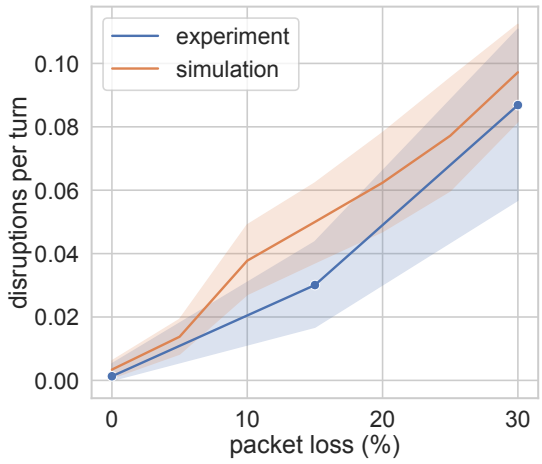

Fig. 5.19 Disruptions per turn in RNV conversations of the CONVSIM dataset (experiment) and simulations at various packet loss probabilities with a burst ratio of 4.0.

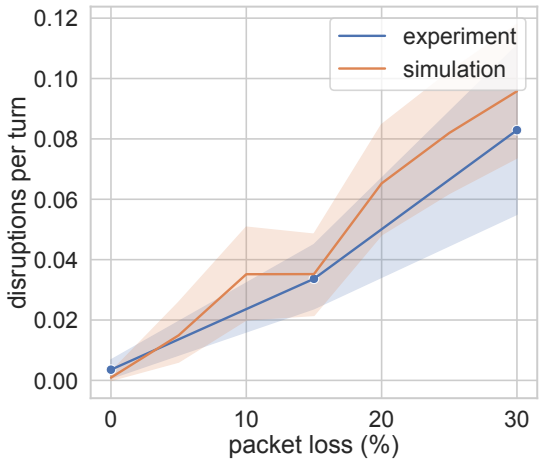

Fig. 5.20 Disruptions per turn in SCT conversations of the CONVSIM dataset (experiment) and simulations at various packet loss probabilities with a burst ratio of 4.0.

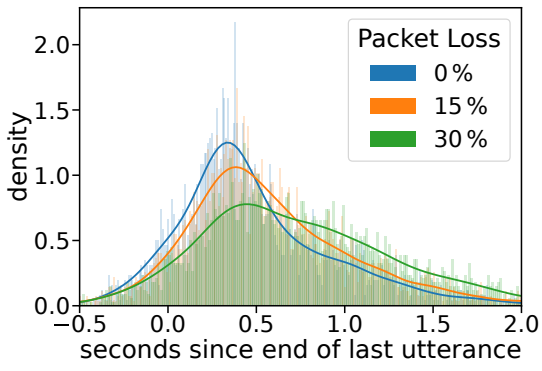

Fig. 5.21 Distribution and kernel density estimations of gaps and overlaps (turn-transitions) for the simulated conversations at 0 %, 15 %, and 30 % bursty packet loss.

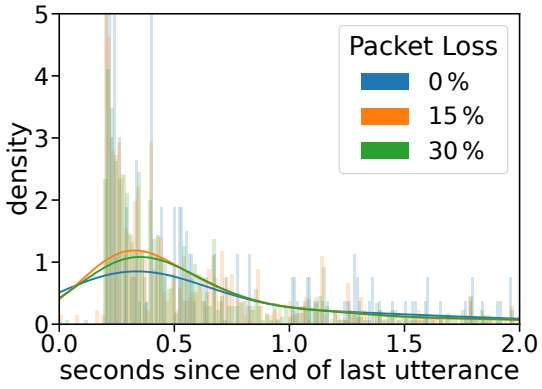

Fig. 5.22 Distribution and kernel density estimations of pauses (turn-continuations) for the simulated conversations at 0 %, 15 %, and 30 % bursty packet loss.

Figure 5.21 shows the distribution of turn-transitions in the form of gaps and overlaps of the simulated conversations. The modeled difference in behavior due to the bursty packet loss results in the slower turn-transitioning at higher packet loss levels. However, the distribution of gaps over 0.6 seconds shows a slight bias toward the longer turn-transition times (compared to Figure 5.13). The distribution of turn-continuations in Figure 5.22 shows a similar distribution as the empirical conversation. Differences in the distributions of the packet loss levels can be seen, even though the turn-continuation algorithms were not adapted, and the empirical data shows no significant effect of packet loss on the pauses. Especially the distribution of pauses at 0 % packet loss is flatter than the distributions at 15 % and 30 % packet loss and the empirical turn-continuation distributions. However, the effect of packet loss on the simulated turn-continuation is small.

With the changes in the turn-taking mechanism and the simulation of conversation disruptions evaluated separately, the overall impact of packet loss on the interactivity of the simulated conversations is evaluated. For this, the changes in the number of speaker alternations per minute are evaluated at each level of bursty packet loss and compared to the empirical data.

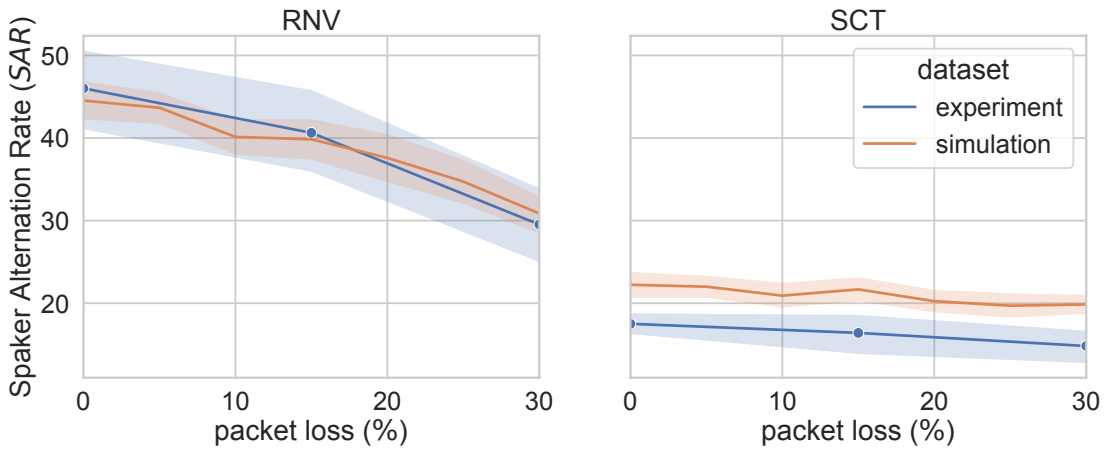

Fig. 5.23 Speaker Alternation Rate for simulated and empirical RNV (left) and SCT (right) conversations at various packet loss probabilities with a burst ratio of 4.0.

Figure 5.23 shows the SAR in speaker alternations per minute for the RNV and SCT conversations. For the RNV conversations, the simulation matches the SAR of the empirical data and is able to replicate the drop in alternations per minute at higher packet loss rates. For SCT conversations, the simulation is able to model the consistency of SAR over the different packet loss levels. However, overall the SAR is modeled slightly too high, with the average SAR of the simulations being 20.5 while the average SAR of the empirical SCT conversations is 18.

Overall the packet-loss-adapted simulation is able to sufficiently replicate the occurrences of conversation disruptions and the changes in turn-continuations at high packet loss levels. While especially the modeling of conversation disruptions relies on information transmitted by the simulated telephone network through a side channel, the agents themselves have no knowledge about the packet loss and burstiness levels. The changes in behavior are modeled only based on the amount of lost speech of each utterance and the probability. These models enable the simulated conversations to replicate the changes in interactivity due to bursty packet loss.

5.5 Summary

In this chapter, the simulation was extended with a conversation disruption model that adds dialogue to repair misunderstandings, as well as a turn-taking model that reflects the changes in interactivity due to bursty packet loss. An analysis of empirical conversation with high amounts of bursty packet loss revealed that the interactivity of a conversation, measured by lengths of silences (state probability MS) and speaker alternation rate,

decreases with increasing packet loss probability. In order to assess the effects of packet-loss-caused misunderstandings on the conversation, the concept of conversation disruption was defined, and it was shown that conversation disruptions significantly correlate with bursty packet loss. A model was created that simulates these conversation disruptions based on the amount of lost speech within one utterance. However, as then changes in conversational interactivity cannot be explained by the conversation disruptions alone, the turn-taking of conversations with packet loss was analyzed. This analysis revealed that the timing of turn-transitions changes with increasing packet loss levels, as the conversation partners do not take turns as quickly. This change was implemented into the turn-taking mechanism by dampening the turn-transitioning timing for every conversation disruption experienced by the simulated agent.

The final analysis showed that the packet-loss-adapted turn-taking model, in combination with the implemented conversation disruption mechanism, is able to reproduce the changes in the conversation structure due to bursty packet loss.

Chapter 6
Conversational Quality Predictions

In the last chapters, a conversation simulation architecture was described and extended for delayed transmission and bursty packet loss. The simulation is able to capture the differences in conversational parameters of two conversation scenarios with distinct interactivity levels and is extended to adapt turn-taking of the simulated agents when a delayed transmission is detected. The resulting conversations capture the differences in interactivity inherent in the conversation scenarios as well as in the changes in interactivity due to delay. The simulation is also extended to simulate *conversation disruptions* that occur due to misunderstanding of packet-loss-affected speech as well as differences in turn-taking due to the degraded speech signal.

In this chapter, the simulated conversations are used to predict the conversational quality based on the parameters of the conversation, as well based on the fullband E-model, with a focus on transmission delay and bursty packet loss. In order to test the suitability of simulated conversations for the prediction, a parameter-based model for the prediction of the impact of delay on a conversation is created and evaluated based on empirical data. Then, the E-model is extended towards interactivity parameters used to predict the effects of transmission delay, as well as towards burstiness parameters to predict the impact of bursty packet loss on a conversation. The adaptions to the E-model are evaluated on empirical data as well as on predictions from the POLQA model. Finally, the created parameter-based model and the extended E-model are utilized to predict the conversational quality based on the simulated conversations, and the results are compared to the empirical data and prediction of the extended E-Model based on general transmission parameters.

The modeling of the impact of transmission delay based on interactivity parameters has been partially published in Michael and Möller (2020b). The delay and packet-loss extension to the fullband E-model have been published in Michael et al. (2020) and Michael et al. (2021). Parts of the evaluation of the simulation by predicting speech quality with the extended E-model has been published in Michael and Möller (2021).

6.1 Predicting Quality of Conversations with Delay

Echo-free delay of speech transmission in a conversational scenario is not audible, but it affects how the interlocutors interact. It slows down the pace with which the two speakers can alternate and causes them to interrupt each other unintentionally

and more frequently. The amount of impact on the interactivity (and thus, on the perceived conversational quality) worsens with increasing overall transmission delay time. However, the degree to which the conversational quality is affected also depends on the interactivity of the conversation that is being held. Thus, to accurately predict the impact of echo-free transmission delay on the conversational quality, the transmission delay, as well as the interactivity of the specific conversation scenario, have to be taken into account. The fullband E-model, which is the most recent ITU-T-standardized parametric model for the prediction of conversational quality, does not include the interactivity of conversations in its calculation of the delay impairment factor. Also, the parametric approach of the E-model requires interactivity parameters to be available for each type of conversation that is studied, as differences in the interactivity of specific conversations are abstracted by generally applicable parameters.

Thus, the impact of transmission delay on the conversational quality of RNV and SCT conversations is predicted with two different approaches. First, based on the delay impairment factor formula of the narrowband E-model, an extension of the fullband E-model adds parameters to the delay formula that reflect the interactivity of the conversation scenario. In a second approach, the conversational quality of conversations impaired with transmission delay is predicted with a linear model that uses interactivity parameters extracted by a P-CA of the conversations under study. Because the model predicts the quality directly from recorded conversations, it predicts the conversational quality of particular conversations and not a MOS averaged over multiple conversations with the same condition. The extended E-model and the interactivity model are evaluated with data from the CONVSIM and UWS datasets.

6.1.1 E-model Extension for Interactivity and Delay

The E-model calculates impairments due to delayed transmission of the speech signal with the impairment factor Id. This includes impairments due to pure, echo-free transmission delay (Idd), impairments due to talker echo ($Idte$), and impairments due to listener echo ($Idle$). For the wideband and fullband versions of the E-model, the impairment factor Idd is calculated only based on the overall one-way delay Ta (see Equations 2.18 and 2.19 for the wideband E-model and Equations 2.22 and 2.23 for the fullband E-model).

Figure 6.1 shows the Idd, FB values for the conversations of the CONVSIM and UWS dataset (assuming delay is the only impairment present in the recorded conversations), split by SCT and RNV conversations. To calculate the Idd, FB impairment factor from the MOS recorded in the experiments, the Equation 2.9 was used and, assuming no other degradation, the resulting transmission rating was converted using Equation 2.21. A significant difference between the impact of transmission delay on the highly interactive RNV conversations and the SCT conversations with low interactivity is present, which shows that the impact of transmission delay on the conversational quality is higher for RNV conversations. The fullband E-model predicts an overall higher Idd, FB than the data points of the CONVSIM and UWS datasets suggest, resulting in an overall more pessimistic prediction.

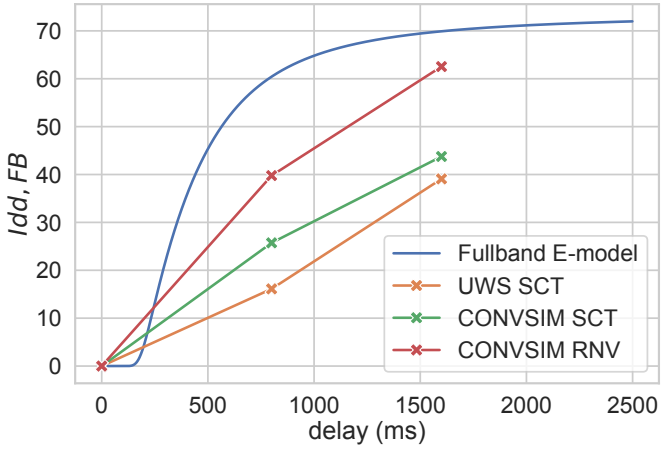

Fig. 6.1 Idd, FB of the UWS SCT, CONVSIM SCT and CONVSIM RNV conversations, as well as predicted by the fullband E-model at different levels of transmission delay.

To extend the fullband E-model to model these differences that stem from the interactivity of the conversations, the parameters sT and mT defined in the narrowband E-model (Equation 2.11 and 2.12) are introduced to the Idd formula of the fullband E-model:

For $Ta \leq mT$:

$$Idd = 0$$

For $Ta > mT$:

$$Idd = 1.48 \cdot 25\{(1 + X^{6 \cdot sT})^{\frac{1}{6 \cdot sT}} - 3(1 + [\frac{X}{3}]^{6 \cdot sT})^{\frac{1}{6 \cdot sT}} + 2\} \tag{6.1}$$

with:

$$X = \frac{\log(\frac{Ta}{mT})}{\log 2} \tag{6.2}$$

Using the sT and mT class "Low" for RNV conversations and "Very low" for SCT conversations as described in Table 2.5, the predictions for the two different conversation types are adapted. Figure 6.2 shows the predictions of the current fullband E-model and the extension using "Low" interactive parameters for delay levels up to 1.6 s and compares them to the RNV conversations of the CONVSIM dataset. The extended version of the E-model predicts the conversational quality better than the current fullband E-model, with the predictions being inside the 95 % confidence interval of the CONVSIM RNV MOS at 0 s, 0.8 s, and 1.6 s delay.

Figure 6.3 shows the predictions of the two E-model versions for SCT conversations, as well as the MOS of SCT conversations of the CONVSIM and UWS datasets. The notably higher MOS for this scenario with low interactivity is reflected in the extended version of the E-model. The predictions of the extension lie within the 95 % confidence interval of the SCT dataset while being outside of the confidence interval of the UWS

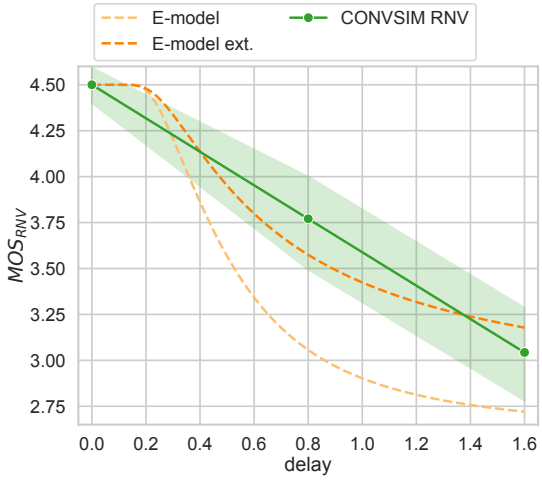

Fig. 6.2 MOS predictions of the current and extended fullband E-model, as well as the MOS of the CONVSIM dataset with 95 % confidence interval at delay levels up to 1.6 s for RNV conversations.

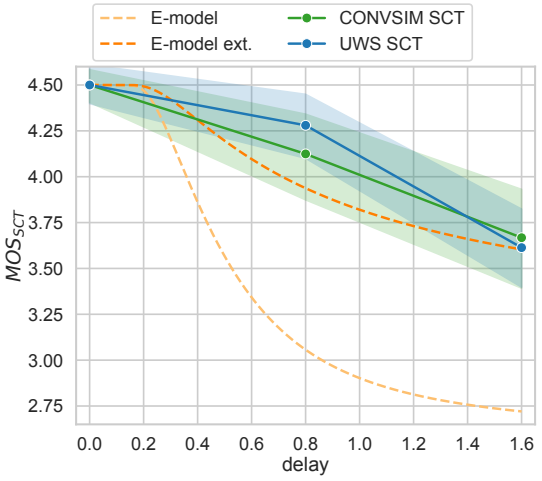

Fig. 6.3 MOS predictions of the current and extended fullband E-model, as well as the MOS of the CONVSIM and UWS datasets with 95 % confidence interval at delay levels up to 1.6 s for SCT conversations.

dataset at 800 ms delay. Especially for this type of conversation, the current E-model formula is overly pessimistic, while the extended Idd formula is able to capture the difference in MOS.

The results show that the extension of the E-model by the mT and sT parameters of the narrowband Idd formula is suited to predict the differences in MOS between SCT and RNV conversations.

6.1.2 Quality Prediction from Interactivity Parameters

In order to create a model for the prediction of conversational quality from a single conversation based on its interactivity parameters and the transmission delay, suitable features need to be selected. To aid this selection, 16 different interactivity parameters are analyzed: the four state probabilities (ms, dt, sa, and sb), the respective sojourn times (st_ms, st_dt, st_sa, and st_sb), the speaker alternation rate, and its corrected version (sar and sarc), the interruption rate, intended interruption rate and unintended interruption rate (ir, iir, and uir), the double talk rate (dtr), pause rate (pr) and the conversational temperature (temp). A correlation matrix of these parameters based on all conversations of the CONVSIM dataset with 0 ms, 800 ms, and 1600 ms of delay are shown in Figure 6.4. The heatmap shows a positive correlation between the state probabilities and their sojourn times. The sojourn time of mutual silence is inversely correlated with the other three states. Both SAR and SARc correlate positively with each other and the conversational temperature. Also, the pause rate correlates positively with the SAR, as conversations with more speaker alternations tend to have more pauses.

Figure 6.5 shows the correlation of these 16 interactivity parameters with the average conversational quality rating of that conversation. Due to the strong variations in the quality ratings of the participants, the maximum positive and negative correlations are

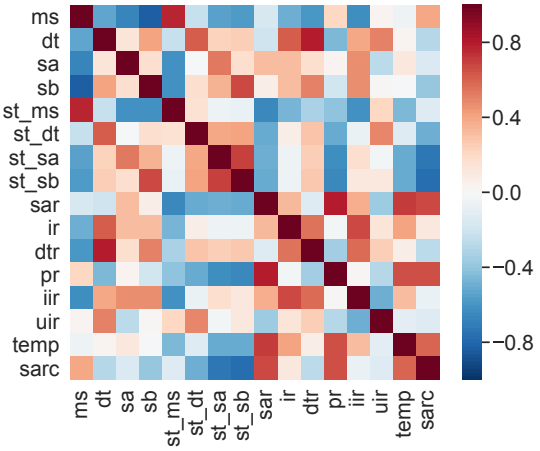

Fig. 6.4 Correlation matrix of the 16 interactivity parameters of conversations in the CONVSIM dataset.

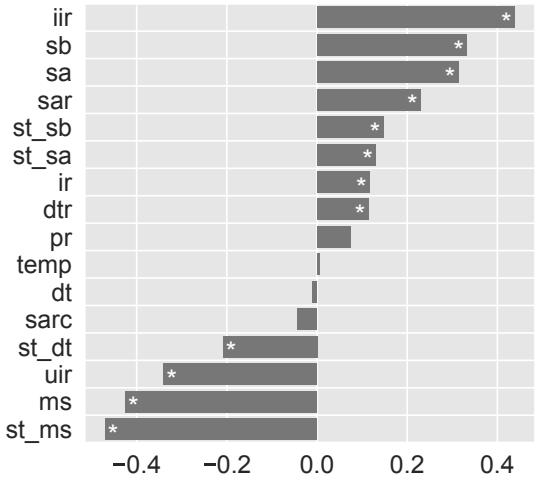

Fig. 6.5 Spearman correlation of each interactivity parameter with the respective quality rating of the conversation. Correlations marked with a "*" are significant.

low, with 0.44 for the intended interruption rate and −0.47 for the sojourn time of the state mutual silence. The pause rate, conversational temperature, the state probability of double talk, and the corrected speaker alternation rate do not correlate significantly with the conversational quality. This lack of correlation is intended for the corrected speaker alternation rate, which should reflect the number of speaker alternations independent of the transmission delay.

Based on the correlation between the interactivity parameters and the correlation with the conversational quality, the following linear model was fitted with the least-squares method:

$$\widehat{CQ} = 4.4824 - 1.4055 \cdot ms - 0.5665 \cdot delay + 0.113 \cdot (sar \cdot delay) \tag{6.3}$$

This model uses the overall one-way delay in seconds, the speaker alternation rate (sar), and the state probability of mutual silence (ms) as input parameters to predict the conversational quality of the conversation. It is fitted on the conversations of the CONVSIM dataset and has an adjusted R^2 of 0.44.

Figure 6.6 compares the predictions of the interactivity model in Equation 6.3 with the per conversation MOS of the CONVSIM dataset and Figure 6.7 shows the comparison with the unseen UWS dataset. On the training data, it has an RMSE of 0.55 and a slightly higher RMSE on the UWS test dataset with 0.58. This high RMSE is to be expected because the labels the models were trained on are noisy due to the high interpersonal variance in the subjective ratings. The evaluation shows that the linear model is able to capture the variances in conversational quality rating on the basis of interactivity parameters of individual conversations. It also shows that the participants' handling of the added transmission delay changes how much the delay degrades the interactivity and, thus, how the overall quality of the conversation is perceived.

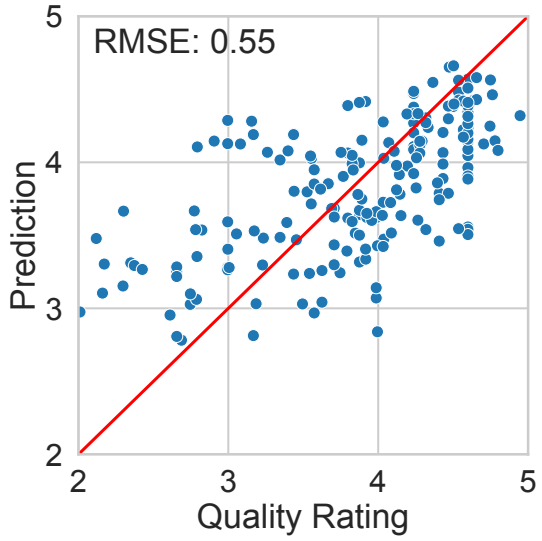

Fig. 6.6 Predictions of the linear interactivity model and conversational quality of conversations in the CONVSIM dataset used for training.

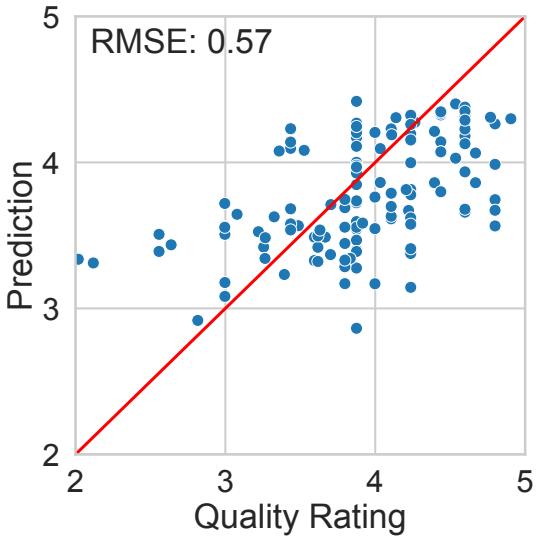

Fig. 6.7 Predictions of the linear interactivity model and conversational quality of conversations in the previously unseen UWS test dataset.

Table 6.1 RMSE for the interactivity-based model ("$\widehat{CQ}$"), the fullband E-model ("E-model"), and the fullband E-model extended by the Idd-formula of the narrowband E-model ("E-model extension") on the training (CONVSIM) and test dataset (UWS).

dataset	delay (s)	scenario	SD data	RMSE $\widehat{CQ}$	RMSE E-Model	RMSE E-Model extension
CONVSIM	0	RNV	0.5079	0.5089	**0.5079**	**0.5079**
		SCT	0.4927	0.5049	**0.4927**	**0.4927**
	0.8	RNV	0.6886	**0.6866**	0.9917	0.7159
		SCT	0.6517	**0.6544**	1.2507	0.6780
	1.6	RNV	0.7548	**0.7408**	0.8209	0.7669
		SCT	0.8042	**0.7947**	1.2425	0.8068
UWS	0	SCT	0.3886	**0.3625**	0.3886	0.3886
	0.8	SCT	0.5697	**0.6522**	1.3495	0.6649
	1.6	SCT	0.6835	0.7072	1.1245	**0.6835**

In order to apply the model on a *per-condition* basis, the predictions of the interactivity-based model were averaged over the levels of transmission delay and the conversation scenarios. Table 6.1 shows the MOS RMSE of the per condition evaluation of the model and compares it to the standard deviation in the data, as well as to the fullband E-model and its delay extension. To accurately compare the models, the RMSE of the E-model were also calculated compared to the MOS of the individual conversations and not to the per-condition MOS. The performance of the three models in terms of their RMSE is comparable, with the current fullband E-model having a slightly worse

performance with an RMSE of over 1 in some conditions. The RMSE of the models is very similar to the Standard Deviation (SD) of the data, indicating that the variance in the data is roughly as high as the uncertainty of the model.

Overall, both the interactivity model and the extended fullband E-model are capable of reflecting the interactivity of the conversation in their prediction and, thus, are suitable to be used to predict the quality of simulated conversations that are impacted by transmission delay.

6.2 Predicting Quality of Conversations with Packet Loss

Packet loss may occur whenever a packet containing coded speech is lost during transmission (e.g., when a routing problem occurs), when a packet arrives too late in order to be included in the decoding process, or when an arriving packet is corrupted, and the retransmission of the same packet would take too long. A speech packet usually contains 20 ms of speech, and thus, a single lost packet either causes the speech to drop out momentarily or, when a codec with PLC is used, the missing part of the signal is estimated, and the lost packet might not even be noticed by the user. This type of packet loss and the resulting audible effects are being taken into account by current quality models like the fullband E-model. However, when packets get lost by the transmission network or a large increase in transmission time occurs, usually multiple consecutive packets are affected. This burstiness of the packet loss results in vastly different audible and conversational effects. When multiple packets are being dropped, a PLC algorithm is no longer able to extrapolate the speech signal, and the speech drops out. Depending on the duration of the burst, the understandability of the transmitted speech might be affected. This shifts the impact of a packet loss occurrence on the conversation from an audible annoyance to an event that changes the flow of the conversation.

Thus, in this section, the codec-related impairment factor of the current fullband E-model is extended to account for bursts in packet loss and evaluated on the CONVSIM dataset. Then, the effects of conversation disruptions caused by highly bursty packet loss on the conversational quality are analyzed, and resulting interactions with transmission delay are discussed.

6.2.1 Bursty Packet Loss E-model Extension

In order to analyze the performance of the current fullband E-model with respect to random packet loss and to extend it towards bursty packet loss, fullband speech data is coded with 16-bit linear PCM at 44.1 kHz (which is used in the CONVSIM dataset) as well as with the commonly used EVS codec in super-wideband mode at 13.2 kbit/s. The coded speech was degraded with bursty packet loss and finally assessed with the well-validated signal-based model POLQA (ITU-T Recommendation P.863, 2014). Based on these predictions, the current codec-related impairment factor of the fullband E-model is evaluated, and an extension is proposed. Finally, this new burstiness extension, together

with the delay and interactivity extension, are evaluated with the CONVSIM dataset by predicting the conditions with bursty packet loss, transmission delay, as well as the combination of the two.

In order to analyze and extend the fullband E-model with respect to bursty packet loss, a set of 18 clean mono speech files with 16-bit linear PCM at 44.1 kHz sampling rate was used. These speech files were then degraded with every combination of 6 packet loss rates (2.5, 5.0, 7.5, 10.0, 20.0, and 30.0 %) and 7 burst ratios (1.0, 1.5, 2.0, 2.5, 3.0, 3.5, and 4.0), resulting in 756 samples for each PCM and EVS. The packet loss in the PCM-coded speech was created by inserting zeros for every lost packet (zero-insertion packet loss), while for the EVS-codec, the native PLC functionality was used. In order to generate the packet loss patterns, a two-state Markov chain was used. Only patterns were allowed that deviate from the targeted Ppl by no more than one percentage point.

For the instrumental assessment with POLQA, the SQuadAnalyzer software in its super-wideband mode and POLQA version 3 was used. Because the range of predicted MOS differs between POLQA and the E-model, the predictions of POLQA were linearly scaled to the range of the E-model (1.0 to 4.5). Additionally, Equation 2.9 with $Rx = R \div 1.48$ was used to transform the POLQA predictions into a transmission rating R. Assuming that the only degradations present in the speech files are codec-related, the effective equipment impairment factor is calculated with $Ie,eff,FB = Ro,FB - R$. Furthermore, the equipment impairment factor Ie,FB is set to 0 for PCM (as by definition, there is no impairment in this case) and to 24.8 for EVS based on the analysis in Mittag et al. (2018).

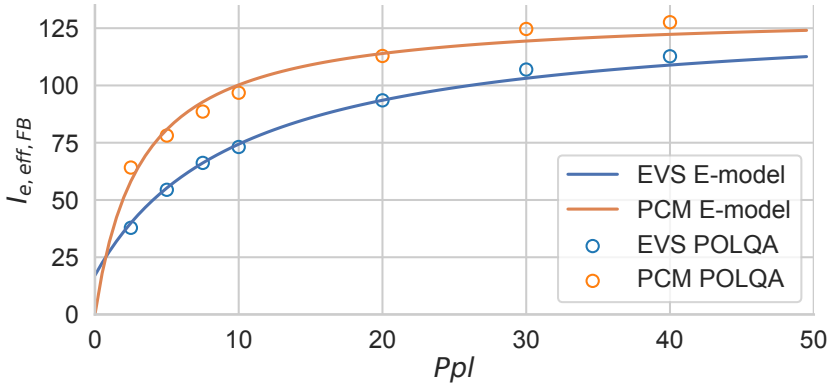

Fig. 6.8 Ie,eff,FB values as predicted by POLQA and the fullband E-model for EVS at 13.2 kbit/s and linear PCM for random packet loss (BurstR = 1.0).

Figure 6.8 shows that the prediction of the E-model and POLQA for both EVS and PCM agree at various packet loss levels when only considering randomly distributed losses. PCM shows lower robustness against packet loss even for very low Ppl values. Figures 6.9 and 6.10 compare the MOS of the current fullband E-model to the predictions of EVS and POLQA respectively, with burst rates visualized in different colors. The E-model in its current form does not include the burst ratio $BurstR$ in the Ie,eff,FB

calculation and thus, the six distinct levels of packet loss are visible in the clusters with the same E-model prediction. However, both predictions have a high Pearson correlation ρ of 0.9796 for EVS and 0.9655 for PCM.

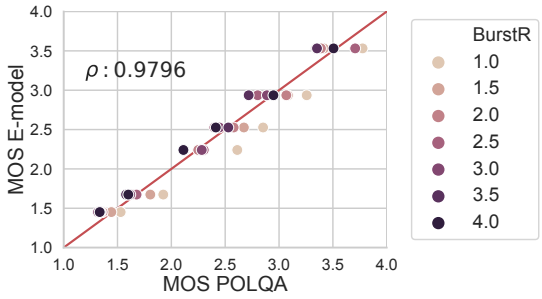

Fig. 6.9 Predicted MOS of POLQA versus the predicted MOS of the fullband E-model for EVS at 13.2 kbit/s.

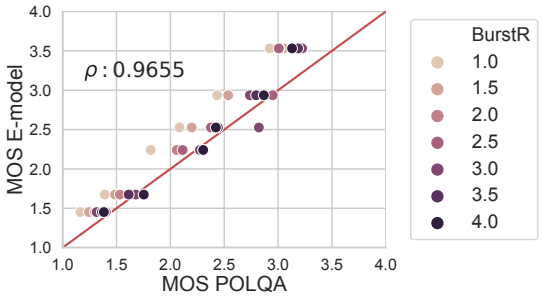

Fig. 6.10 Predicted MOS of POLQA versus the predicted MOS of the fullband E-model for PCM.

To accommodate for the burstiness of packet loss, the narrowband E-model introduces the BurstR parameter defined in Equation 2.16. This approach penalizes the burstiness of packet loss by dividing the Ppl-term in the divisor by the burst ratio (see Equation 2.15). However, as the narrowband E-model states, this penalization is too strong for $BurstR > 2$ when $Ppl > 2$. Thus, the proposed extension for bursty packet loss includes $BurstR$ in the dividend of the equation and introduces a new "burstiness robustness factor" Brf:

$$Ie,eff,FB = Ie,FB + (132 - Ie,FB)\frac{Ppl - \frac{1-BurstR}{Brf}}{Ppl + Bpl} \tag{6.4}$$

For randomly distributed packet loss ($BurstR = 1.0$), this extension results in the current Ie,eff,FB as shown in Equation 2.24. With increasing burst ratio, the Ie,eff,FB increases, but in contrast to the narrowband version of the formula (Equation 2.15), the penalization of burstiness is independent of the packet loss probability. The amount of penalization in Equation 6.4 can be regulated with the Brf parameter, with higher values indicating higher robustness against burstiness. The introduced parameter Brf is fitted together with the Bpl value and may be different for every codec, packet size, and packet loss concealment used.

Table 6.2 shows the Bpl, Brf, and RMSE values for the EVS and PCM codecs. The Bpl value for EVS is similar to the value calculated in Mittag et al. (2018) and the according Brf value suggests only slight robustness against burstiness. For PCM the robustness factor is negative, indicating that the quality of bursty packet loss is rated higher than the quality of randomly distributed packet loss. This is in line with the predictions of POLQA seen in Figure 6.10, where samples with $BurstR = 1.0$ have a lower predicted MOS than the samples with higher burstiness.

The improved accuracy of the modified E-model can be seen in Figure 6.11 for the EVS codec. The Pearson correlation improves to $\rho = 0.9891$ compared to the current fullband E-model. The positive Brf-term decreases the MOS for increasing burst ratios. While this effect is strong for low Ppl values, it gets weaker for higher packet loss probabilities. In contrast, for the PCM codec with zero-insertion packet loss

shown in Figure 6.12, the negative Brf value shows an increase in MOS for higher values of $BurstR$. While the extended E-model is able to describe this behavior, the overall prediction quality is slightly lower with a Person correlation of $\rho = 0.9540$.

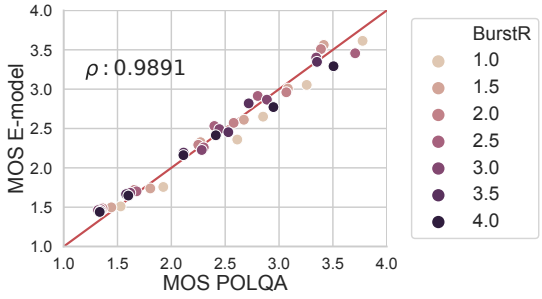

Fig. 6.11 Predicted MOS of POLQA versus the predicted MOS of the extended E-model for EVS at 13.2 kbit/s.

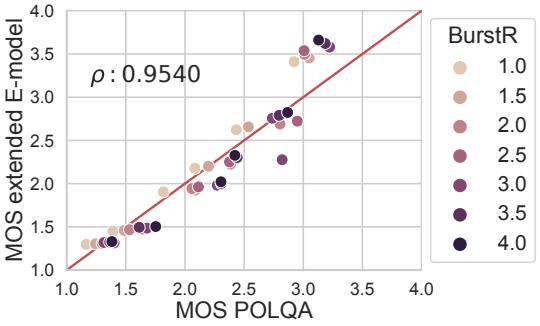

Fig. 6.12 Predicted MOS of POLQA versus the predicted MOS of the extended E-model for PCM.

However, due to the fact that PCM coded speech is not used in common speech transmission systems, the specific agreement of the E-model and POLQA should not be over-interpreted.

To evaluate this burstiness extension, as well as the extension for interactivity and delay described in Equation 6.1 and 6.2, the predictions of the extended E-model are compared to the MOS of the CONVSIM dataset, where each combination of 0 %, 15 %, and 30 % packet loss at a burst ratio of 4.0 with 0 ms, 800 ms, and 1600 ms of transmission delay was rated. The Bpl and Brf values for PCM with zero insertion packet-loss were derived using – in addition to the CONVSIM dataset – the results of a separate conversation experiment with 0 %, 5 %, 15 %, 25 %, and 35 % PCM-coded zero-insertion packet loss and a burst ratio of 1.0. The resulting Bpl of 21.79 and Brf of −6.9 lead to a more optimistic prediction than the POLQA-fitted parameters shown in Table 6.2. For the predictions with transmission delay, the dataset is split into SCT and RNV conversations. As recommended by the narrowband E-model, the interactivity parameters were set to $sT = 0.4$ and $mT = 150$ for SCT predictions and $sT = 0.55$ and $mT = 120$ for RNV predictions. The packet loss in the CONVSIM dataset has a constant burst ratio of 4.0, so the $BurstR$ parameter of Equation 6.4 is set to that value.

Figure 6.13 shows the extended E-model predictions and CONVSIM MOS values for RNV conversations, and Figure 6.14 for SCT conversations, all at a burst ratio of 4.0. The predictions for conversations without transmission delay are modeled well by the extensions. Also, the difference in the interactivity of SCT and RNV conversations

Table 6.2 Bpl, Brf, and according RMSE values as calculated with the extended equipment impairment factor formular in Equation 6.4.

Codec	Bpl	Brf	RMSE
EVS	8.96	2.03	7.45
linear PCM	5.01	−4.35	11.75

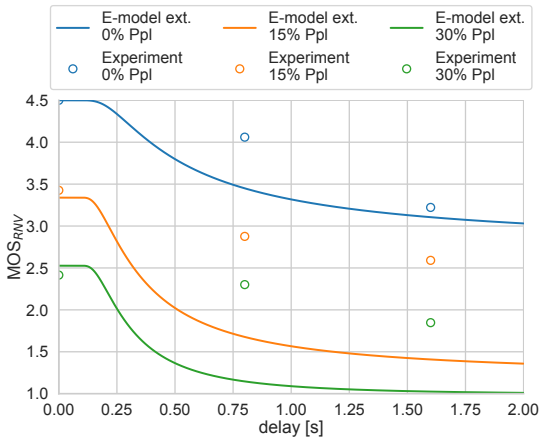

Fig. 6.13 Predictions of the extended E-model using Equation 6.4 for bursty packet loss and Equations 6.1 and 6.2 for delay, compared to RNV conversations of the CONVSIM dataset.

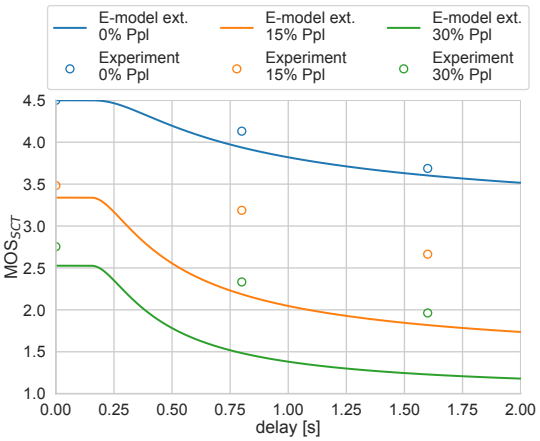

Fig. 6.14 Predictions of the extended E-model using Equation 6.4 for bursty packet loss and Equations 6.1 and 6.2 for delay, compared to SCT conversations of the CONVSIM dataset.

at 0 % packet loss is modeled accurately (shown in blue). However, the combination of both degradations is predicted to be substantially lower than the subjective MOS. The quality ratings in the CONVSIM dataset at 30 % packet loss are less affected by delay than predicted by the extended E-model. Similarly, the ratings or conversations at 1600 ms delay are also higher than the predictions. Slightly pessimistic worst-case predictions are expected in the E-model, as too optimistic predictions would put the planning process at risk.

Overall, the newly introduced formula for bursty packet loss is able to model the differences in quality as seen in the CONVSIM dataset and in the POLQA predictions. However, the combination of high transmission delay and highly bursty packet loss leads to a very pessimistic prediction of the E-model.

6.2.2 Conversation Disruptions and Quality

In Chapter 5 it was shown that highly bursty packet loss causes misunderstandings that the interlocutor needs to repair in order to continue the conversation. These *conversation disruptions* increase with the packet loss probability and its burstiness. Because these occurrences disrupt the flow of the conversation, the SAR of especially highly interactive conversations, like the RNV task, decreases with higher packet loss rates. It has also been shown that measured by the number of conversation disruptions per minute (CDR, see Figure 5.7), the RNV is more affected than SCT conversations. However, the average number of disruptions per turn (see Figure 5.8) does not significantly differ between the two conversational scenarios.

Figure 6.15 shows the MOS of the conversations of the CONVSIM dataset at 0 %, 15 %, and 30 % packet loss, split by SCT and RNV scenarios. No differences between the two scenarios are visible at the different packet loss levels. Thus, while a difference in CDR between SCT and RNV conversations is present, it is not reflected in the mean overall quality rating of the participants. Figure 6.16 shows the number of conversation

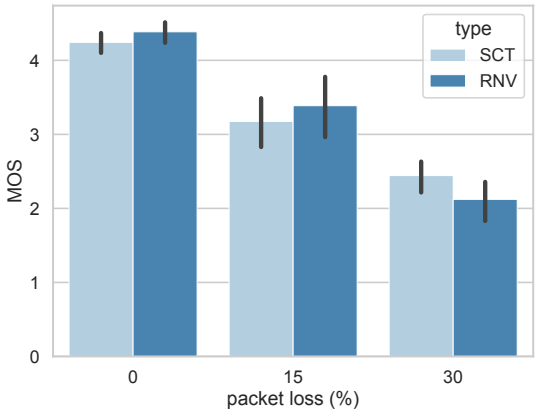

Fig. 6.15 Mean Opinion Score of SCT and RNV conversations at 0 %, 15 %, and 30 % packet loss. Error bars indicate the 95 % confidence interval.

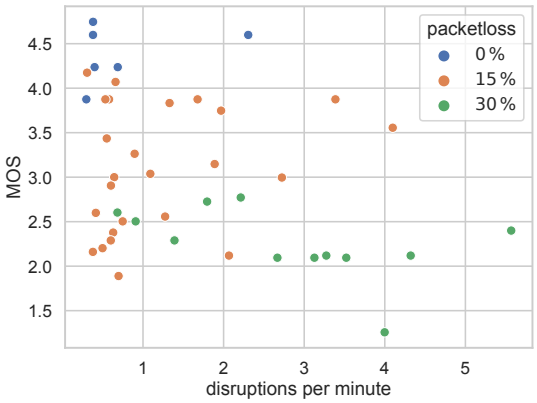

Fig. 6.16 Scatter plot of MOS versus conversation disruptions per minute at 0 %, 15 %, and 30 % packet loss.

disruptions per minute against the MOS per conversation (excluding all conversations where no disruption was present). The per conversation MOS correlates significantly with the disruptions per minute with $\rho = -0.61$. This indicates that the number of conversation disruptions can be a strong indicator of conversational quality. Because these disruptions occur mostly when the packet loss is highly bursty, it cannot be used as the only indicator of the degradation of conversational quality due to packet loss.

Due to the fact that the interactivity of the conversation has a negligible influence on the quality rating of the packet loss degradation, an extension of the E-model to include interactivity parameters seems unpromising. A parameter-based model that predicts the quality per conversation from the number of conversation disruptions is also not performed here due to the lack of an annotated evaluation dataset. However, the presented E-model extension towards burstiness of packet loss is able to capture the changes in quality described here.

6.2.3 Interaction between Delay and Packet Loss

Even though the interactivity of a conversation does not directly impact the quality rating of conversations affected by bursty packet loss, it has been shown in Chapter 5 that an increase in conversation disruptions leads to a higher state probability of mutual silence (Figure 5.4) and to a decrease of SAR in RNV conversations (Figure 5.3). As the interactivity of a conversational scenario directly affects the degree to which delay impacts the turn-taking and thus the conversational quality, this change in conversational structure due to packet loss influences the quality indirectly.

As dialogue that repairs misunderstood utterances is significantly slower than the highly interactive dialogue of the RNV conversations, the packet loss leads to a slower interaction between the two interlocutors and more pauses. Thus, the focus shifts away from the delay of speech signal transmission. Especially in the predictions of the extended E-model of conversations with both delay and packet loss (see Figure 6.13 and

6.14), the combination of the two degradations was too pessimistic compared to the actual quality ratings of the CONVSIM dataset. This behavior of the E-model can be partly attributed to the strictly additive nature of the impairment factors and its use in transmission planning, where a too pessimistic prediction is favored over a too optimistic one. However, the effects of packet loss on the interactivity of the conversation may reduce the effects of transmission delay, which may also play its part in the less severe quality ratings.

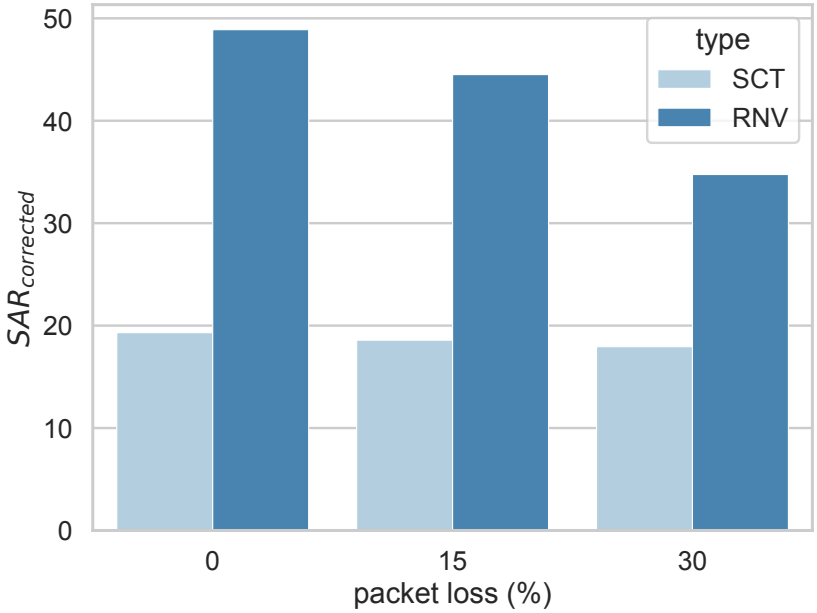

Fig. 6.17 SARc of SCT and RNV conversations in the CONVSIM dataset, calculated over conversations with 0 ms, 800 ms, and 1600 ms delay and split by levels of 0 %, 15 %, and 30 % bursty packet loss.

In the narrowband E-model and the fullband delay extension, the interactivity parameters for the minimal perceivable delay (mT) and delay sensitivity (sT) are calculated with the Corrected Speaker Alternation Rate (SARc) as shown in Equations 2.13 and 2.14. Figure 6.17 shows the SARc of conversations with 0 ms, 800 ms, and 1600 ms transmission delay at the three levels of packet loss only decreases slightly for the SCT conversations, with 19.3 alternations per minute at 0 % packet loss and 17.9 at 30 %. However, for RNV conversation, the SAR reduces from 48.9 alternations per minute at 0 % packet loss down to 44.5 alternations per minute at 15 % and 34.8 alternations per minute at 30 % packet loss. This translates into a reduction of delay sensitivity and an increase of minimal perceivable delay.

As the simulation of conversation disruptions in conversations with bursty packet loss replicates the changes in interactivity, the simulated conversations can be utilized to predict the interaction effects of delay and packet loss with the extensions of the fullband E-model.

6.3 Predicting Quality from Simulations with Delay

In Chapter 4, the simulation was adapted to reflect the turn-taking and interaction parameters of conversations with different interactivity (namely, SCT and RNV conversations) and later on extended to reflect the changes in turn-taking due to transmission delay. These simulated conversations can now be used to predict the conversational quality with the interactivity model or with the extended E-model. The linear interactivity model directly utilizes the interactivity parameters of the simulation, while the extended E-model calculates the minimal perceivable delay and delay sensitivity from the simulated conversations.

For the evaluation with both quality models, 30 SCT and 30 RNV conversations from 0 ms transmission delay up to 2000 ms transmission delay in 100 ms time steps were simulated, resulting in 1260 simulated conversations.

6.3.1 Prediction from Interactivity Parameters of Simulations

In order to apply the interactivity model in Equation 6.3, a P-CA is used to extract the interactivity parameters ms (state probability of mutual silence) and sar (the Speaker Alternation Rate) from the simulated conversations. Together with the one-way overall delay in seconds, the quality is predicted for each simulated conversation.

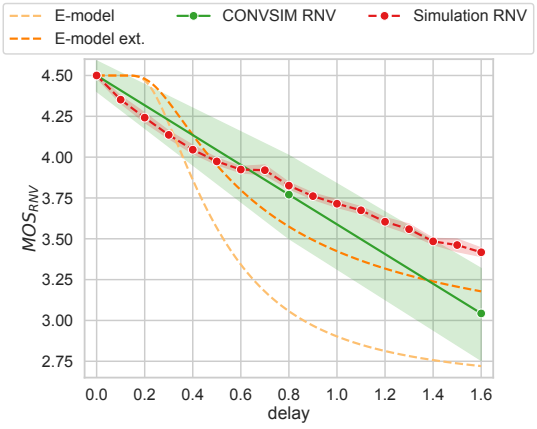

Fig. 6.18 MOS predictions of the current and extended fullband E-model, as well as the MOS of the CONVSIM dataset (CONVSIM RNV) and the prediction from the simulated RNV conversations with the linear interactivity model (Simulation RNV) at delay levels up to 1.6 s.

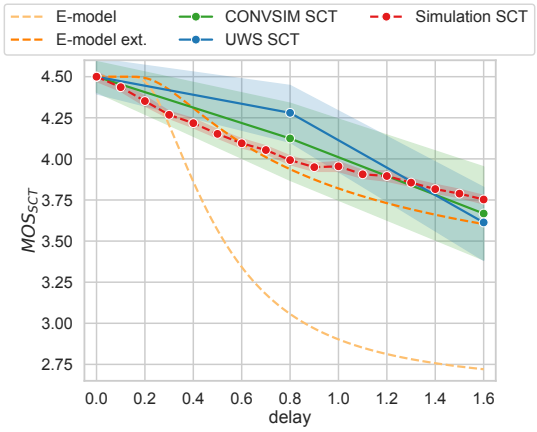

Fig. 6.19 MOS predictions of the current and extended fullband E-model, as well as the MOS of the CONVSIM and UWS dataset, as well as the prediction from the simulated SCT conversations with the linear interactivity model (Simulation SCT) at delay levels up to 1.6 s.

Figure 6.18 and 6.19 show the quality prediction from the simulated RNV and SCT conversation, respectively, as estimated by the interactivity model. For RNV conversations, the interactivity model becomes too optimistic for high delay levels but mostly stays inside the 95 % confidence interval of the empirical MOS. A slightly too low state

probability MS might have a large impact on the prediction quality of the interactivity model, as this parameter is correlated negatively with the overall conversational quality. For SCT conversations, the prediction stays inside the 95 % confidence interval of the CONVSIM dataset, but between 500 ms and 1000 ms, it is slightly below the CI of the UWS dataset. Generally, the difference in the degradation of SCT and RNV conversation is replicated well by the simulation approach.

Overall, the simulated conversations yield a good prediction quality when used with the interactivity model in Equation 6.3. In contrast to the prediction based on parameters of the transmission alone, the simulations provide variance in the underlying conversations. This results in the possibility of reporting the standard deviation and confidence intervals of the predictions. However, the validity of these results needs to be investigated further, as both the interactivity model and the simulation are partly based on the very limited CONVSIM dataset.

6.3.2 Prediction from Extended E-Model

The delay-extended Idd formula of the E-model shown in Equation 6.1 and 6.2 utilize the minimal perceivable delay mT and the delay sensitivity sT of a conversation in order to predict the impact of transmission delay on conversations with different interactivity. While the narrowband E-model provides standardized mT and sT parameters (shown in Table 2.5), they can also be calculated with Equations 2.13 and 2.14 using the Corrected Speaker Alternation Rate (SARc).

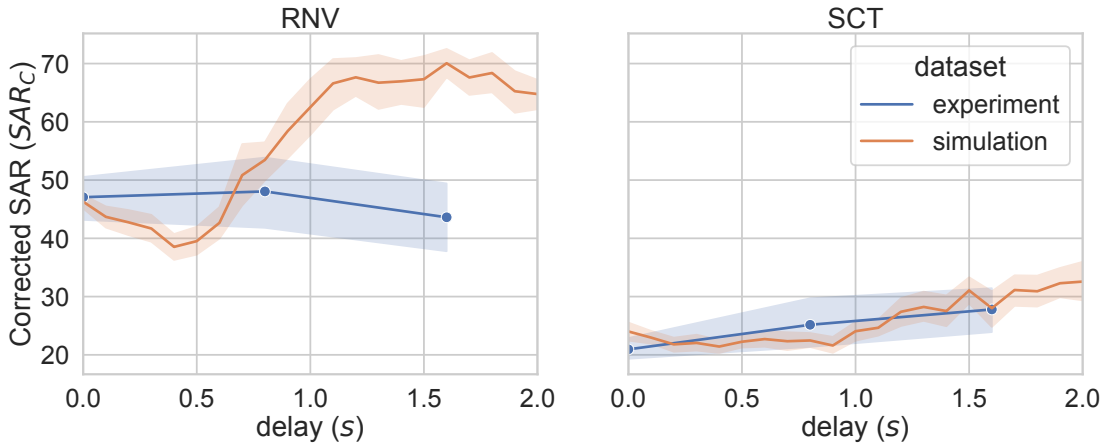

Fig. 6.20 Corrected speaker alternation rate for empirical and simulated RNV conversations (left) and SCT conversations (right) at various delay levels.

The SARc is extracted from the simulated conversation based on the formula in Equation 2.3. Figure 6.20 shows the SARc for the simulated RNV and SCT conversations and compares them to the SARc of the conversations in the CONVSIM dataset. The stability of the corrected SAR is only reproduced for the SCT simulations. For the simulated RNV conversations, the SARc overestimates the delay-adjusted Speaker Alternation Rate for delay levels above 800 ms. As the SAR of the simulated conversations match the empirical data (as shown in Figure 4.23), the discrepancy of the SARc lies in the

length of the simulated RNV conversations. While the empirical conversations tend to increase in length with higher levels of delay, the simulated conversations' increase in length is much lower. This results in an overestimation of the interactivity by the SARc formula. From these SARc values, the mT and sT parameters of each simulated conversation is calculated with the Equations 2.13 and 2.14 respectively. Finally, the MOS for each simulated conversation is predicted using the extended Idd formula given in Equations 6.1 and 6.2.

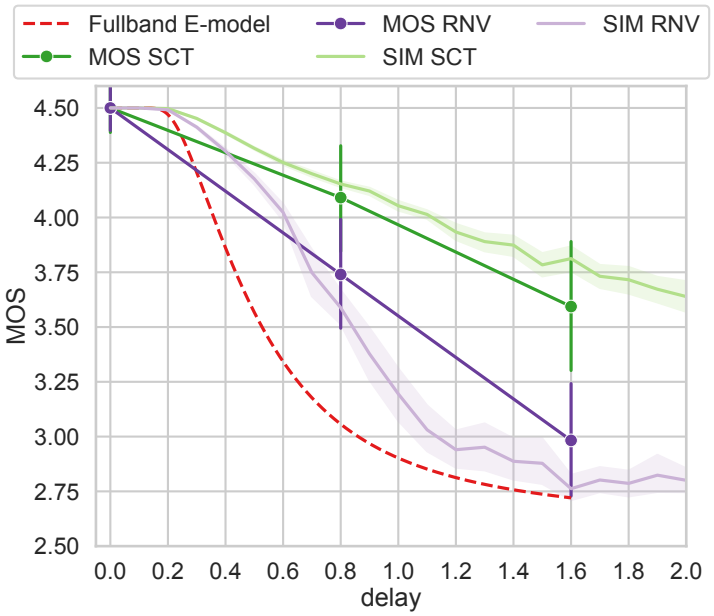

Fig. 6.21 Conversational Quality MOS from the conversation experiment, as well as the predictions from the current fullband E-model (red) and the predictions based on the extended E-model and simulations (light colors). For the experiment and the simulation, the MOS is split by RNV (purple) and SCT conversations (green).

Figure 6.21 shows the predicted MOS from the simulation and compares them to the MOS of the CONVSIM dataset, both split by conversation type and the prediction of the fullband E-model. As expected, due to the overestimation of the SARc of RNV conversations, the extended E-model assumes that the RNV conversations are more interactive than they are, which results in a pessimistic quality prediction. The prediction of the quality of SCT conversations is slightly too optimistic. However, both SCT and RNV predictions are inside the 95 % confidence intervals of the empirical MOS.

Overall, the approach of combining the simulations with the extended E-model is able to reproduce the behavior of SCT and RNV conversations and results in acceptable quality estimations. Again the variance in the simulated data produces standard deviations and confidence intervals of the quality predictions, which may be used to qualify the accuracy of the prediction further.

6.4 Predicting Quality from Simulations with Packet Loss and Delay

In Chapters 4 and 5, the simulation was adapted to model the changes in the interactivity of SCT and RNV conversations due to delay and bursty packet loss, as well as the changes in the conversation due to packet-loss-induced conversation disruptions. Because the effects of delay and packet loss are modeled independently of each other and are based on the incoming speech alone, the potential interaction between the added turn-taking due to conversation disruptions and the impact of transmission delay can be simulated.

For a final evaluation of the simulation approach, the two extensions of the full-band E-model are utilized to predict the quality of simulated conversations with both transmission delay and bursty packet loss. With the impact of packet-loss-induced conversation disruptions on the interactivity of conversations, the delay impairment factor of the extended E-model will take into account the impact of the packet loss. For this evaluation, 30 SCT and 30 RNV conversations with each combination of 0 %, 5 %, 10 %, 15 %, 20 %, 25 % and 30 % of packet loss with a burst ratio of $BurstR = 4.0$, as well as 0 ms, 400 ms, 800 ms, 1200 ms, and 1600 ms of transmission delay were simulated, resulting in 2100 simulated conversations.

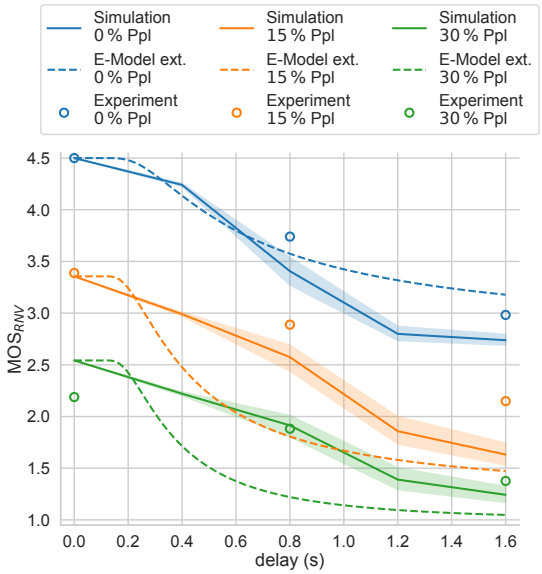

Fig. 6.22 MOS of RNV conversations of the CONVSIM dataset (circles) and as predicted by the extended E-model with general parameters (dashed lines) and with parameters from the simulation (solid lines) at 0 % (blue), 15 % (orange), and 30 % (green) bursty packet loss and various levels of overall one-way delay.

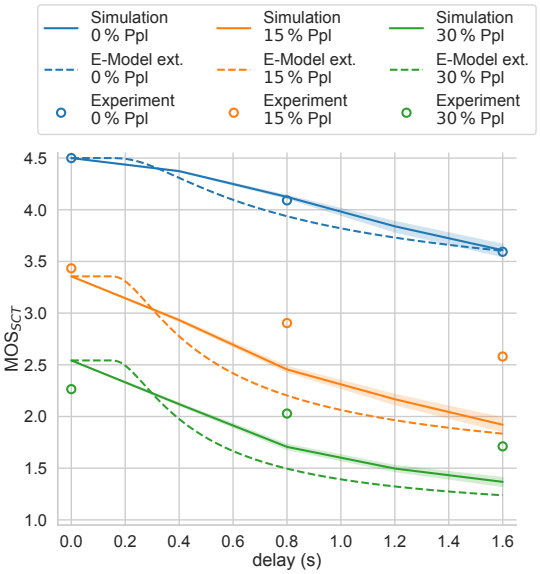

Fig. 6.23 MOS of SCT conversations of the CONVSIM dataset (circles) and as predicted by the extended E-model with general parameters (dashed lines) and with parameters from the simulation (solid lines) at 0 % (blue), 15 % (orange), and 30 % (green) bursty packet loss and various levels of overall one-way delay.

Figures 6.22 shows the performance of this approach when simulating RNV conversations and compares the results to the MOS obtained from the CONVSIM dataset and the predictions of the extended E-model, but using the general parameters provided in Table 2.5 together with the parameters of the transmission system. At 0 ms delay, the predictions of the three packet loss levels are independent of the parameters extracted

from the simulations, as the Idd impairment factors amount to 0.0. With increasing overall one-way delay, the prediction at 0 % packet loss drops below the empirical MOS and the predictions of the extended E-model. This behavior is in line with the evaluation shown in Figure 6.21 and is due to the overestimation of the Corrected Speaker Alternation Rate. For 15 % and 30 % packet loss, the reduction in interactivity dampens the degradation due to the transmission delay, which results in a more optimistic quality prediction than with just the extended E-model. While at 15 % packet loss, the prediction is still too pessimistic compared to the MOS of the CONVSIM dataset, at 30 % packet loss, the prediction matches the empirical data.

Figure 6.23 shows the predictions of the model for SCT conversations. Again, at 0 ms delay, the model just uses the parameters of the transmission, and at 0 % packet loss, the simulation performs similar to the delay-only evaluation shown in Figure 6.21. At 15 % and 30 % bursty packet loss, the prediction through the simulated conversations is also more optimistic than the extended E-model with the general transmission parameters. However, due to the fact that the SCT conversations are not as strongly impacted by the transmission delay, the effect is much less pronounced than for RNV conversations. This results in the simulation predictions being too pessimistic for 15 % and 30 % packet loss at higher delay levels.

The evaluation of quality predictions of simulations with delay and packet loss has shown that the estimation of interactions between both degradations can be simulated and that the resulting conversations can be used for the prediction of conversational quality. Here, the advantage of simulations over the purely transmission-parameter-based E-model is visible, as effects of degradations on the conversation itself can be modeled, and the resulting changes can be used to improve the quality estimation.

6.5 Summary

In this chapter, the simulation was evaluated by predicting conversational quality from simulated conversations. To be able to predict changes in conversational quality due to delay from simulated conversations, a new interactivity model was presented, and the fullband E-model was extended to include interactivity parameters. The interactivity model utilizes parameters extracted from a P-CA to predict the average overall quality of a specific conversation (as opposed to the MOS from a set of transmission parameters). The E-model extension is based on the delay formula of the narrowband E-model and uses two interactivity parameters to change the impact of transmission delay on the impairment factor. Both models are evaluated on empirical data.

In order to predict the changes in overall conversational quality in conversations affected by bursty packet loss, the formula of the effective equipment impairment factor of the E-model was extended. Here, an addition of the burst ratio and a burstiness robustness factor specific to each codec is used to model the changes in quality ratings. The extension is validated with predictions of the POLQA model, as well as with empirical conversations from a conversation test.

Finally, the simulation approach was validated by predicting the conversational quality with the interactivity model, as well as the extended E-model, based on simulated conversations. An evaluation of simulated conversations with a delay has shown that

both the interactivity model as well as the delay-extended E-model are able to replicate good prediction accuracy based on simulations. In a final evaluation of the simulation, the quality of conversations impaired with overall delay and bursty packet loss was predicted with the extended E-model. It was shown that the simulation is able to replicate the interaction of the changes in interactivity from bursty packet loss with the impaired turn-taking due to delay. Thus, the resulting prediction based on the simulation outperforms the classical prediction based solely on the parameters of the transmission.

Chapter 7
Conclusions and Future Work

The quality of conversations in remote VoIP contexts is – now more than ever – an important research topic. In such real-time communication scenarios, degradations like delay and bursty packet loss impact the way we interact with each other and thus degrade the interactivity of the conversation. Current models that predict speech quality either do not consider the interactivity or include them as abstract parameters that require extensive subjective testing before they can be applied. Thus, in this thesis, the concept of simulating conversations for the prediction of transmission quality is explored. The simulation of conversations allows for modeling the impact of degradations on the conversation level itself, which is beneficial for non-audible impairments like pure delay.

This approach to speech quality prediction is completely novel, and thus, the work presented in this thesis was split into two separate research objectives: the general feasibility and architecture of simulating a conversation between two humans and the application of the simulation approach to the prediction of conversational quality. As part of these two objectives, five research questions were answered, each of which is a substantial extension of current research in the domain of speech quality, but also useful in the field of spoken dialogue systems. In the following, the answers to these research questions will be summed up and discussed.

Question 1: *How can the models and methods of dialogue and user simulation from the area of Spoken Dialogue Systems be applied to the simulation of conversations between two humans?*

In order to simulate the content of a conversation, as well as the interactions between interlocutors, a simulation needs to be performed on the signal, text, and dialogue act level. In order to model the turn-taking in conversations, the simulation needs to implement the incremental nature of human speech and dialogue processing.

In Chapter 3, a simulation architecture was presented that utilizes an incremental dialogue processing framework specially developed for the simulation. The architecture uses training data from previously recorded conversations to model the dialogue based on a goal-oriented agenda dialogue manager. The agents in the simulation communicate through a simulated VoIP network that is able to insert impairments like delay and

packet loss. An evaluation of the simulation framework has shown that the ITU-T standardized SCT and RNV conversations, which exhibit distinct levels of interactivity, can be modeled on the semantic level. An evaluation of the interaction parameter has shown that a simulation in turn-steps, while showing some difference in the two conversation scenarios, is insufficient to accurately model the differences in turn-taking interactivity of everyday remote conversations.

Question 2: *How can the smooth taking of turns in natural VoIP conversations of different levels of interactivity be replicated in a simulation?*

Modeling turn-transitions and turn-continuations as two competing processes can implement turn-taking in a conversation simulation. The currently speaking agent is determining when to continue talking based on a probabilistic distribution, and the listening agent is determining when to start talking relative to the end of the current utterance. The distribution that both agents decide their turn-taking behavior on is selected based on the dialogue act that is currently uttered.

As a simulation in turn steps is not able to replicate the interactivity of a conversation, Chapter 4 introduces a turn-taking mechanism that can be independently employed in the two agents of the simulated conversation. The underlying model of the turn-taking mechanism is based on the distribution of turn-transitions, as modeled by gaps and overlaps between speaker turns, and turn-continuations, as modeled by pauses in between the turns of the same speaker. For this, the turn-taking behavior of real SCT and RNV conversations are analyzed, and, based on the resulting distributions, multiple transitioning and continuation models are defined. In the simulation, the agents then decide which distribution to use, as determined by the type of dialogue act the active agent is producing. Finally, to determine the beginning of their next turn relative to the current turn's end, the agents randomly select values from the appropriate cumulative distribution function. The evaluation of the turn-taking mechanism showed that it enables the simulation to reproduce the differences in turn-taking between conversations of different conversational interactivities in terms of pauses, double talk, speaker alternation rate, and length of the conversation. The turn-taking and conversation structure of the two conversation scenarios arise naturally from the implemented turn-taking mechanism and do not need to be modeled explicitly.

Question 3: *How is turn-taking affected by transmission delay, and what rules and models can be employed to replicate these changes in a simulation?*

An analysis of conversations through transmissions with transmission delay has shown that humans mostly change their turn-continuation behavior when they detect that transmission delay is present. In order to model this, the turn-taking mechanism is extended by dampening the turn-continuation probability by a constant factor every time an agent is interrupted unintendedly.

When introducing transmission delay into the simulation, the resulting conversations show degradations in the interaction due to the naturally slower arrival of speech signals. However, when comparing these simulations with empirical data, not all changes can be modeled by applying the turn-taking mechanism modeled without transmission delay in a delayed conversation. An analysis of the previously recorded SCT and RNV conversation degraded by transmission delay showed that in conversations with high delay levels, the participants changed the behavior of their turn-continuations. Specifically, with increased delay levels, the currently active speaker made significantly longer pauses, assumably for the delayed speech signal of their interlocutor to arrive. Because the participants of a conversation with transmission delay do not have knowledge about the exact amount of transmission delay, the changes in turn-taking behavior were modeled by dampening the turn-continuation model by a constant factor each time an unwanted interruption was detected. A final evaluation of the turn-taking with transmission delay showed that the simulation is able to model the different reactions to transmission delay of the SCT and RNV conversation scenarios in terms of speaker alternation rate and unintended interruptions.

Question 4: *What impact has bursty packet loss on the understandability of speech in a conversation, and how can it be modeled in a simulation?*

The occurrences of misunderstandings due to packet loss were made measurable by formalizing *conversation disruptions* and measuring them in conversations affected by bursts of lost packets. It was shown that the conversation disruption rate is different for the two conversational scenarios SCT and RNV. Furthermore, it was shown that the conversation disruptions as a result of bursty packet loss have an impact on the interactivity of RNV conversation, as measured by the speaker alternation rate. Finally, the conversation disruptions were modeled on an utterance level, and changes in turn-taking were implemented to reflect the changes in interactivity.

In Chapter 5, the interactivity of conversations with highly bursty packet loss was analyzed. It was shown that conversations with high interactivity (i.e., RNV conversations) reduce interactivity with increasing levels of bursty packet loss. The term *conversation disruption* was defined, which makes misunderstandings in a conversation, as an effect of the loss of speech signal, measurable. It was shown that the rate of conversation disruptions is higher for RNV conversations than for SCT conversations, but also that the number of disruptions per turn is consistent between the two conversation types. It was also shown that in conversations with high packet loss, the turn-taking behavior of the participants changed. Specifically, the turn-transitions happen more slowly, with fewer overlaps between the speaking turns. Finally, the occurrences of conversation disruptions were implemented into the simulation that models conversation disruption based on the percentage of lost speech in an utterance. With the new behavior implemented, the turn-transitioning model was dampened by a constant factor for every conversation disruption that occurred. The evaluation of the new simulation behavior showed that

the resulting conversation model simulated the number of disruptions, the frequency of disruptions, as well as the changes in turn-taking and conversational interactivity quite well.

Question 5: *How well can conversational parameters and the overall quality be predicted with this new approach?*

For the prediction of conversational quality based on the simulated conversations, the parametric E-model was extended to incorporate interactivity parameters in the delay calculation and burstiness in the packet loss formula. A new model to predict the quality of individual conversations based on interactivity parameters has been presented. Both the extended E-model and the interactivity model are trained and evaluated on empirical conversation data. An evaluation of the models with simulated conversations revealed that this new approach is able to replicate the changes due to delay and packet loss and thus, can be used to predict the overall quality. Especially with both delay and packet loss present, the prediction based on the simulation is significantly better than a transmission-parameter-based approach.

In Chapter 6, a model that predicts the conversational quality based on interactivity parameters of the conversation was presented and evaluated. Also, two extensions of the fullband E-model were proposed. First, the extension of the Idd delay formula is based on the equation of the narrowband E-model and adds the two interactivity parameters for the minimal perceivable delay and delay sensitivity. These parameters can be adjusted depending on the expected interactivity of the conversations carried out with the transmission system under study. The second extension of the E-model concerns the Ie, eff, FB formula, which captures impairments related to coding. Here, the burst ratio $BurstR$ is added to the equation, together with a burstiness robustness factor. The interactivity model, as well as the E-model extensions, are validated with conversations recorded in laboratory experiments. Finally, the two models are used on simulated conversations with delay and a combination of delay and packet loss. The evaluation shows that the simulated conversations with transmission delay improve the parametric prediction of the E-model by providing more accurate interactivity parameters. When bursty packet loss is added in the simulation, the effects on the interactivity can be extracted, and the E-model predictions further improve over a prediction without simulations.

One major drawback of the current approach to simulating a conversation is the limited variance compared to empirical conversations. This smaller range of characteristics of the simulation is the result of the small amount of training data that is used for the simulation but is also influenced by the way an interlocutor is simulated. While the overall 30 annotated conversations of each scenario are enough to replicate the interactivity parameters of the conversations, it is not enough to replicate the variance of interactions and resulting quality predictions. The simulation approach described here models the interlocutors of each conversation with the same set of behavioral characteristics. Properties of different speakers like variations in turn-taking behavior,

speech speeds and styles, and even accents and choice of vocabulary can influence the resulting conversation. As the agents of the simulation replicate the distribution of behaviors seen in the datasets, the resulting behavior reflects an average interlocutor that might not exist in the training data. Increasing the training data alone might not be sufficient for resolving this issue and modeling of distinct speaker behavior is needed. Generally, the amount of data and type of modeling presented in this thesis results in simulated conversations that are able to reproduce the main parameters of impairments of current speech communication systems.

In conclusion, the simulation of conversations described in this thesis has proven to model key features of VoIP communication and is able to replicate changes in the conversation due to the common impairments delay and packet loss. It is a universal scientific tool that can be used especially for the prediction of conversational quality.

7.1 Future Work

The simulation of human-to-human conversations presented in this thesis is the first of its kind and a step towards a universally applicable model for the prediction of speech quality, as well as a possible tool for other areas of research. Nevertheless, the presented approach leaves room for future work, which can be separated into three different directions. First, the presented simulation framework should be improved and validated. Second, the simulation can be extended for additional impairments and for use in quality monitoring. Finally, the new approach can be extended for use in other scientific fields.

The primary validation of the system and its architecture has been provided in this thesis. For this first modeling step, the underlying data and the chosen levels of delay and packet loss impairments have been very strong. Additional validation of the simulation with more moderate levels is thus advised. While in over-the-top VoIP communication platforms, the transmission delay is generally higher than in classical telephone systems, delay levels as high as $1800\,ms$ are not to be expected in everyday scenarios. However, especially for services that require large jitter buffers, the modeling of higher levels of overall delay is a useful scenario. Also, packet loss levels of $15\,\%$ and above are uncommon in everyday speech communication. Still, packet loss with strong bursts tends to be realistic, as network problems often affect more than one packet. Especially for the trade-off between the size of the jitter buffer (thus, implicitly, the overall delay) and the probability of packet loss can be evaluated with this approach. Especially the effect of burstiness on the likelihood of conversation disruptions needs to be investigated further.

A possible extension for the simulation approach is the modeling of additional impairments. For example, the system could be used to simulate the behavior of the listener and talker echo on the conversation. Here, phenomena like slower speaking rates or Lombard speech can be simulated and used for the prediction of perceived quality. Another speech quality-related extension may be to utilize other quality models than the E-model. For example, the double-ended speech quality prediction model POLQA might be used on the degraded and clean signal of the simulated conversation to estimate the listening quality. The current simulation architecture allows for the prediction of conversational quality for planning purposes. By extending the simulation to transmit

the conversational speech over a real transmission network, it could also be used for quality monitoring purposes. To better replicate the variance of parameters and quality ratings of real conversations, the simulation needs to model the difference in individual conversations rather than trying to simulate an average interlocutor. For the simulation of different speaking styles, choice of dialogue acts and selection of vocabulary-specific communication profiles may be created. The difference in communication and the interaction between the communication profiles could then replicate the greater range of variances seen in the empirical data.

The third area of future work might be the use of the simulation approach in other areas of research. As the simulation tries to replicate the behavior of humans in conversational scenarios, it can be used as a tool to model and validate spoken dialogue systems. Single incremental modules of the simulation may be replaced with models under study, or a simulation agent may be replaced with a dialogue system. This would allow for fast and cost-effective training and evaluation of dialogue systems and their parts. Also, for use in phonetics and linguistics, the simulation approach may be a useful foundation. For example, different conversational phenomena like Lombard speech during conversations with background noise, hesitations, changes in articulation, or prosody can be modeled, and interactions between these characteristics can be examined.

Finally, the work presented here is discussed in the Study Group 12 of the International Telecommunications Union ITU-T. There, a working item P.CONVSIM has been established that focuses on standardizing the architecture and models used in this simulation approach.

Appendix A
Short Conversation Test (SCT)

Scenario 3: Information on flights

Your name: Matthew *Parker*

Intended journey: Sydney → Abu Dhabi

Date: June 23rd
Morning flight
Direct flight preferred

Departure : _____ h

Arrival : _____ h

Flight number : _____

Reservation : One seat
Economy Class

Address : 47 Rawson Street, Sydney
☎ 02955 0833

What are some tourist attractions in Abu Dhabi?

_____

Scenario 3: Information on flights

Your name: *Sydney flight information*

Flight schedule: Sydney → Abu Dhabi

		Etihad	British Airways	Qantas Airways
Flight number		E 615	BA 381	QF 413
Sydney	*Dep.*	6:30 h	6:35 h	8:20 h
Perth	*Arr.*		10:35 h	
Perth	*Dep.*		11:15 h	
Abu Dhabi	*Arr.*	21:35 h (*daily*)	21:55 h (*daily*)	23:25 h (*daily*)

Reservation:

Name : _____

Address : _____

: _____

Telephone number : _____

Number of seats : _____

Class ☐ Business ☐ Economy

_____

Scenario 11: Pizza service

Your name: Jeremy *Clemence*

One large pizza

For 2 people
Vegetarian pizza preferred

Topping : _____

_____

Price : _____ A$

Delivery to : 109 George Street
Adelaide
☎: 08212 7320

How long will we have to wait for the pizza to be delivered?

_____

Scenario 11: Pizza service

Your name: *Pizzeria Roma*

Pizza

Pizzas	1 person	2 persons	4 persons
Toscana (ham, mushrooms, tomatoes, cheese)	A$ 12.20	A$ 14.95	A$ 19.50
Tonno (tuna, onions, tomatoes, cheese)	A$ 12.95	A$ 15.50	A$ 22.95
Fabrizio (salami, ham, tomatoes, cheese)	A$ 13.20	A$ 15.95	A$ 23.95
Vegetarian (spinach, mushrooms, tomatoes, cheese)	A$ 13.50	A$ 16.50	A$ 24.95

Delivery to: Name : _____

Address : _____

_____

Telephone number : _____

_____

Appendix B
Random Number Verification (RNV) Task

Example scenarios for random number verification tasks

Instructions: "Your conversation partner is also provided with such a list. Some of the numbers in your list do not correspond with those of your conversation partner. Find the wrong numbers as quickly as possible by taking turns reading them line by line. Acknowledge by saying "yes" or "no", and cross out the wrong numbers. You will read the bold red numbers and your conversation partner will read the non-bold blue ones".

You are the "caller".

18	**88**	**80**	**74**	**55**	**7**
15	29	14	37	17	82
20	**95**	**36**	**77**	**34**	**83**
46	84	30	67	25	99
28	**27**	**36**	**96**	**60**	**97**
55	10	87	53	43	98

Instructions: "Your conversation partner is also provided with such a list. Some of the numbers in your list do not correspond with those of your conversation partner. Find the wrong numbers as quickly as possible by taking turns reading them line by line. Acknowledge by saying "yes" or "no", and cross out the wrong numbers. You will read the bold blue numbers and your conversation partner will read the non-bold red ones".

Please wait to be called.

18	84	80	74	55	7
15	**29**	**14**	**67**	**17**	**82**
36	95	36	77	53	83
46	**88**	**30**	**37**	**25**	**99**
28	27	20	96	60	97
55	**10**	**87**	**34**	**43**	**98**

Appendix C
Agenda of Simulated Agents

Shown here are the `.ini`-files used as the agenda for the simulation of the SCT scenario 11 and the RNV scenario 1. The information provided in these agenda files correspond to the information given in the short conversation test scenarios shown in Appendix A. The information is split into categories denoted by square brackets (e.g., `[General]`). For each category, information has to be either requested from the interlocutor (when only the name of the information variable is given) or information has to be given to the interlocutor (when the information variable is set to a specific value). If an information is split over multiple lines (e.g., the telephone number), it may (but does not have to) be split over multiple turns.

```
[General]
callee_name=Pizzeria Roma

[Reason]
reason

[Additional]
num_of_persons
pizza_type

[Offer]
pizza_name=Pizza Vegetaria

[CalleeInformation]
toppings=spinach
  mushrooms
  tomatoes
  cheese
price=17 Euro

[CallerInformation]
caller_name
address
telephone

[Improv]
delivery_duration=<improvised>
```

Listing C.1 Configuration file of the agenda of the callee in the SCT 11 scenario (Pizzeria Roma).

```
[General]
callee_name

[Reason]
reason=1 large pizza

[Additional]
num_of_persons=2
pizza_type=vegetarian

[Offer]
pizza_name

[CalleeInformation]
toppings
price

[CallerInformation]
caller_name=Jeremy Clemens
address=Gluecksburger Str.
  41
  Bochum
telephone=0
  8
  1
  1
  7
  3
  4
  2
  0

[Improv]
delivery_duration
```

Listing C.2 Configuration file of the agenda of the caller in the SCT 11 scenario (Jeremy Clemens).

```
[Numbers]
string0
string1=41
  7
  86
  24
  56
  38
string2
string3=17
  56
  76
  20
  77
  34
```

Listing C.3 Configuration file of the agenda of the callee in the RNV 1 scenario.

```
[Numbers]
string0=31
  85
  17
  73
  44
  59
string1
string2=11
  81
  85
  36
  37
  78
string3
```

Listing C.4 Configuration file of the agenda of the caller in the RNV 1 scenario.

References

3GPP Technical Specification 26.071 (1999). *Mandatory speech codec speech processing functions; AMR speech Codec; General description*. 3GPP, Sophia Antipolis Valbonne, France.

3GPP Technical Specification 26.171 (2001). *Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; General description*. 3GPP, Sophia Antipolis Valbonne, France.

3GPP Technical Specification 26.441 (2014). *Codec for Enhanced Voice Services (EVS); General overview*. 3GPP, Sophia Antipolis Valbonne, France.

American National Standards Institute (1997). *American National Standard: Methods for Calculation of the Speech Intelligibility Index*. Acoustical Society of America.

Appel, R. and Beerends, J. G. (2002). On the quality of hearing one's own voice. *Journal of the Audio Engineering Society*, 50(4):237–248.

Baumann, T. (2008). Simulating Spoken Dialogue With A Focus on Realistic Turn-Taking. *13th ESSLLI Student Session*, pages 17–25.

Baumann, T., Buß, O., and Schlangen, D. (2010). InproTK in action: Open-source software for building german-speaking incremental spoken dialogue systems. In *21. Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*. TUDpress, Dresden.

Baumann, T. and Schlangen, D. (2012a). INPRO_iSS: A component for just-in-time incremental speech synthesis. In *Proceedings of the ACL 2012 System Demonstrations*, pages 103–108.

Baumann, T. and Schlangen, D. (2012b). The INPROTK 2012 release. In *NAACL-HLT Workshop on Future Directions and Needs in the Spoken Dialog Community: Tools and Data*, pages 29–32. Association for Computational Linguistic.

Benoît, C., Grice, M., and Hazan, V. (1996). The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. *Speech communication*, 18(4):381–392.

Bodden, M. and Jekosch, U. (1996). Entwicklung und Durchführung von Tests mit Versuchspersonen zur Verifizierung von Modellen zur Berechnung der Sprachübertragungsqualität. *Final report to a project funded by Deutsche Telekom AG (unpublished), Institut für Kommunikationsakustik, Ruhr Universität, Bochum*.

Brady, P. T. (1968). A statistical analysis of on-off patterns in 16 conversations. *Bell System Technical Journal*, 47(1):73–91.

Cavanaugh, J. R., Hatch, R. W., and Sullivan, J. L. (1976). Models for the subjective effects of loss, noise, and talker echo on telephone connections. *Bell System Technical Journal*, 55(9):1319–1371.

Côté, N., Gautier-Turbin, V., and Möller, S. (2007). Influence of loudness level on the overall quality of transmitted speech. In *Proceedings of the 123rd Convention of the Audio Engineering Society*, New York, NY, USA.

Cuayáhuitl, H., Renals, S., Lemon, O., and Shimodaira, H. (2005). Human-computer dialogue simulation using hidden markov models. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 290–295. IEEE.

Eckert, W., Levin, E., and Pieraccini, R. (1997). User modeling for spoken dialogue system evaluation. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 80–87. IEEE.

Egger, S., Schatz, R., and Scherer, S. (2010). It takes two to tango-assessing the impact of delay on conversational interactivity on perceived speech quality. In *Prcoeedings of INTERSPEECH 2010*, pages 1321–1324. ISCA.

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

Ekstedt, E. and Skantze, G. (2020). Turngpt: a transformer-based language model for predicting turn-taking in spoken dialog. *arXiv preprint arXiv:2010.10874*.

Engelbrecht, K.-P., Quade, M., and Möller, S. (2009). Analysis of a new simulation approach to dialog system evaluation. *Speech Communication*, 51(12):1234–1252.

Ferrer, L., Shriberg, E., and Stolcke, A. (2002). Is the speaker done yet? faster and more accurate end-of-utterance detection using prosody. In *Seventh International Conference on Spoken Language Processing*.

Fischer, K., Naik, L., Langedijk, R. M., Baumann, T., Jelínek, M., and Palinko, O. (2021). Initiating human-robot interactions using incremental speech adaptation. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '21 Companion, page 421–425, New York, USA. Association for Computing Machinery.

Ford, C. E. and Thompson, S. A. (1996). *Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns*, page 134–184. Studies in Interactional Sociolinguistics. Cambridge University Press.

Guéguin, M., Le Bouquin-Jeannès, R., Faucon, G., and Barriac, V. (2006). Towards an objective model of the conversational speech quality. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I.

Guéguin, M., Le Bouquin-Jeannès, R., Gautier-Turbin, V., Faucon, G., and 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*. Forschungszentrum Telekommunikation Wien.

Hammer, F., Reichl, P., and Raake, A. (2004). Elements of Interactivity in Telephone Conversations. In *Proc. Interspeech 2004*, pages 1741–1744, Jeju Island, Korea.

Hammer, F., Reichl, P., and Raake, A. (2005). The well-tempered conversation: Interactivity, Delay and Perceptual VoIP Quality. In *IEEE International Conference on Communications*, volume 1, pages 244–249. Institute of Electrical and Electronics Engineers (IEEE).

Heeman, P. A. and Lunsford, R. (2017). Turn-taking offsets and dialogue context. In *Proc. Interspeech 2017*, pages 1671–1675.

Heldner, M., Edlund, J., Hjalmarsson, A., and Laskowski, K. (2011). Very short utterances and timing in turn-taking. In *Proceedings of INTERSPEECH 2011*. Citeseer.

Hillmann, S. (2017). *Simulation-Based Usability Evaluation of Spoken and Multimodal Dialogue Systems*. Springer.

Hillmann, S. and Engelbrecht, K.-P. (2015). Modelling goal modifications in user simulation. In *International Workshop on Future and Emergent Trends in Language Technology*, pages 149–159. Springer.

Hornsby, B. (2004). The Speech Intelligibility Index: What is it and what's it good for? *The Hearing Journal*, 57(10):10–17.

Huo, L. (2015). *Attribute-based Speech Quality Assessment-Narrowband and Wideband*. Shaker-Verlag.

ITU-T Contribution SG12-C35-E (1997). Development of scenarios for short a conversation test. Geneva: Internation Telecommunication Union.

ITU-T Recommendation G.107 (2015). *The E-model: a computational model for use in transmission planning*. International Telecommunication Union, Geneva, Switzerland.

ITU-T Recommendation G.107.1 (2015). *Wideband E-model*. International Telecommunication Union, Geneva, Switzerland.

ITU-T Recommendation G.107.2 (2019). *Fullband E-model*. International Telecommunication Union, Geneva, Switzerland.

ITU-T Recommendation G.113 (2007). *Transmission impairments due to speech processing*. International Telecommunication Union, Geneva, Switzerland.

ITU-T Recommendation G.113 Amendment 1 (2009). *Revised Appendix IV – Provisional planning values for the wideband equipment impairment factor and the wideband packet loss robustness factor*. International Telecommunication Union, Geneva, Switzerland.

ITU-T Recommendation G.113 Amendment 2 (2019). *New Appendix V – Provisional planning values for the fullband equipment impairment factor and the fullband packet loss robustness factor*. International Telecommunication Union, Geneva, Switzerland.

ITU-T Recommendation G.114 (2003). *One-way transmission time*. International Telecommunication Union, Geneva, Switzerland.

ITU-T Recommendation G.711 (1988). *Pulse Code Modulation (PCM) of Voice Frequencies*. International Telecommunication Union, Geneva, Switzerland.

ITU-T Recommendation G.722 (2012). *7 kHz audio-coding within 64 kbit/s*. International Telecommunication Union, Geneva, Switzerland.

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

ITU-T Recommendation P.10 (2017). *Vocabulary for performance, quality of service and quality of experience*. International Telecommunication Union, Geneva, Switzerland.

ITU-T Recommendation P.50 (1999). *Artificial Voices*. International Telecommunication Union, Geneva, Switzerland.

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

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

ITU-T Recommendation P.59 (1993). *Artificial Conversational Speech*. International Telecommunication Union, Geneva, Switzerland.

ITU-T Recommendation P.800 (1996). *Methods for subjective determination of transmission quality*. International Telecommunication Union, Geneva, Switzerland.

ITU-T Recommendation P.800.1 (2016). *Mean opinion score (MOS) terminology*. International Telecommunication Union, Geneva, Switzerland.

ITU-T Recommendation P.804 (2017). *Subjective diagnostic test method for conversational speech quality analysis*. International Telecommunication Union, Geneva, Switzerland.

ITU-T Recommendation P.805 (2007). *Subjective Evaluation of Conversational Quality*. International Telecommunication Union, Geneva, Switzerland.

ITU-T Recommendation P.806 (2014). *A subjective quality test methodology using multiple rating scales*. International Telecommunication Union, Geneva, Switzerland.

ITU-T Recommendation P.808 (2021). *Subjective evaluation of speech quality with a crowdsourcing approach*. International Telecommunication Union, Geneva, Switzerland.

ITU-T Recommendation P.862 (2001). *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, Geneva, Switzerland.

ITU-T Recommendation P.863 (2014). *Perceptual objective listening quality assessment*. International Telecommunication Union, Geneva, Switzerland.

Janarthanam, S. and Lemon, O. (2009). A two-tier user simulation model for reinforcement learning of adaptive referring expression generation policies. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 120–123. Association for Computational Linguistics.

Jekosch, U. (2005). *Voice and speech quality perception: assessment and evaluation*. Springer.

Johannesson, N. O. (1997). The ETSI computation model: A tool for transmission planning of telephone networks. *IEEE Communications Magazine*, 35(1):70–79.

Jokinen, K. and McTear, M. (2009). Spoken dialogue systems. *Synthesis Lectures on Human Language Technologies*, 2(1):1–151.

Jurafsky, D. and Martin, J. H. (2009). *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J.

Kawahara, T., Yamaguchi, T., Inoue, K., Takanashi, K., and Ward, N. (2016). Prediction and Generation of Backchannel Form for Attentive Listening Systems. In *Proc. Interspeech 2016*, pages 2890–2894.

Kennington, C., Kousidis, S., and Schlangen, D. (2014). Inprotks: A toolkit for incremental situated processing. *Proceedings of SIGdial 2014: Short Papers*.

Kennington, C., Moro, D., Marchand, L., Carns, J., and McNeill, D. (2020). rrsds: Towards a robot-ready spoken dialogue system. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 132–135.

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

Köster, F. (2018). *Multidimensional Analysis of Conversational Telephone Speech*. Springer.

Köster, F., Guse, D., Wältermann, M., and Möller, S. (2015). Comparison between the discrete ACR scale and an extended continuous scale for the quality assessment of transmitted speech. *Fortschritte der Akustik-DAGA*.

Köster, F. and Möller, S. (2014). Analyzing perceptual dimensions of conversational speech quality. In *Proceedings of INTERSPEECH*, pages 2041–2045, Singapore, Singapore.

Köster, F. and Möller, S. (2015). Perceptual speech quality dimensions in a conversational situation. In *Proceedings of INTERSPEECH*, pages 2544–2548, Dresden, Germany.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Lecomte, J., Vaillancourt, T., Bruhn, S., Sung, H., Peng, K., Kikuiri, K., Wang, B., Subasingha, S., and Faure, J. (2015). Packet-loss concealment technology advances in EVS. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5708–5712. IEEE.

Lee, H. and Un, C. (1986). A study of on-off characteristics of conversational speech. *IEEE Transactions on Communications*, 34(6):630–637.

Liu, C., Ishi, C., and Ishiguro, H. (2017). Turn-Taking Estimation Model Based on Joint Embedding of Lexical and Prosodic Contents. In *Proceedings of INTERSPEECH 2017*, pages 1686–1690.

Lunsford, R., Heeman, P. A., and Rennie, E. (2016). Measuring turn-taking offsets in human-human dialogues. In *Proceedings of INTERSPEECH 2016*, pages 2895–2899, San Francisco, U.S.A.

Michael, T. (2020). Retico: An incremental framework for spoken dialogue systems. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 49–52.

Michael, T. and Ibrahim, O. (2022). Lexical frequency and listener's response to packet loss in telephone conversations. In *33. Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, pages 74–80. TUDpress, Dresden.

Michael, T., Mittag, G., Bütow, A., and Möller, S. (2021). Extending the Fullband E-Model Towards Background Noise, Bursty Packet Loss, and Conversational Degradations. In *Proc. Interspeech 2021*, pages 2391–2395.

Michael, T., Mittag, G., and Möller, S. (2020). Analyzing the fullband e-model and extending it for predicting bursty packet loss. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6.

Michael, T. and Möller, S. (2018). Simulating Human-to-Human Conversations for the Prediction of Conversational Quality. *Fortschritte der Akustik-DAGA*.

Michael, T. and Möller, S. (2019). Retico: An open-source framework for modeling real-time conversations in spoken dialogue systems. In *30. Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, pages 134–140. TUDpress, Dresden.

Michael, T. and Möller, S. (2020a). Effects of Delay and Packet-Loss on the Conversational Quality. *Fortschritte der Akustik-DAGA*, pages 945–948.

Michael, T. and Möller, S. (2020b). Interactivity-based Quality Prediction of Conversations with Transmission Delay. In *Proceedings of the 22nd International Conference SPECOM*, pages 336–345.

Michael, T. and Möller, S. (2020c). Simulating Turn-Taking in Conversations with Delayed Transmission. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 157–161.

Michael, T. and Möller, S. (2020d). Simulating Turn-Taking in Conversations with varying Interactivity. In *31. Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, pages 93–100. TUDpress, Dresden.

Michael, T. and Möller, S. (2021). Predicting conversational quality from simulated conversations with transmission delay. In *14th ITG Conference on Speech Communication*. VDE.

Michael, T. a. (2021). Intelligibility in Telephone Conversations with Packet Loss. In *32. Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, pages 311–318. TUDpress, Dresden.

Mittag, G. (2022). *Machine Learning Based Speech Quality Prediction*. Springer International Publishing.

Mittag, G., Möller, S., Barriac, V., and Ragot, S. (2018). Quantifying quality degradation of the EVS super-wideband speech codec. In *Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE.

Mittag, G. and Möller, S. (2020). Full-reference speech quality estimation with attentional siamese neural networks. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 346–350.

Mittag, G., Naderi, B., Chehadi, A., and Möller, S. (2021). Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. *arXiv preprint arXiv:2104.09494*.

Möller, S. (2004). *Quality of telephone-based spoken dialogue systems*. Springer Science & Business Media.

Möller, S., Köster, F., and Weiss, B. (2017). Modelling speech service quality: From conversational phases to communication quality and service quality. In *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE.

Möller, S., Mittag, G., Michael, T., Barriac, V., and Aoki, H. (2019). Extending the e-model towards super-wideband and fullband speech communication scenarios. In *Proceedings of INTERSPEECH 2019*, pages 3436–3440, Graz, Austria.

Möller, S., Raake, A., Kitawaki, N., Takahashi, A., and Wältermann, M. (2006). Impairment factor framework for wide-band speech codecs. *IEEE Transactions on Audio, Speech and Language Processing*, 14(6):1969–1976.

Möller, S. (2000). *Assessment and prediction of speech quality in telecommunications*. Kluwer Academic Publishers.

Möller, S. and Skowronek, J. (2004). An analysis of quality prediction models for telephone-based spoken dialogue systems. *Acta Acustica united with Acustica*, 90:1112–1130.

Niebuhr, O., Görs, K., and Graupe, E. (2013). Speech reduction, intensity, and f0 shape are cues to turn-taking. In *Proceedings of the SIGDIAL 2013 Conference*, pages 261–269.

Osaka, N. and Kakehi, K. (1986). Objective evaluation model of telephone transmission performance for fundamental transmission factors. *Electronics and Communications in Japan (Part I: Communications)*, 69(2):18–27.

Padilha, E. G. (2006). *Modelling turn-taking in a simulation of small group discussion*. PhD thesis, University of Edinburgh.

Pietquin, O. and Hastie, H. (2013). A survey on metrics for the evaluation of user simulations. *The knowledge engineering review*, 28(1):59–73.

Raake, A. (2006). Short-and long-term packet loss behavior: towards speech quality prediction for arbitrary loss distributions. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):1957–1968.

Raake, A., Möller, S., Wältermann, M., Cote, N., and Ramirez, J.-P. (2010). Parameter-based prediction of speech quality in listening context—Towards a WB E-model. In *2010 Second International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 182–187. IEEE.

Raake, A., Schoenenberg, K., Skowronek, J., and Egger, S. (2013). Predicting speech quality based on interactivity and delay. In *Proceedings of INTERSPEECH*, pages 1384–1388, Lyon, France.

Rafla, A. and Kennington, C. (2019). Incrementalizing rasa's open-source natural language understanding pipeline. *arXiv preprint arXiv:1907.05403*.

Reichl, P. and Hammer, F. (2004). Hot discussion or frosty dialogue? Towards a temperature metric for conversational interactivity. In *Eighth International Conference on Spoken Language Processing*.

RFC 6716 (2012). *Definition of the Opus Audio Codec*. Internet Engineering Task Force (IETF), Fremont, CA, USA.

Sacks, H., Schegloff, E., and Jefferson, G. (1974). A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, 50(4):696–735.

Schatzmann, J., Thomson, B., Weilhammer, K., Ye, H., and Young, S. (2007). Agenda-based user simulation for bootstrapping a pomdp dialogue system. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 149–152. Association for Computational Linguistics.

Schatzmann, J., Weilhammer, K., Stuttle, M., and Young, S. (2006). A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *The knowledge engineering review*, 21(2):97–126.

Schatzmann, J. and Young, S. (2009). The hidden agenda user simulation model. *IEEE Transactions on Audio, Speech, and Language Processing*, 4(17):733–747.

Scheffler, T., Roller, R., and Reithinger, N. (2009). Speecheval – evaluating spoken dialog systems by user simulation. In *Proceedings of the 6th IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, pages 93–98, Pasadena.

Schlangen, D. and Skantze, G. (2009). A general, abstract model of incremental dialogue processing. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 710–718. Association for Computational Linguistics.

Schlangen, D. and Skantze, G. (2011). A general, abstract model of incremental dialogue processing. *Dialogue and Discourse*, 2(1):83–111.

Schoenenberg, K. (2015). *The Quality of Mediated-Conversations under Transmission Delay*. PhD thesis, TU Berlin.

Scholz, K. (2008). *Instrumentelle Qualitätsbeurteilung von Telefonbandsprache beruhend auf Qualitätsattributen*. Shaker.

Selfridge, E. O. and Heeman, P. A. (2012). A temporal simulator for developing turn-taking methods for spoken dialogue systems. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 113–117. Association for Computational Linguistics.

Sen, D. and Lu, W. (2012). Objective evaluation of speech signal quality by the prediction of multiple foreground diagnostic acceptability measure attributes. *The Journal of the Acoustical Society of America*, 131(5):4087–4103.

Siegert, I., Sinha, Y., Jokisch, O., and Wendemuth, A. (2020). Recognition performance of selected speech recognition apis–a longitudinal study. In *International Conference on Speech and Computer*, pages 520–529. Springer.

Skantze, G. (2017). Towards a General, Continuous Model of Turn-taking in Spoken Dialogue using LSTM Recurrent Neural Networks. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 220–230.

Uhrig, S., Michael, T., Möller, S., Keller, P. E., and Voigt-Antons, J.-N. (2018). Effects of delay on perceived quality, behavior and oscillatory brain activity in dyadic telephone conversations. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE.

W3C Recommendation WebRTC (2021). *WebRTC 1.0: Real-Time Communication Between Browsers*. World Wide Web Consortium, Cambridge, MA, USA.

Wältermann, M. (2012). *Dimension-based quality modeling of transmitted speech*. Springer.

Wältermann, M., Raake, A., and Möller, S. (2010). Quality dimensions of narrowband and wideband speech transmission. *Acta Acustica united with Acustica*, 96(6):1090–1103.