

Dissertation

Quality Aspects of Packet-Based Interactive Speech Communication

ausgeführt zum Zwecke der Erlangung des akademischen Grades
eines Doktors der technischen Wissenschaften

eingereicht an der
Technischen Universität Graz
Fakultät für Elektrotechnik und Informationstechnik

von
Dipl.-Ing. Florian Hammer

Wien, Juni 2006



Supervisor

Prof. Dr. Gernot Kubin

Signal Processing and Speech Communication Laboratory

University of Technology at Graz, Austria

Examiner

PD Dr.-Ing. Sebastian Möller

Deutsche Telekom Laboratories

Berlin, Germany

ftw. Dissertation Series

Florian Hammer

**Quality Aspects of Packet-Based
Interactive Speech Communication**



telecommunications research center vienna

This work was carried out with funding from **Kplus** in the ftw.
projects A0/B0/B1/N0/U0.

This thesis has been prepared using L^AT_EX.

August 2006

1. Auflage

Alle Rechte vorbehalten

Copyright © 2006 Florian Hammer

Herausgeber: Forschungszentrum Telekommunikation Wien

Printed in Austria

ISBN 3-902477-05-9

Abstract

Voice-over-Internet Protocol (VoIP) technology provides the transmission of speech over packet-based networks. The transition from circuit-switched to packet-switched networks introduces two major quality impairments: packet loss and end-to-end delay. This thesis shows that the incorporation of packets that were damaged by bit errors reduces the effective packet loss rate, and thus improves the speech quality as perceived by the user. Moreover, this thesis addresses the impact of transmission delay on conversational interactivity and on the perceived speech quality. In order to study the structure and interactivity of conversations, the framework of Parametric Conversation Analysis (P-CA) is introduced and three metrics for conversational interactivity are defined. The investigation of five conversation scenarios based on subjective quality tests has shown that only highly structured scenarios result in high conversational interactivity. The speaker alternation rate has turned out to represent a simple and efficient metric for conversational interactivity. Regarding the two-way speech quality, it was found that echo-less end-to-end delay up to half a second does not cause impairment, even for highly interactive tasks.

Kurzfassung

Voice-over-Internet Protocol (VoIP) Technologie unterstützt die Übertragung von Sprache über paketvermittelte Netzwerke. Der Übergang von leitungsvermittelten zu paketvermittelten Netzwerken führt zu zwei wichtigen Faktoren, die die Sprachqualität beeinträchtigen: Paketverluste und Ende-zu-Ende-Verzögerung. Diese Arbeit zeigt, dass die Verwendung von Sprachpaketen, die durch Bitfehler gestört wurden, die effektive Paketverlustrate verringert und damit die vom Benutzer wahrgenommene Sprachqualität verbessert. Weiters widmet sich diese Arbeit dem Einfluß der Übertragungsverzögerung auf die Konversationsinteraktivität und auf die vom Benutzer wahrgenommene Qualität. Um die Interaktivität von Gesprächen untersuchen zu können, wird ein Konzept für eine parametrische Konversationsanalyse vorgestellt und drei Metriken für Konversationsinteraktivität definiert. Die Untersuchung von fünf Konversationszenarien auf der Basis von subjektiven Qualitätstests hat gezeigt, dass nur stark strukturierte Szenarien zu Gesprächen mit hoher Konversationsinteraktivität führen. Die Sprecherwechselrate hat sich dabei als ein einfaches und effizientes Maß für Konversationsinteraktivität herausgestellt. Bezüglich der Sprachqualität wurde festgestellt, dass echofreie Übertragungsverzögerungen bis zu einer halben Sekunde sogar bei einem stark interaktiven Szenario keine Beeinträchtigung der Qualität darstellen.

Acknowledgements

My work was supported by a lot of people. Therefore, I thank

Prof. Gernot Kubin for his supervision and for the inspiring and constructive discussions, and PD Sebastian Möller for carefully co-supervising my thesis.

Peter Reichl for his his great deal of guidance, enthusiasm, inspiration, and encouragement, and Tomas Nordström for his guidance and support, especially during the first half of my thesis work.

Christoph Mecklenbräuker for his guidance during the early phase of my work.

My colleagues Joachim Wehinger, Thomas Zemen, Elke Michlmayr, Peter Fröhlich, Ed Schofield, Ivan Gojmerac, Driton Statovci, Thomas Ziegler, Eduard Hasenleithner and Horst Thiess for all kinds of discussions, help, encouragement, inspirations and entertainment.

My colleagues Alexander Raake and Ian Marsh for the inspiring and fruitful discussions and the collaborations.

Markus Kommenda and Horst Rode for providing such a friendly and flexible research environment.

James Moore for his spirit, guidance and advice.

My parents and sisters for their love.

Petra for her light, patience and understanding.

Contents

1	Introduction	15
1.1	Motivation	15
1.2	Thesis Overview and Contributions	17
1.3	End-to-End Delay	20
1.4	Speech Quality Measurement	24
2	Corrupted Speech Data Considered Useful	31
2.1	Introduction	31
2.2	Background	33
2.2.1	UDP-Lite	33
2.2.2	Adaptive Multi-Rate Speech Coding	35
2.2.3	Robust Header Compression	36
2.2.4	Speech Quality Evaluation	37
2.2.5	Related Work	39
2.3	Alternative Strategies for VoIP Transport	39
2.4	Simulations	41
2.4.1	Simulation Environment	41
2.4.2	Bit Error Model	42
2.5	Results and Discussion	43
2.5.1	Estimated Perceived Speech Quality vs. Bit Error Rate	43
2.5.2	Gender Dependency	45
2.5.3	PESQ vs. TOSQA	47
2.6	Summary	47
3	Modeling Conversational Interactivity	49
3.1	Introduction	49
3.2	Related Work	50
3.2.1	Definitions of Interactivity	50
3.2.2	Conversational Parameters	52
3.2.3	The Concept of Turn-Taking	53
3.2.4	Conversation Scenarios	53
3.3	Parametric Conversation Analysis	54
3.3.1	Conversation Model	54
3.3.2	Conversational Events	55

3.3.3	Impact of Transmission Delay on Conversational Structure . . .	55
3.4	Models for Conversational Interactivity	57
3.4.1	Speaker Alternation Rate	57
3.4.2	Conversational Temperature	58
3.4.3	Entropy Model	60
3.5	Experiment 1	62
3.5.1	Objective	62
3.5.2	Measurement Setup and Test Procedure	62
3.6	Experiment 2	66
3.6.1	Objective	66
3.6.2	Selection of Conversation Scenarios	66
3.6.3	Measurement Setup and Test Procedure	67
3.7	Results and Discussion	70
3.7.1	Comparison SCT vs. iSCT (no delay)	70
3.7.2	The Effect of Delay on iSCTs	74
3.7.3	Comparison of various Conversation Scenarios	77
3.8	Summary	84
4	Impact of Transmission Delay on Perceptual Speech Quality	85
4.1	Introduction	85
4.2	Related Work	86
4.3	Experiment 1	90
4.3.1	Objective	90
4.3.2	Measurement Setup and Test Procedure	90
4.4	Experiment 2	90
4.4.1	Objective	90
4.4.2	Measurement setup and test procedure	91
4.5	Results and Discussion	94
4.5.1	Quality Impairment using the iSCT scenario	94
4.5.2	Influence of Conversation Scenarios	95
4.6	Summary	98
5	Conclusions and Outlook	101
5.1	Conclusions	101
5.2	Outlook	103
A	Acronyms	105
B	Scenarios	107
B.1	Random Number Verification	108
B.2	Short Conversation Test	110
B.3	Interactive Short Conversation Test	112

B.4 Asymmetric Short Conversation Test	114
B.5 Free Conversation	116
C Algorithms	117
C.1 Conversational Temperature	117
C.2 Entropy Rate	118
D E-model Parameters	119
Bibliography	121

1 Introduction

1.1 Motivation

In 1995, when Voice over Internet Protocol (VoIP) technology was commercially introduced [95], nobody could foresee the success and popularity it has reached during the last ten years. Peer-to-peer service providers like Skype [101] provide a state-of-the-art software client of very good quality. PC-to-PC calls are free (provided that high-speed internet access with a flat rate including a sufficient amount of data transfer is available). Additionally, an interconnection to the Plain Old Telephone System (POTS) is available at low fares. Moreover, an increasing number of companies merge their telephone system into their data network using VoIP technology. A converged network is more cost-effective and easier to manage than two separate networks. However, the incorporation of a real-time communication system into a network that has primarily been designed for pure data transmission implies strong requirements with regard to quality-of-service (QoS) and security, only to mention two important issues. Yet, as opposed to the POTS and due to its packet-switched nature, VoIP requires a specific configuration of Quality of Service (QoS) parameters, and still causes impairments regarding the QoS as perceived by the end user. Originally, packet-switched networks were constructed for data transmission, hence the major requirement on the network was a reliable transmission, so no data would be lost. Therefore, data transmission protocols such as TCP (transmission control protocol [80]) assure that every data packet is received at the destination. If any packet is lost in the network, the source must repeat sending the lost packet until it is finally received. The transport reliability results in severe latency caused by the transport protocol. Thus, TCP cannot be used for real-time applications like VoIP. In VoIP, the requirement of low latency does not allow packets to be resent, thus, speech packets are sent in real-time using the user datagram protocol (UDP, [78]). Figure 1.1 outlines a VoIP system. The “heart” of the system is the IP backbone which provides connectivity over various distances. The access network facilitates the “last mile” connection, i.e., the connection between the backbone network and the end user. Access networks are implemented in either wireline or wireless technology. Wireline technology includes DSL (Digital Subscriber Line [13]) or cable access and wireless technology is represented by UMTS (Universal Mobile Telecommunication System [6, 2]) and WIMAX (Worldwide Interoperability for Microwave Access, IEEE 802.16 [75, 40]). In order to be able to use a VoIP service, each participant

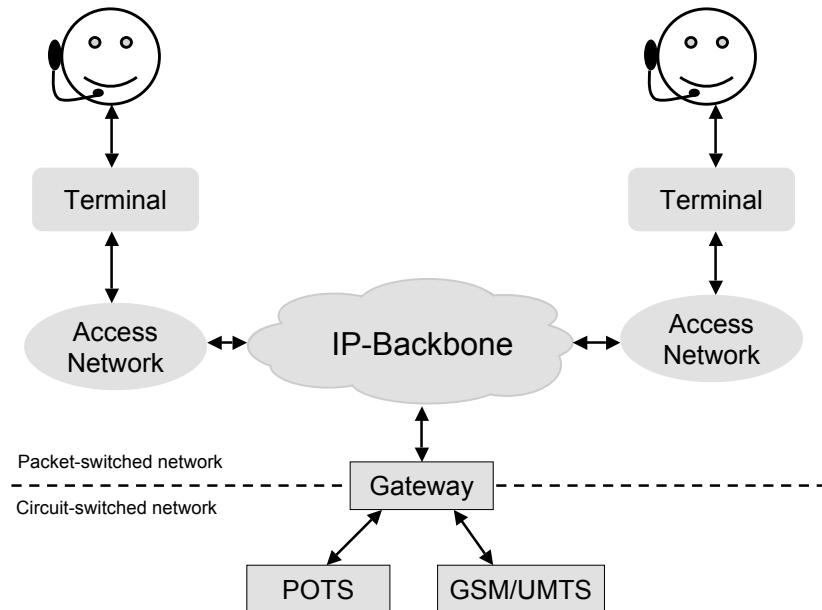


Figure 1.1: Overview: Voice-over-IP network.

needs a terminal which may either be a VoIP phone (hardware) or a computer with a VoIP software client. The technical fundamentals of VoIP are described in [39]. ITU-T Rec. E.800 [46] gives a formal definition of Quality of Service (QoS):

The collective effect of service performance which determines the degree of satisfaction of a user of the service. (ITU-T Rec. E.800 [46])

The service performance is characterized by four combined aspects: *Service support performance* describes the ability of an organization to provide a service and assist in its utilization. *Service operability performance* indicates the ability of a service to be successfully and easily operated by a user. *Service availability performance* indicates the ability of a service to be obtained within specified tolerances and other given conditions when requested by the user and continue to be provided without excessive impairment for a requested duration. Finally, *service security performance* specifies the protection provided against unauthorized monitoring, fraudulent use, malicious impairment, misuse, human mistake and natural disaster.

From a technical point of view, I distinguish two types of QoS: Network and terminal QoS. *Network QoS* is described by the parameters that determine the level of performance of the underlying network. In case of a VoIP-network, these parameters are packet loss rate, packet transmission delay, and delay jitter. On the other hand, the level of VoIP *terminal QoS* is based on the speech codec in

use, the packet loss concealment algorithm, the playout buffer mechanism, acoustic properties of the terminal, and echo cancelation.

The QoS parameters of the network and the terminal provide a description of the technical performance of the VoIP system. However, this description does in no way represent the QoS that is perceived by the persons who actually use the telephone system. Therefore, in [58], the term “Quality of Experience” (QoE) is defined as

A measure of the overall acceptability of an application or service, as perceived subjectively by the end-user. (ITU-T SG12 Contrib. D.197 [58])

QoE explicitly focuses on the users subjective perception. Since user acceptability is a crucial issue in order to provide a successful VoIP service, the network and service providers need to maintain an acceptable level of QoE. Perceived speech quality is determined from end to end, i.e., from mouth to ear. Standard methods for the measurement of the perceived speech quality are presented in Section 1.4.

In this thesis, I investigate the influence of the network QoS parameters on the perceived VoIP speech quality. As packet loss and delay are considered the most important VoIP Network QoS-parameters, I will focus on their perceptual impact¹. In addition to these technical parameters, the conversational context (e.g., type of conversation) influences the quality perception. For example, an important business call leads to higher demands on the quality of the connection than an everyday conversation with a friend in which one might tolerate a certain level of quality degradation.

Therefore, in this thesis I investigate the following topics:

- The reduction of packet discarding in error-prone transmission systems by using corrupted speech packets.
- The impact of delay on the conversational structure and on the speech quality for different conversation scenarios².

1.2 Thesis Overview and Contributions

In this thesis, I address two major quality aspects of VoIP. Firstly, I present a method for speech quality improvement by avoiding packet losses in bit-error-prone links. Secondly, I introduce three metrics for conversational interactivity and explore the impact of transmission delay on the perceived quality for a number of conversation scenarios. In the following, I briefly describe my contributions concerning these aspects.

¹Packet loss and delay are the only VoIP network QoS parameters included in the E-Model [61] which is the standard telephone network planning model (cf. Section 1.4).

²Throughout this thesis, I refer to the end to end delay as *absolute delay*

In Chapter 2, I present a method for improving the speech quality for access links which exhibit bit errors. I investigate the usefulness of keeping speech data that has been damaged by bit errors instead of dropping them and using the corrupted data for the reconstruction of the speech signal at the receiver. I simulated different levels of tolerance regarding the incorporation of erroneous speech data and evaluated the resulting speech quality using instrumental speech quality assessment methods. The unexpected results show that using all of the damaged speech data at the receiver for decoding the speech signal provides improved speech quality when compared to strategies which partly, or not at all, incorporate corrupted data. This contribution has been published in the *Acta Acustica* journal:

Florian Hammer, Peter Reichl, Tomas Nordström and Gernot Kubin, “Corrupted Speech Data Considered Useful: Improving Perceived Speech Quality of VoIP over Error-Prone Channels”, *Acta Acustica united with Acustica, special issue on Auditory Quality of Systems*, 90(6):1052-1060, Nov/Dec, 2004 [33].

A revised version of this article is reprinted in Chapter 2.

In chapters 3 and 4, I focus on the conversational interactivity of telephone conversations and the impact of end-to-end delay on conversational speech quality. If there is no echo on the line, the delay represents a degradation which is not “audible” in the sense of an impairment which is noticeable in a listening-only situation. In other words, the delay can only be noticed in a situation exhibiting some degree of interaction between the communication parties. I refer to this degree of interaction as interactivity. In this regard, I identify the following factors that have an influence on the perception of delay impairment: *Conversational Situation*, e.g., the type and purpose of the call and the user’s environment), *human factors*, e.g., age, experience, the users’ needs and behavior, and the *actual structure/interactivity of the conversation*.

This leads us to the following questions:

1. *How can we characterize the conversation structure/conversational interactivity?*
2. *What is the relation between end-to-end delay, conversational interactivity and perceived speech quality?*

In chapter 3, I seek an answer to the question about conversational interactivity. I start by formalizing the structure of conversations using the new framework of Parametric Conversation Analysis (P-CA) which defines conversational parameters and events. Moreover, I develop and investigate three metrics for conversational interactivity. This concept is applied to the recordings of conversations held in

speech quality tests. I provide an analysis of five different conversation scenarios which were used in the tests. Comparing the conversational interactivity of the scenarios, I find that the three metrics yield very similar results. One of the scenarios is highly interactive and the remaining four scenarios result in about the same amount of interactivity. I conclude that the speaker alternation rate is a simple and efficient metric to describe the conversational interactivity.

This work resulted in the following papers:

- Florian Hammer, Peter Reichl, and Alexander Raake, “Elements of Interactivity in Telephone Conversations”, 8th International Conference on Spoken Language Processing (ICSLP/INTERSPEECH 2004), Jeju Island, Korea, October 2004 [34].
- Peter Reichl and Florian Hammer, “Hot Discussion or Frosty Dialogue? Towards a Temperature Metric for Conversational Interactivity”, 8th International Conference on Spoken Language Processing (ICSLP/INTERSPEECH 2004), Jeju Island, Korea, October 2004 (Best Student Paper Award) [88].
- Florian Hammer and Peter Reichl, “How to Measure Interactivity in Telecommunications”, Proc. 44th FITCE Congress 2005, Vienna, Austria, September 2005 [31].
- Peter Reichl, Gernot Kubin and Florian Hammer, “A General Temperature Metric Framework for Conversational Interactivity”, Proc. 10th International Conference on Speech and Computer (SPECOM 2005), Patras, Greece, October 2005 [89].
- Florian Hammer, “Wie interaktiv sind Telefongespräche?”, 32. Deutsche Jahrestagung für Akustik - DAGA 06, Braunschweig, Germany, March 2006 [30].

Chapter 4 deals with question number two: The impact of end-to-end delay on the perceived speech quality. In two subjective quality tests, I collected data about the quality ratings in a variety of conversation situations of different conversational interactivity. Our results show that even in highly interactive situations, the speech quality is hardly degraded by the delay. Moreover, I tested two approaches for measuring “perceived interactivity” by means of a perceived speaker alternation rate and perceived conversation flow.

Parts of Chapter 4 have been published in the following paper:

Florian Hammer, Peter Reichl, and Alexander Raake, “The Well-tempered Conversation: Interactivity, Delay and Perceptual VoIP Qual-

ity”, Proc. IEEE Int. Conf. Communications ICC 2005, Seoul, Korea, May 2005 [35].

In the following sections, I provide background information for the interested reader regarding the sources which contribute to the end-to-end delay between two telephone terminals, and I give an overview of the standard methods for perceptual speech quality measurement.

1.3 End-to-End Delay

Absolute delay is made up of a number of individual components in both the IP-network and the terminal in use. In the following, I identify these components.

- **Speech coding delay.** Coding a digitized speech signal into a bitstream takes processing time which depends on the coding technique. Most of today’s codecs collect a block of samples resulting in a certain basic delay time. The algorithms of codecs like ITU-T G.729 or the AMR additionally take a further 5 ms “lookahead”-block of samples into account. Table 1.1 presents the algorithmic delays of various codecs. In addition, processing delay adds the same amount of delay as algorithmic delay. The decoding process takes at least the length of one block of coded data.

Codec type	Bit-rate [kb/s]	Block size [ms]	Lookahead delay [ms]	Algorithmic delay [ms]
ITU-T G.711 [41]	64	0.125	none	0.125
ITU-T G.723.1 [48]	5.3/6.3	30	7.5	37.5
ITU-T G.729 [47]	8	10	5	15
GSM EFR [21]	12.2	20	none	20
AMR [25]	4.75-10.20	20	5	25
	12.20	20	none	20
AMR-wb [20]	6.6-23.85	20	5	25
iLBC [90]	13.33	30	none	30

Table 1.1: Commonly used speech codecs and their associated coding delays.

- **Packetization.** The packetization delay represents the time needed to prepare the speech frames for RTP/UDP/IP transport (cf. Section 2.2.1). It partly depends on the packet length, which is, e.g., 60 bytes for ITU-T G.729, 90 bytes for iLBC and 200 bytes for G.711 (all: 20 ms frame including 40 bytes of RTP/UDP/IP header information). Additional time is needed for various checksum calculations. Multiple speech frames may be included in a VoIP packet, however, the delay of one speech frame is added for processing [56].

Bandwidth	Packet size		
	60 bytes	90 bytes	200 bytes
10 Mb/s	0.05	0.07	0.16
1.1 Mb/s	0.44	0.66	1.25
512 kb/s	0.94	1.41	3.13
256 kb/s	1.88	2.81	6.25
64 kb/s	7.50	11.25	25.00

Table 1.2: Serialization delays, in ms, for different transmission rates and packet sizes.

- **Serialization.** Serialization delay is the fixed amount of time needed to transmit packet frames of a certain size over a link at a certain bandwidth. Table 1.2 provides values of this kind of delay for different packet sizes and rates.
- **ADSL transmission/processing delays.** ADSL provides a “fast path” or a “slow path” for data transmission. In the slow path, an interleaver is used to improve the protection against burst noise on the DSL link. The delay produced by interleaving depends on the interleave depth (lower bound: 4.25 ms).
- **Radio link delay** The GSM radio link introduces 95 ms one-way delay from the acoustic reference point to the PSTN point of interconnect [18]. Thus, deducting the coding delay (GSM-EFR) of 40 ms from the total radio link delay, the channel coding delay and serialization delay of a radio link is about 55 ms.
- **Propagation delay (backbone).** Due to mean one-way delays of $\approx 5\mu\text{s}/\text{km}$ for optical fibre systems, and for copper the propagation delay remains low for low/middle distance calls. As an example, a connection over a distance of 600 km results in 3 ms of propagation delay. Table 1.3 presents one-way delay values for various transmission media.
- **Queueing delay.** In routers and gateways, voice frames are queued for transmission. Due to the variable states of the queues, the queueing delay is variable and contributes essentially to delay jitter.
- **VoIP gateway delay.** VoIP gateways connect IP networks with other networks like PSTN or GSM. Due to the use of different voice coding algorithms, the speech information has to be converted (transcoded) into an appropriate format (e.g., G.729-G.711 or G.711-GSM-EFR). The transcoding not only results in additional delay, but also degrades the speech quality.

Transmission media	Mean one-way delay	Remarks
Terrestrial coaxial cable or radio relay system; FDM and digital transmission	4 μ s/km	Allows for delay in repeaters and regenerators
Optical fibre cable system	5 μ s/km	Allows for delay in repeaters and regenerators
Submarine coaxial cable system	6 μ s/km	Allows for delay in repeaters and regenerators
Satellite system, 1 400 km	12 ms	Distance delay between earth stations only
Satellite system, 14 000 km	110 ms	Distance delay between earth stations only
Satellite system, 36 000 km	260 ms	Distance delay between earth stations only

Table 1.3: Transmission media delay (from [18], FDM... Frequency Division Multiplexing).

- **User Terminal.** Assuming a PC as the user terminal, an essential amount of latency is introduced by the computer equipment and software. This amount includes the playout buffering (min. packet size) sound-card latency (20-180ms, values from [69]), operating system latency, and the potential delay of the sound wave from the loudspeaker to the ear of the user (3 ms per meter).

Call setup delay is a factor that is not directly related to the end-to-end delay. It represents the time the user has to wait for a connection to be established after dialing a phone number and influences the user's communication experience. Call setup delay may distort the perceived impact of delay on speech quality.

Table 1.4 illustrates the decomposition of the end-to-end delay into its components for a typical example. I assume the transmission of two G.729 speech frames (2x10 ms frames + 5 ms lookahead + 20 ms processing delay = 45 ms) per IP/UDP/RTP packet which results in a packet size of 80 Bytes every 20 ms. Furthermore, I assume a 64 kbps link and a transmission distance of 1000 km. The decoding delay of about 2 ms has been neglected here. Further delay may have to be added for further processing like transcoding at VoIP gateways. From the resulting end-to-end delay of 180 ms, I may conclude that packet-switching introduces more delay than traditional circuit-switched land-line telephony ($T_a < 20$ ms).

Which are the requirements on a voice network regarding transmission delay? As one of the most important standardization organizations in telecommunications, the

Delay type	Delay value [ms]
Coding	45
Packetization	20
Serialization	10
Propagation	5
Queueing/Forwarding	10 (var.)
Playout buffer	60
User terminal	30
Total	180

Table 1.4: Decomposition of the end-to-end delay into its components.

International Telecommunication Union (ITU) dedicates one of their recommendations to the issues of one-way transmission time (ITU-T Rec. G.114 [56]). Major points in this standard concern the need to consider the delay impact in today's telecommunications applications, and the avoidance of delay whenever possible. G.114 recommends three areas of limits for one-way transmission delay provided that the echo of the connection is adequately controlled (table 1.5).

Delay range [ms]	Description
0-150	Acceptable for most user applications
150-400	Acceptable provided that administrations are aware of the transmission time impact on the transmission quality of user applications
above 400	Unacceptable for general network planning purposes; however it is recognized that in some exceptional cases this limit will be exceeded

Table 1.5: One-way end-to-end transmission delay limits [56]

The first area represents one-way delay times up to 150 ms, which basically do not influence a telephone conversation (except for highly interactive tasks [56]). Further up to 400 ms delay, transmission quality can be accepted for international connections with satellite hops, and one-way delays beyond 400 ms are generally unacceptable, except for unavoidable double satellite links or international video-telephony over satellites.

1.4 Speech Quality Measurement

Before describing the methods for measuring speech quality, I present the definition of the term “quality” according to Jekosch [63]:

Quality is the result of the judgement of the perceived composition of an entity with respect to its desired composition. (Jekosch, [63])

In the context of this thesis, the entity to be judged is the speech transmission system. Thus, the quality of the system is reflected in the relation between what the user expects and what she perceives. Jekosch distinguishes *quality elements* and *quality features*. Quality elements are the characteristics of a system or service which are related to their design, implementation or usage. Examples of quality elements in a VoIP system are the codec in use and network parameters such as packet loss and delay. Quality features represent the perceptual characteristics that contribute to the users’ quality perception. As an example, the distortion caused by the quality element “codec” may result in a degradation with regard to the quality feature “intelligibility”.

Before I give an overview of standard methods for perceptual speech quality measurement, I define two terms which are often used synonymously, but need to be distinguished: assessment and evaluation.

*An **evaluation** is the “determination of the fitness of a system for a purpose – will it do what is required, how well, at what cost etc. Typically for prospective user, may comparative or not, may require considerable work to identify user’s needs”* (Jekosch, [63])

*An **assessment** is the “measurement of system performance with respect to one or more criteria.”* (Jekosch, [63])

Figure 1.2 illustrates a general classification of speech quality measurement methods. On top of Figure 1.2, I distinguish between auditory and instrumental methods. Auditory methods include all kinds of methods that are based on tests involving test persons (subjective testing) either in listening-only or conversational situations. Subjective tests are time-consuming, expensive, and require appropriate test facilities. In order to reduce this effort and facilitate efficient and cost-effective quality measurement, instrumental measurement methods have been developed. Instrumental methods use perceptually motivated algorithms for estimating the speech quality based on either a speech signal (signal-based models), or on instrumentally measurable parameters of the system (parameter-based models). In the next sections, I describe these methods for speech quality measurement. A comprehensive elaboration on the assessment, evaluation and prediction of speech quality can be found in the books of Möller [72] and Jekosch [63], Raake [84] particularly addresses the speech quality of VoIP.

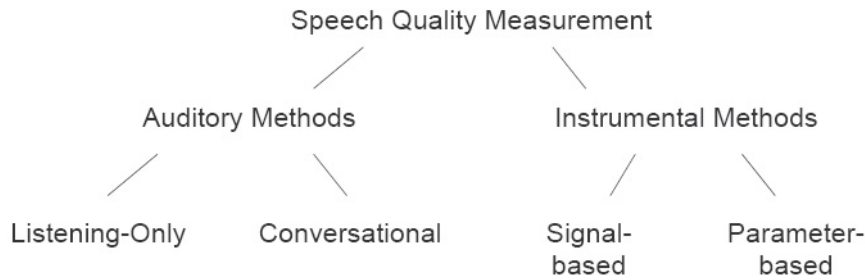


Figure 1.2: Classification of speech quality measurement methods (partly adapted from Raake [84]).

Auditory Methods

Auditory, or subjective, quality measurement methods as standardized in ITU-T Rec. P.800 [49] require test sessions with test persons. The choice of an appropriate measurement method depends on the impairments to be tested. In listening-only tests, degradations that directly impair the speech signal, e.g., noise or speech coding degradation, can be measured. However, the measurement of impairments that only occur in conversation situations, i.e., delay and echo, requires conversational testing. The results of auditory quality tests are typically presented as Mean Opinion Scores (MOS) which represent the mean ratings given by the test subjects.

Listening-Only Tests

In listening-only tests, the subjects listen to a series of speech samples and rate the quality based on an appropriate rating scale. The Absolute Category Rating (ACR) procedure is used to determine the (absolute) perceived quality of individual degraded speech samples. Table 1.6 depicts the 5-point absolute category rating scale.

Quality of the speech	Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

Table 1.6: Absolute Category Rating (ACR) scale for listening-quality (from [49]).

The Degradation Category Rating (DCR) method is used to distinguish among good-quality transmission systems for which the ACR method lacks sensitivity. In

DCR tests, the degradation of samples that passed through the system under test is rated against a high quality reference. The speech samples are either presented in pairs (A-B) or in repeated pairs (A-B-A-B) where A represents the reference sample and B represents the degraded sample. The DCR scale ranges from “inaudible” to “very annoying” as shown in Table 1.7.

5	Degradation is inaudible.
4	Degradation is audible but not annoying.
3	Degradation is slightly annoying.
2	Degradation is annoying.
1	Degradation is very annoying.

Table 1.7: Degradation Category Rating (DCR) scale (from [49]). The degradation of the second sample is rated in comparison to the first (reference) sample.

While in the DCR method the reference sample is always to be presented first, in the Comparison Category Rating (CCR) method, the order of the processed and reference sample is randomly chosen for each trial. In half of the trials, the reference sample is presented first, and in the rest of the trials, the processed signal is presented first. After each trial, the test persons are required to rate the quality of the second sample in comparison to the quality of the first using the scale presented in Table 1.8. The advantage of the CCR method over the DCR method is the possibility to measure the impact of speech processing that either impairs or improves the quality. Listening-only tests require high-quality speech material. ITU-T Rec. P.800 recommends to use samples spoken by male *and* female speakers because sophisticated processes often affect male and female voices differently [49].

3	Much Better
2	Better
1	Slightly Better
0	About the Same
-1	Slightly Worse
-2	Worse
-3	Much Worse

Table 1.8: Comparison Category Rating (CCR) scale (from [49]). The subjects rate the quality of the second sample compared to the quality of the first sample.

Conversation Tests

In a conversation test, two test persons have a series of conversations over a real-time telephone test system in a controlled laboratory environment. The subjects fulfill the tasks of a given conversation scenario. After each conversation, the subjects rate the quality of the connection they have been using on a five-point scale from “Excellent” to “Bad” (cf. Table 1.6, ACR-scale). Conversation tests are especially required for measuring the quality degradations of network parameters such as transmission delay or echo. Test scenarios are presented in Section 3.2.4.

Instrumental Measurement

This section presents methods for instrumental³ quality measurement following the classification provided by Raake [84]. None of methods presented in this section measure, or predict, the perceived speech quality directly, but always require either a speech signal or a number of quality elements within a speech transmission system. Instead, in *signal-based* methods, a model of the human auditory system is applied to the degraded received speech signal and, the perceived quality is estimated from a similarity measure that has been calculated in from a psychoacoustic representation of the received speech signal. In contrary, *parameter-based* methods predict the perceived quality based on the characteristics of the transmission system. Moreover, I present models that can be used for monitoring the speech quality of existing networks.

Signal-Based Models

The principle of intrusive signal-based speech quality estimation is illustrated in Figure 1.3. A reference speech signal is transmitted through the network under test. Both the reference speech signal and the resulting degraded speech signal are preprocessed (level and time alignment) and converted into a representation that models the human auditory system. In the perceptual domain, the signal representations are compared and similarity measures are calculated. From the similarity measures, the perceived speech quality can be estimated. Note that the quality estimation calculation is based on a large set of results from subjective quality assessments. In order to obtain meaningful assessment results, a set of at least four speech samples (two female and two male) should be applied (cf. ITU-T Rec. P.862 [54]). Intrusive signal-based speech quality estimation corresponds to paired comparison listening-only tests as described in Section 1.4. Examples for intrusive speech quality measurement are PESQ (“Perceptual Evaluation of Speech Quality”, ITU-T Rec. P.862 [54]) which is useful for measurements at the electrical

³Since these methods are based on data gathered from subjective speech quality tests, I use the term “instrumental” instead of “objective” (cf. [84]).

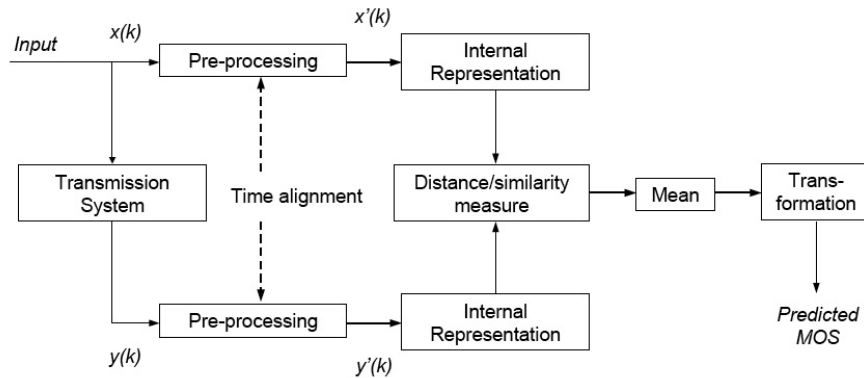


Figure 1.3: Signal-based instrumental speech quality assessment (from [38]).

interface (before the speech signal is played out acoustically), and TOSQA [5, 51] which is capable of measuring the speech quality at the acoustic interface (e.g., handsets and headsets).

A method for signal-based single-ended (non-intrusive) speech quality measurement has been standardized in ITU-T Rec. P.563 [60]. Starting from a received degraded speech sample, the algorithm reconstructs an artificial reference signal. Similar to intrusive measurement (cf. Figure 1.3, in the single-ended algorithm $x(k)$ is unknown) the degraded signal is compared with the artificial reference, and the quality is estimated. Signal-based single-ended speech quality measurement corresponds to the ACR procedure in subjective tests since the quality ratings given by users are based on their internal reference which mostly results from their experience.

Parameter-Based Models

As pointed out above, parameter-based estimation of the speech quality is based on the characteristics of the transmission system. In the following, I describe the E-model (ITU-T Rec. G.107 [61]) which represents the current network planning model recommended by the ITU-T.

The E-model allows for the prediction of QoE based on the QoS parameters. A large amount of data from auditory quality tests (intrusive offline measurements) is needed for the modeling. The E-model is based on the assumption that “(...) evaluation of psychological factors (not physical factors) on a psychological scale is additive” [44, 3]. The overall quality of the network under consideration is estimated as follows:

$$R = R_0 - I_s - I_d - I_{e,eff} + A. \quad (1.1)$$

R represents the transmission rating factor corresponding to the predicted quality. R ranges from 0 to 100, with 0 indicating worst quality and 100 the best quality.

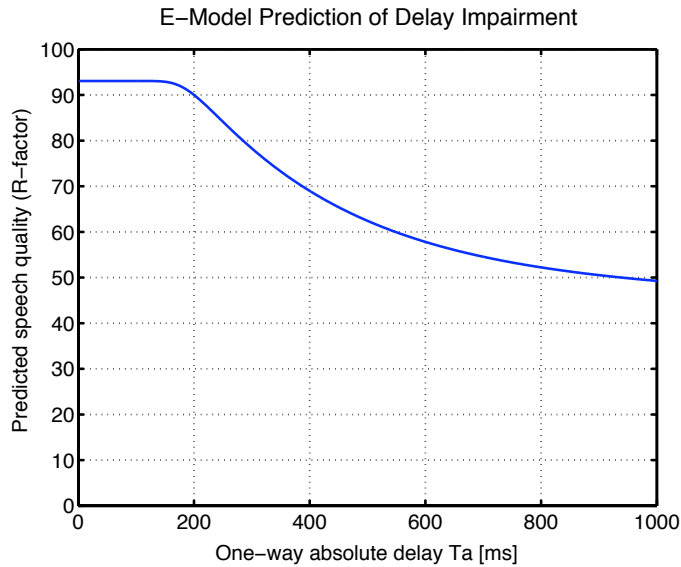


Figure 1.4: Impact of transmission delay on speech quality as predicted by the E-model [61].

R_0 stands for the basic signal-to-noise ratio, including circuit noise and room noise. The Simultaneous Impairment Factor I_s sums up all further impairments which may occur simultaneously on the voice signal, e.g., non-optimum overall loudness ratings, impairment caused by non-optimal side-tone and listener echo). The Delayed Impairment Factor I_d represents the delay impairment and combines impairment caused by echo and absolute delay. $I_{e,eff}$ describes the effective equipment impairment caused by, e.g., low-bitrate speech coding and packet loss concealment. The advantage factor A incorporates factors that provide additional benefit to the user which are not directly related to quality. Examples for such factors are mobility or access to hard-to-reach locations. In my work, I am mainly interested in the performance of the E-model regarding the impact of transmission delay. The state-of-the-art prediction of delay impairment is illustrated in Figure 1.4. At 150 ms, the predicted quality continuously decreases with delay. For illustrating the meaning of the R-factor, in Table 1.9, I present the relation between the E-model's quality ratings R and categories of speech transmission quality as given in ITU-T Rec. G.109 [53]. Note that connections with E-model ratings R below 50 are not recommended.

In the next chapter, I study the question whether the use of speech data that has been damaged by bit errors may help improve the perceived speech quality.

Range of E-Model Rating R	Speech transmission quality category	User satisfaction
$90 \leq R < 100$	Best	Very satisfied
$80 \leq R < 90$	High	Satisfied
$70 \leq R < 80$	Medium	Some users dissatisfied
$60 \leq R < 70$	Low	Many users dissatisfied
$50 \leq R < 60$	Poor	Nearly all users dissatisfied

Table 1.9: Definition of categories of speech transmission quality (from [53]).

2 Corrupted Speech Data Considered Useful

The provisioning of an appropriate level of perceptual speech quality is crucial for the successful deployment of Voice over the Internet Protocol (VoIP). Today’s heterogeneous multimedia networks include links that introduce bit errors into the voice data stream. These errors are detected by the IP packet transport protocol and result in packet losses which eventually degrade the speech quality. However, modern speech coding algorithms can either conceal packet losses or tolerate corrupted packets.

In this chapter, we investigate to which extent it makes sense to keep corrupted speech data for the special case of uniformly distributed bit errors. We simulate different transport strategies that allow the incorporation of damaged speech data into the speech decoding process. The results from an instrumental speech quality assessment show that keeping as much damaged data as possible leads to superior performance with regard to the perceptual speech quality¹.

2.1 Introduction

The advent of the Internet in the 1990s has launched an increasing interest in packet-based telephony. First packet voice transport experiments have been accomplished in the mid-1970s, but it took about 20 years to introduce an application to the public [95]. Besides the existence of a well-established circuit-switched telephone network, one of the major reasons for the slow evolution of Internet telephony may be that the Internet as such has primarily been designed to support the transmission of non-interactive, non-realtime data. In contrast, interactive applications like telephony require reliable *and* in-time data delivery, otherwise no user would accept the service. Therefore, the network providers need to maintain a certain level of quality of service (QoS) [17, 26].

Compared to the public (circuit-) switched telephone network (PSTN), the “packet-nature” of VoIP exhibits transmission impairments of its own, like packet

¹This chapter has appeared earlier as a journal paper [33], and has been revised. Hence, in this chapter I use “we” instead of “I” in order to recognize my co-authors Peter Reichl, Tomas Nordström and Gernot Kubin.

loss, packet delay, and packet delay jitter. Packet loss results from one of the major obstacles within an IP network, i.e., congestion: if too many users send lots of data at once, router queues become overloaded, and packets need to be dropped. In addition, time-varying traffic causes variations of the packet delay, the so-called jitter, which influences the probability that packets cannot be incorporated into the speech reconstruction process because they have not been received in time. Adaptive buffers can alleviate this problem by buffering packets and delaying their playout time to compensate for the varying network delay [86, 73], while however, the absolute end-to-end delay must be limited to allow a fluent conversation. For a more detailed description of VoIP speech impairments we refer to [84].

In this chapter, we are concerned about a network impairment that occurs mainly in the so-called access network, connecting the user's fixed or mobile terminal to an IP backbone network. Here, even in state-of-the-art broadband access technologies like Digital Subscriber Lines (DSL, wireline) or the Universal Mobile Telecommunication System (UMTS, wireless), transmission impairments introduce bit errors. In fact, the amount of errors represented by the *bit error rate* (BER) serves as an indicator for the quality of the transmission channel. It is important to note that losing only one bit within a packet may have dramatic consequences, as the IP voice packet transport network is designed to simply drop erroneous packets, which results in the loss of the entire information within such packets. In case of data transmission, the transmission control protocol (TCP, [80]) running on top of the Internet Protocol (IP, [79]) cares for this problem by employing a retransmission mechanism, but at the cost of increased transmission delay. In order to avoid this effect and to meet the real-time constraints, VoIP is based on the user datagram protocol (UDP, [78]) which does not retransmit lost packets.

Speech decoders deal with the packet loss problem by substituting a lost speech entity according to a packet loss concealment (PLC) algorithm [77, 105], e.g., by repeating the last received packet. The loss concealment allows to decrease the perceptual impact caused by the loss of information, but lost packets anyhow degrade the speech quality. On the other hand, modern speech codecs can tolerate a certain amount of damaged (but nevertheless delivered) data, especially if the speech bits are ordered according to their perceptual importance and if less important bits are damaged only.

This chapter presents a performance evaluation of such mechanisms. We have simulated and compared the performance of traditional and modified transport schemes as introduced in [32], where the latter either employ selected parts or even all of the damaged data. The perceptual speech quality resulting from this alternative approach has been evaluated with instrumental quality measurement methods, where "instrumental" refers to the fact that the quality is estimated

by computer algorithms instead of being assessed by test persons². In this way, we compare the modified transport schemes with traditional VoIP transport and show that keeping even all of the damaged data results in superior performance.

The remainder of this chapter is structured as follows: In Section 2.2, we present the techniques that we have used for incorporating damaged speech data and for evaluating the resulting perceptual speech quality. Furthermore, we briefly review related work. In Section 2.3, we specify three strategies that utilize the techniques for VoIP transport over error-prone links as introduced above. Section 2.4 presents the framework of the environment in which the transport strategies have been simulated. Our results are presented and discussed in Section 2.5. Finally, we draw conclusions from our work in Section 2.6.

2.2 Background

In this section, we explain the techniques that facilitate error-tolerant VoIP transport. Based on these techniques, we will propose various strategies for transmitting voice data over error-prone links in Section 2.3. Firstly, we present UDP-Lite, a UDP modification allowing bit errors in the payload. We then introduce the adaptive multi-rate (AMR) speech coding algorithm with its ability to substitute lost packets and to distinguish bits concerning their perceptual sensitivity. Robust header compression can save bandwidth by reducing the huge amount of header information resulting from the IP real-time transmission protocol stack. Then, we present the instrumental speech quality measurement methods we have applied to compare our transmission strategies in terms of perceptual quality. Finally, we briefly review some related work concerning this topic.

For convenience, the layer model we use is depicted in Figure 2.1. PHY and LL represent the physical layer and link layer, respectively. These lower layers handle the physical transmission of data either over a wireline or radio link. In this chapter, we are concerned about the layers above the link layer.

2.2.1 UDP-Lite

IP-telephony is based on the real-time transport protocol (RTP, [96]) and the user datagram protocol (UDP, [78]). The structure of a Voice-over-IP packet is illustrated in Figure 2.2. Note that the 20 Bytes of IP header contain amongst others a length field that indicates the total length of the IP/UDP/RTP packet, and a checksum that may be used to detect errors in the IP header itself (but *not* in the IP payload, i.e., the carried data). In contrast, UDP protects both header *and* payload

²However, such methods depend on information obtained from subjective listening tests. Thus, we avoid the term “objective” measurement which is widely used in the literature (see, e.g., [22]).

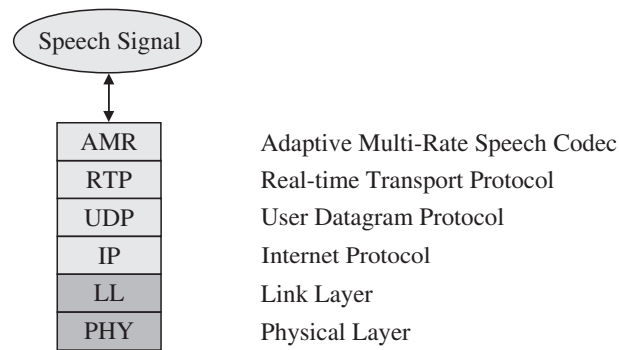


Figure 2.1: Layer model.

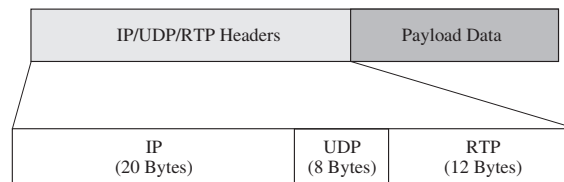


Figure 2.2: IP/UDP/RTP packet structure.

by calculating and adding a checksum to each of the packets at the sender side. Thus, routers can detect bit errors by recalculating the checksum and comparing it with the original. A difference between these numbers indicates that one or more bits in the packet have been corrupted, and as a consequence, the packet is stopped from being forwarded. Furthermore, UDP does not retransmit a packet if it got lost along its way to the receiver because, for real-time traffic, there is no time to wait for a retransmitted packet. Therefore, any packet corrupted by bit errors gets lost. For a more detailed illustration, Figure 2.3 shows the header of UDP containing the source and destination port numbers, the packet length (including the 8 UDP header bytes), and a checksum that is calculated over both header and payload data. The functionality of the length and checksum fields has been slightly varied by Larzon et al. [66]. Their proposal, the so-called UDP-Lite, allows for checksums that cover the payload only partially. To this end, the length field is substituted by a field that defines the checksum coverage size, as depicted in Figure 2.4. Therefore, only the first part of the payload is covered by the checksum, whereas bit errors are allowed towards the tail end of the payload, assuming that the link layer supports the forwarding of damaged information. Note, that the total IP/UDP/RTP packet length is given in the IP-header.

UDP encapsulates the RTP header, including a sequence number and a times-

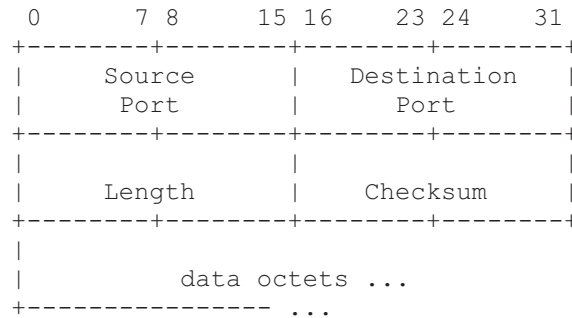


Figure 2.3: UDP header format [78]. The header fields are placed in lines of 32 bits.

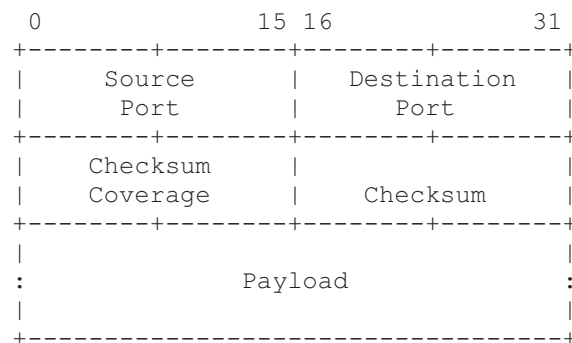


Figure 2.4: UDP-Lite header format [66]. The header fields are placed in lines of 32 bits.

tamp, and the RTP payload, i.e., the actual speech data. The RTP protocol provides mechanisms for end-to-end transport of real-time data such as voice or video. RTP is integrated in the leading VoIP signaling protocols SIP (IETF RFC 3261 [92]) and H.323 (ITU-T Rec. H.323 [57]).

2.2.2 Adaptive Multi-Rate Speech Coding

The IP/UDP-Lite/RTP protocol stack provides the transmission of speech data over the Internet (cf. Figure 2.1). In this section, we will describe important features of the adaptive multi-rate (AMR) speech codec which transforms the speech signal into a set of data frames and vice versa. Moreover, the AMR codec is especially suited for our investigations because the Internet Engineering Task Force (IETF) has defined an RTP payload format that allows for its employment in an all-IP system.

The AMR speech codec [25, 16, 105] was originally developed for the (circuit-switched) global system for mobile communications (GSM) and has then been chosen as a mandatory codec for third generation (3G) cellular systems [2]. Speech signals, sampled at 8 kHz, are processed in frames of 20 ms, and coded to bitrates ranging from 4.75 to 12.2 kbps. Thus, in a circuit-switched mobile communication system,

the codec can adapt its bitrate and the corresponding error protection according to the quality of the wireless transmission channel. The worse the channel quality, the lower the bitrate chosen, and the higher the respective error protection.

Like most of today's speech codecs, the AMR codec features an internal packet loss concealment (PLC) method, discontinuous transmission, and unequal error protection (UEP). Thus, it provides the flexibility and robustness needed for deployment in packet-based networks. For our explorations, we are mainly interested in the AMR codec's capability of providing UEP, and in its PLC algorithm.

Unequal error protection is provided at the coder's side by ordering the speech data bits of a frame according to their perceptual importance. The importance levels are referred to as class A (most sensitive), class B, and class C (least sensitive). If an entire speech frame is lost or if A-bits are corrupted during the transmission, it is recommended to forget about the corrupted packet and to use the internal PLC algorithm [23] instead. Otherwise, the damaged B/C-bits may be used. This bit ordering feature is fundamental for the construction of our strategies for error-tolerant speech data transport.

For our purposes, we have chosen the 12.2 kbps mode of the AMR codec which produces 244 speech bits per frame. These bits are divided into 81 A-bits, 103 B-bits, and 60 C-bits ([24], cf. Figure 2.5).

The *packet loss concealment* algorithm [23] works as follows. If a speech frame has been lost, the PLC algorithm substitutes this frame by utilizing adapted speech parameters of the previous frames. In principle, the gain of the previous speech frame is gradually decreased, and the past line spectral frequencies are shifted towards the overall mean of the previous frames. It is important to note that the AMR codec maintains a set of state variables that include the samples required for long-term and short-term prediction, and a memory for predictive quantizers. Therefore, aside from missing speech information, packet losses may lead to the de-synchronization of the encoder and the decoder which results in error propagation. In other words, the decoder needs some time to recover from the lost data.

The standard-compliant transport of AMR frames over IP has been defined by the IETF by specifying corresponding RTP payload formats [100]. An RTP payload format consists of the RTP payload header, payload table of contents, and payload data. Payloads may contain one or more speech frames. For our simulations, we have chosen the "bandwidth efficient mode" payload format that is illustrated in Figure 2.5. The H and T fields represent the payload header and TOC (table of contents) field, respectively, and sum up to 10 bits.

2.2.3 Robust Header Compression

IP/UDP(-Lite)/RTP transport of speech data results in a major drawback regarding the transmission efficiency. The protocol headers in total form a 320 bits large cluster



Figure 2.5: AMR RTP payload format: Bandwidth efficient mode [100]. Note that the H and T fields constitute the RTP *payload header* and are not part of the RTP header.

- H ... RTP payload header (4 bits)
- T ... RTP payload table of contents (6 bits)
- A ... 81 Class A speech bits
- B ... 103 Class B speech bits
- C ... 60 Class C speech bits
- P ... 2 Padding bits

(20 Bytes IP, 8 Bytes UDP, 12 Bytes RTP) of administrative overhead. Assuming that a packet carries only one speech frame and contains 256 actual payload bits, this overhead comprises more than half of the total packet size. Hence, the majority of packets are lost due to bit errors in the headers when sent over an error-prone serial link.

Robust Header Compression (ROHC, [9]) resolves this problem by utilizing redundancy between header fields within the header and in particular between consecutive packets belonging to the same packet stream. In this way, the overhead can be reduced to a minimum. The term “robust” expresses that the scheme tolerates loss and residual errors on the link over which header compression takes place without losing additional packets or introducing additional errors in decompressed headers. ROHC profiles for UDP-Lite are defined in [76].

In our simulations, we are able to reduce the header size from 40 to 4 Bytes using this technique, i.e., 10% of its original size. Figure 2.6 illustrates the efficiency of the compression. Note that the 10 bits of RTP payload header and TOC are additionally included in the headers.

2.2.4 Speech Quality Evaluation

Instead of quality of service as characterized by technical parameters, users are first of all concerned about the quality of service as perceived by themselves, because they want to communicate in a comfortable way without having to care about the underlying technology (cf. Section 1.1).

Perceived quality is primarily measured in a subjective way. To this end, test persons rate the quality of the media either in listening-only or conversational tests. The subjective assessment of speech quality is addressed in ITU-T Recommendation P.800 [49]. Absolute Category Rating (ACR) is the most common rating method

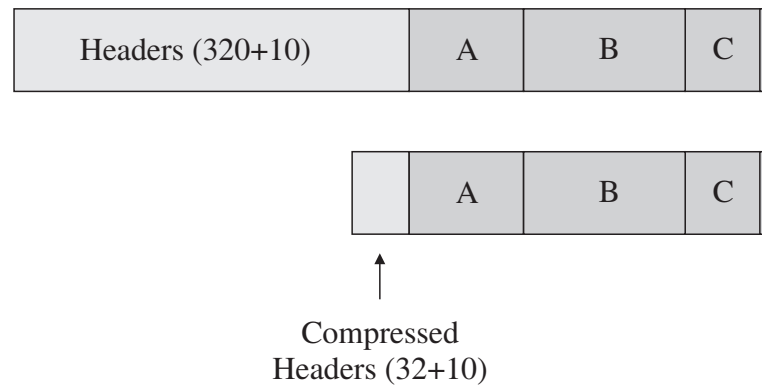


Figure 2.6: Efficiency of robust header compression.

for speech quality listening tests. It is based on the listening-quality scale shown in Table 1.6. The quantity evaluated from the score averaged over the complete set of test persons is called Mean Opinion Score (MOS).

Currently, a lot of research effort is invested in developing algorithms that derive an instrumental measure of the perceived quality, often referred to as “objective” measure. So-called “intrusive” instrumental speech quality assessment algorithms compare a degraded speech signal with its undistorted reference in the perceptual domain, and estimate the corresponding speech quality. This principle is shown in Figure 2.7 (see also Figure 1.2 in Section 1.4). In comparison, “non-intrusive” instrumental assessment methods do not require a reference signal, but estimate the perceptual speech quality by measuring network parameters.

In our experiments, we evaluate the perceptual quality of the speech samples resulting from our simulations by using ITU-T Rec. P.862 “Perceptual Evaluation of Speech Quality” (PESQ, [54]) and the “Telecommunication Objective Speech

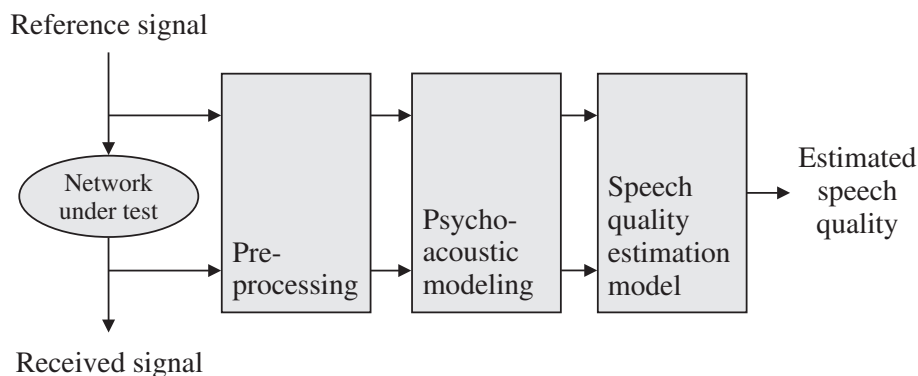


Figure 2.7: “Intrusive” instrumental perceptual speech quality assessment.

Quality Assessment” (TOSQA, [51]). Compared to PESQ, an important feature of TOSQA is a modification of the reference signal by utilizing an estimated transfer function of the spectral distortions. Therefore, some of the effects of linear distortions can be balanced.

2.2.5 Related Work

After presenting the technical background on which we base our investigations, we give a brief overview of related work. The application of UDP-Lite for video transmission over wireless links has been explored by Singh et al. [99]. In that work, GSM radio frame error traces have been collected in a cellular IP testbed, and have then been used to simulate the transmission of video streams over a wireless link. Compared to traditional UDP, the use of UDP-Lite provides 26% less end-to-end delay, constant inter-arrival time of the packets, slightly higher throughput, and 50% less effective packet losses. The perceptual quality is claimed to be significantly higher, but neither subjective nor instrumental quality assessment has been accomplished.

In addition to packet loss concealment, forward error correction (FEC) can be used to compensate for packet losses [77]. At the cost of bandwidth and delay, either Reed-Solomon (RS) block coded data [7] or low bit-rate redundancy data (LBR), i.e., a low quality version of the same speech signal, are added as redundant information within one of the following voice packets or in a separate packet. Jiang and Schulzrinne [64] show that LBR performs worse, with regard to the perceptual speech quality, than the use of FEC in terms of RS-codes.

However, in our study we aim to investigate the impact of bit errors on the transmitted speech data, so we do not apply any additional FEC or channel coding for our simulations, with the exception of perceptual bit ordering.

2.3 Alternative Strategies for VoIP Transport

This section introduces the strategies following [32]. We explore the impact of using erroneous speech data on the perceived speech quality by defining three strategies which handle corrupted packets in different ways. The strategies are based on IP/UDP-Lite/RTP transport of AMR speech frames facilitating different UDP-Lite checksum coverage which is illustrated in Figure 2.8 and Table 2.1. In addition, we apply ROHC to the headers.

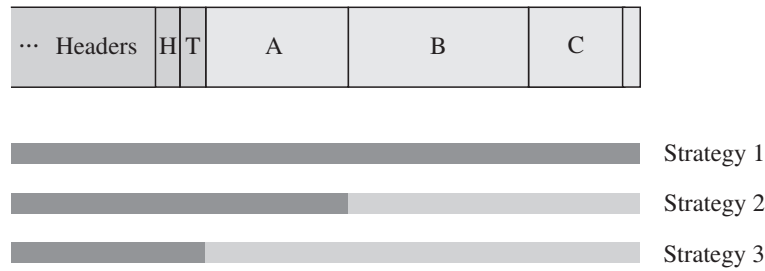


Figure 2.8: UDP-Lite checksum coverage. The dark gray shading indicates the regions in which the speech data is protected by the checksum. In contrary, light gray shading indicates unprotected speech data.

We define the strategies as follows:

- *Strategy 1* simply corresponds to traditional IP transport, hence the UDP-Lite checksum covers the entire UDP payload. If any data is corrupted, the packet is lost and substituted by the receiver’s PLC. Including the traditional transport method into the simulations constitutes a reference with regard to the speech quality performance.
- In accordance with the AMR standard, *strategy 2* permits B- and C-bits to be faulty, but detect errors within the header and the class A bits. Thus, a reasonable amount of packets with erroneous B- and C-bits can be saved.
- *Strategy 3* exhibits the most tolerant behavior. All of the payload data are allowed to be corrupted, consequently a packet is only dropped when the header is corrupted. All of the corrupted speech data can be incorporated in the reconstruction of the speech signal.

Under any strategy, the IPv4 header is protected by its own checksum.

In order to further characterize the strategies, we introduce the *coverage degree* α as a parameter that corresponds to the relation of the checksum coverage N' to the total packet length N (including the headers),

$$\alpha = \frac{N'}{N}. \quad (2.1)$$

Hence, a coverage degree of $\alpha = 0$ means that none of the data is covered by the checksum, and a coverage degree of $\alpha = 1$ indicates that the entire packet is covered by the checksum. Note that the smaller the coverage degree α , the less packets are discarded due to bit errors (cf. Section 2.4.2).

Table 2.1 summarizes the properties of the three proposed strategies and Table 2.2 provides the corresponding values of the coverage degree α . Note for the extreme case of strategy 3, the use of ROHC may reduce the coverage degree down to 0.15.

Corrupted part of packet	Strategy		
	1	2	3
Header	drop	drop	drop
A-bits	drop	drop	keep
B/C-bits	drop	keep	keep

Table 2.1: Packet drop strategies.

Strategy	No ROHC		ROHC	
	[bits]	α	[bits]	α
1	576	1	288	1
2	411	0.71	123	0.43
3	330	0.57	42	0.15

Table 2.2: UDP-Lite checksum coverage and coverage degrees.

2.4 Simulations

2.4.1 Simulation Environment

The simulation environment, as depicted in Figure 2.9, represents an example for the interworking between signal processing and networking methods³. It contains the following parts: the speech database, AMR speech encoding and decoding, a Matlab [71] module simulating the strategies specified in Section 2.3 for different bit error rates, and the perceptual speech quality assessment unit.

After coding a speech sample, the voice bitstream is processed corresponding to each of the three transport strategies. To obtain a good resolution of the area of decreasing speech quality, we have chosen 13 bit error rates between 10^{-5} and 10^{-3} . The three bitstreams are then decoded, and instrumental measurements estimate the perceptual quality of the degraded speech samples. This procedure is repeated 24 times per bit error rate per speech sample in order to get good average values of the resulting speech quality. For our investigations, we have chosen the PHON-DAT speech sample database [27] that contains phonetically rich German sentences recorded in studio-quality (16 bit/16 kHz). We have selected 12 sentence pairs spoken by 4 speakers (2 female and 2 male). The speech samples were down-sampled to 8 kHz, modified-IRS [50] filtered and normalized to an active speech level of -26 dBov (units of dB relative to overload, [45]).

³This simulation environment represents a special case of our method of evaluating the speech quality of transmission channels using error traces as presented in [36].

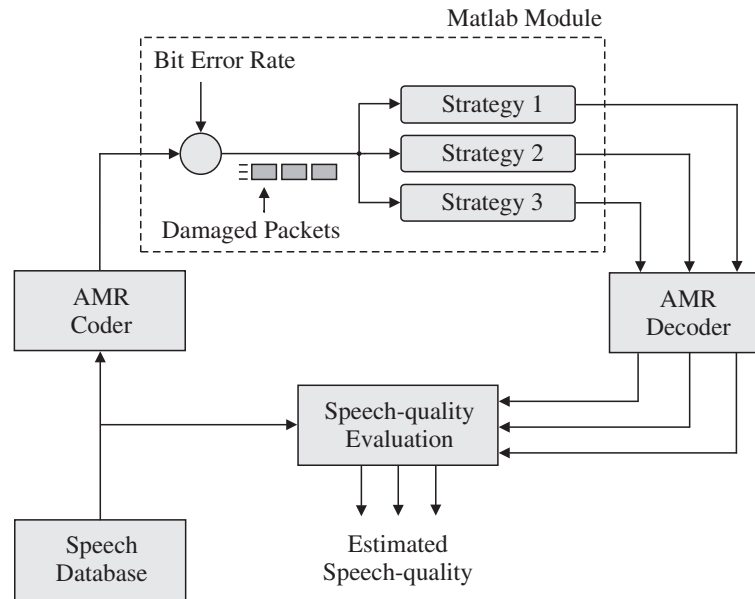


Figure 2.9: Simulation environment.

2.4.2 Bit Error Model

As already mentioned in the introduction, digital transmission of data over wire-line or wireless access networks can result in a certain amount of bit errors. The amount and the distribution of the bit errors can be controlled by channel coding. In this study, we assume the special case that the channel coding at the physical link provides uniformly distributed bit errors. We further assume that the lower system layers provide support for UDP-Lite by forwarding erroneous data to the upper layers.

Based on these assumptions, the number of bit errors X that occur in an actual packet is calculated using the binomial distribution

$$X \sim B(N, p), \quad (2.2)$$

where N represents the packet size [bits] and p represents the bit error rate. The location of the erroneous bits within the packet is then uniformly distributed over the packet.

To be able to present the effects of keeping damaged data in detail, we deal with bit error rates ranging from 10^{-5} to 10^{-3} which can be expected for wireless channels. For Digital Subscriber Lines (DSL), the BER is typically controlled at 10^{-7} . However, our choice of bit error rates might be relevant for “customized” wireline techniques like “Channelized Voice over DSL” (referred to as “voice over data service” in ITU-T Rec. G.992.3 “ADSL2” [55]).

At this range of the bit error rate, the behavior of the strategies highly affects the

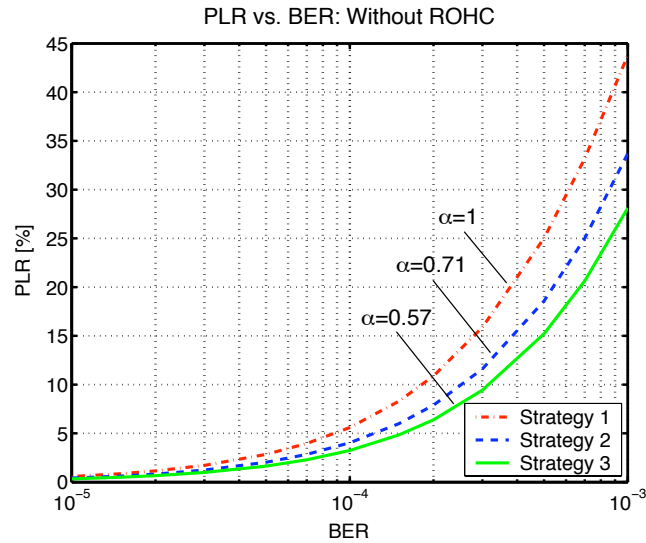


Figure 2.10: Relation between packet loss rate and bit error rate: Without ROHC.

amount of packets lost due to bit errors. The packet loss rate, PLR , depends on the bit error rate p according to

$$PLR(p) = 1 - (1 - p)^{\alpha N}, \quad (2.3)$$

where α represents the checksum coverage degree as defined in Equation (2.1). Figures 2.10 and 2.11 depict the packet loss relations among the three strategies without and with ROHC, respectively. The graphs illustrate that the loss rate is substantially reduced by compressing the header and by reducing the checksum length.

2.5 Results and Discussion

2.5.1 Estimated Perceived Speech Quality vs. Bit Error Rate

At first, we compare the performance of the strategies with regard to the perceptual speech quality estimated by PESQ as a function of the bit error rate. We have chosen PESQ as the main tool for the quality evaluation, since it is widely used and has been standardized by the ITU-T. The results for the non-header compressed case are shown in Figure 2.12. The differences in quality are noticeable for strategies 2 and 3 compared to strategy 1. Strategy 3 performs best, although the average improvement compared to strategy 2 is only marginal. The standard deviations of the speech quality MOS estimates are around 0.14, 0.25, and 0.19 at bit error rates of 10^{-5} , 10^{-4} , and 10^{-3} , respectively, for all strategies. In less than 1% of the test cases strategy 2 performs better than strategy 3 for the non-header compressed

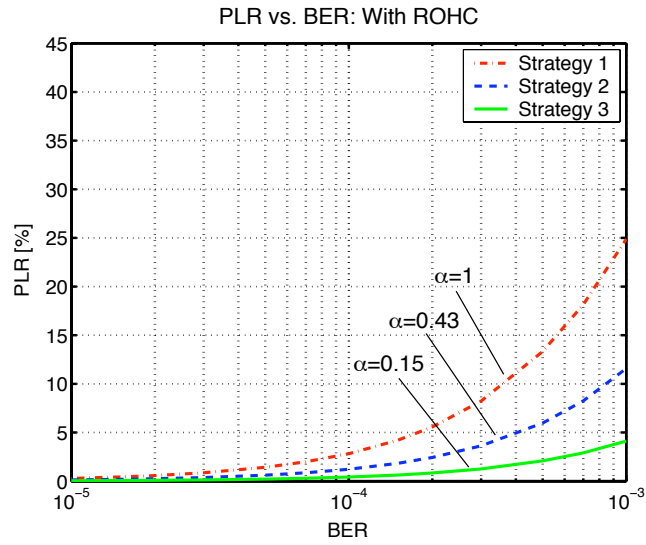


Figure 2.11: Relation between packet loss rate and bit error rate: Using ROHC.

case. However, this behavior can only be observed at very low bit error rates. When the packet header is compressed, strategy 3 significantly outperforms strategy 2. Figure 2.13 shows that at a bit error rate of 10^{-3} , strategy 3 results in an estimated perceptual quality that is half a MOS point higher compared to strategy 2, and an increase of more than one MOS point compared to strategy 1. The standard deviations of the MOS estimates for strategies 1/2/3 are 0.25/0.20/0.14 for a bit error rate of 10^{-4} , and 0.21/0.24/0.23 for a bit error rate of 10^{-3} . Similar to the

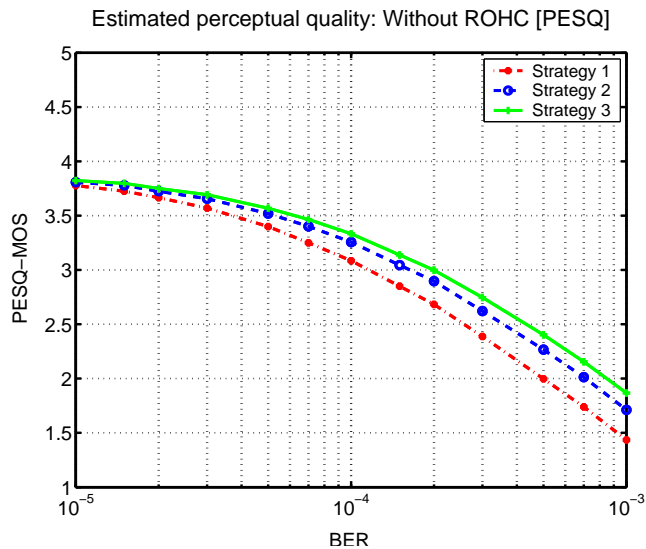


Figure 2.12: Relation between PESQ-MOS and bit error rate: Without ROHC.

non-header compressed case, strategy 2 performs better than strategy 3 at low bit error rates in only 0.7% of the test cases. This underlines the consistent trend of the results. As a result, we conclude that applying the packet loss concealment in case of erroneous A-bits performs worse than keeping them for decoding. This result may reflect the fact that employing corrupted data saves a considerable amount of packets from being dropped (cf. Section 2.4.2). As the codecs maintain an internal state, they need some time to recover from a lost packet. Employing damaged speech data introduces artifacts but avoids such error propagation. We conclude that for a certain bit rate, a damaged speech data packet is of significantly higher “perceptual value” than its substitution by the loss concealment. We regard this conclusion to be one of the central results of our investigations.

2.5.2 Gender Dependency

The dependency of the estimated speech quality on the gender of the speakers for the header compressed case is shown in Figures 2.14 and 2.15 for PESQ and TOSQA, respectively. The PESQ results indicate a significant difference between samples of female and male speakers. At low bit error rates, male voices are rated about 0.23 MOS points higher than female voices. For strategies 1 and 2, this difference decreases with increasing bit error rate, while for strategy 3 it slightly increases. At 10^{-3} , the differences are 0.18, 0.21, and 0.25 for strategies 1, 2, and 3, respectively. In contrast, quality evaluation using TOSQA results in marginally better quality for female voices for all strategies at low bit error rates. Additionally, at high bit error rates, female voices are rated slightly better than male voices. This small difference

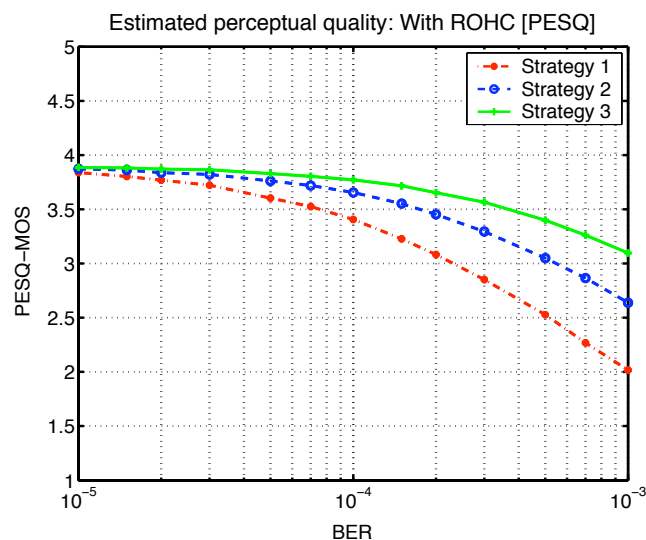


Figure 2.13: Relation between PESQ-MOS vs. bit error rate: Using ROHC.

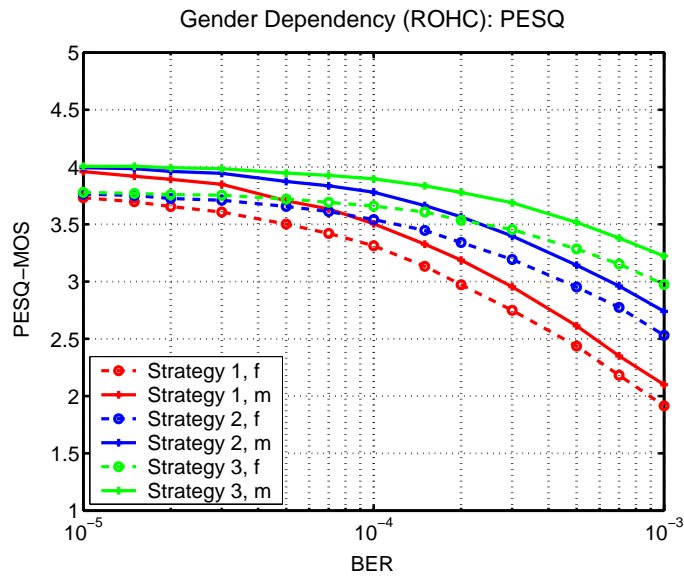


Figure 2.14: Gender dependency of the speech quality estimated by PESQ when ROHC is used.

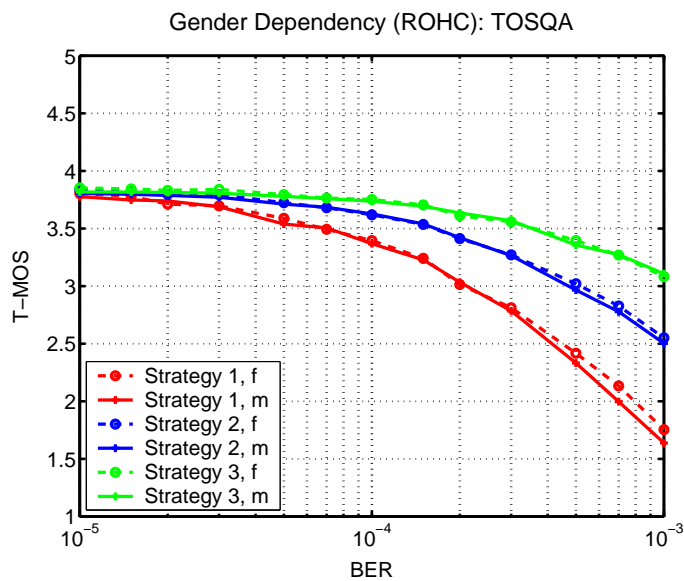


Figure 2.15: Gender dependency of the speech quality estimated by TOSQA when ROHC is used.

decreases with an increasing amount of damaged data that is incorporated into the speech decoding process.

The difference between the PESQ and TOSQA evaluation results may have two reasons:

- The *speech coding algorithm* results in different speech quality for female and male speakers.
- The *instrumental speech quality assessment methods* behave in different ways.

From subjective tests conducted at AT&T for AMR characterization [19], we imply that there is no significant speaker dependency for the AMR codec.

A closer look at the input filter responses at the preprocessing stages of PESQ and TOSQA (cf. Figure 2.7) offers a possible explanation for its different behavior. PESQ cuts the signal energy below 250 Hz and applies an IRS receive filter [42] to the input signal. In comparison, TOSQA uses an input frequency response that is based on acoustic handset measurements [5]. This frequency response is, especially at the lower frequencies, more bandlimited than the frequency response of the PESQ input filter. The different input filter responses may be the main reason for the difference of the results regarding the gender of the speakers. However, this issue is out of the scope of this thesis.

2.5.3 PESQ vs. TOSQA

As a final result, we present a comparison of the mean speech quality estimates given by PESQ and TOSQA. The results of both methods are given in Figure 2.16 for the header compressed case. The major observation is that for strategies 1 and 2, TOSQA estimates a lower MOS compared to PESQ at high bit error rates. At a BER of 10^{-3} , the differences in estimated quality are 0.3 and 0.1 for strategies 1 and 2, respectively. On the contrary, PESQ as well as TOSQA provide equal quality results at higher bit error rates for strategy 3. As we can observe from Figures 2.14 and 2.15, the difference between the results of PESQ and TOSQA seems to be caused mainly by the different ratings obtained for speech samples of male speakers.

In any case, we may conclude that the TOSQA results approve the trend, indicated by the PESQ results, that the use of all corrupted data results in superior speech quality.

2.6 Summary

In this chapter, we have simulated a VoIP system making use of speech data that have been corrupted due to bit errors. We have distinguished traditional VoIP transport which drops damaged packets, the use of corrupted data that is perceptually less sensitive, and the incorporation of all available erroneous data into the speech decoding process.

The results of an instrumental perceptual speech quality assessment clearly indicate that keeping all damaged speech packets in combination with robust header compression results in superior performance compared to dropping the damaged

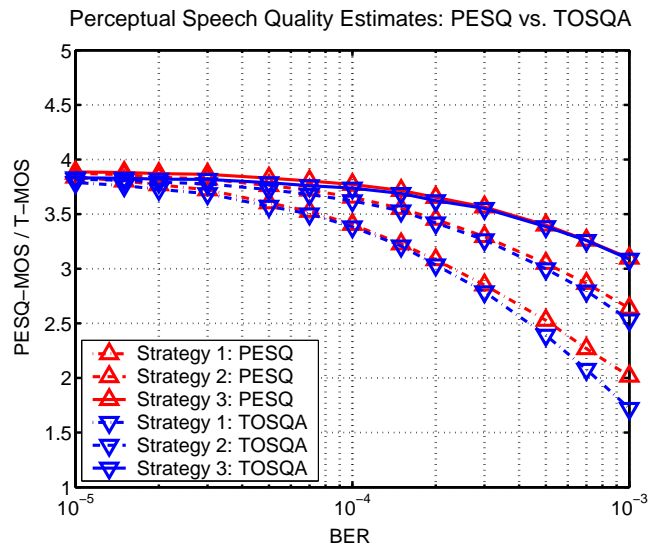


Figure 2.16: A comparison of PESQ and TOSQA results (Using ROHC).

packets and utilizing the packet loss concealment algorithm at the receiver. It is especially remarkable that the dropping of packets which contain perceptually sensitive corrupted speech bits does not yield a gain in quality compared to the use of all erroneous data. Thus, in fact all corrupted speech data have to be considered useful.

The study presented in this chapter does not at all incorporate the impact of transmission delay on the conversational speech quality, since PESQ is only intended to estimate the listening-only speech quality.

After successfully investigating the potential quality improvement that results from the avoidance of packet losses by considering damaged speech packets, in the next chapter, we model the interactivity of telephone conversations in order to be able to distinguish conversation scenarios.

3 Modeling Conversational Interactivity

3.1 Introduction

In the previous chapter, I have described a method which improves the perceived one-way speech quality. If I want to measure conversational speech quality (two-way), I need to take the end-to-end delay into account which I have identified as one of the key QoS parameters of VoIP in the introduction. In the absence of echo, which would make delay perceivable in terms of a listening and talking quality parameter¹, the quality degradation caused by delay cannot be “heard” by the user. Rather, the latency may impair the users’ ability to interact with the conversation partner. The potential impairment on this interaction depends on the conversation context. We define the *conversation context* as

the set of circumstances in which a conversation occurs. The conversation context covers the conversational situation and human factors.

The *conversational situation* arises from the purpose of the call, the callers environment (e.g., quiet vs. noisy surroundings), the number of participants, and the distance between the participants. *Human factors* include the user’s intentions, experience in communicating and in using the technology, the user’s character, e.g., patience and aggressiveness, her behavioral nature in communicating, e.g., offensive vs. defensive, and the age of the user. Note that, within the conversational context, some situational factors and human factors are tightly interweaved, e.g., the purpose of the call and the users intention.

Regarding the amounts of delay that users would accept, ITU-T Rec. G.114 summarizes the bounds for one-way delay (cf. Section 1.3). Based on these bounds, the E-model [61] (cf. Section 3) models the impact of transmission delay via the delay impairment factor I_d which takes the absolute delay T_a into account (cf. Equation 1.1). However, this modeling approach does not consider the conversational context by means of conversation situation. In conversational speech quality measurement, the situation is represented by the scenarios on which the conversations held by the test subjects are based (cf. Section 3.2.4). An instrumentally measurable metric for

¹An instrumental approach to estimate talking quality is presented in [4], i.e., “Perceptual Echo and Sidetone Quality Measure” (PESQM).

conversational interactivity would be beneficial for taking the conversational *situation* (scenario) into account within the E-model. My definition of “conversational interactivity” will be given in Section 3.2.1.

In this chapter, I aim at a detailed investigation of the conversational interactivity at different delay conditions, and with regard to the conversation context, i.e., conversational situations modeled by different conversation scenarios. Firstly, I define a set of parameters describing the structure of a given conversation. I define conversational interactivity and introduce three metrics for this parameter. Secondly, I evaluate these metrics by applying them to conversations recorded during conversational quality tests.

This chapter is structured as follows. After presenting related work and giving a definition of conversational interactivity in Section 3.2, I describe the framework of parametric conversation analysis in Section 3.3. Section 3.4 presents a selection of models for conversational interactivity. In Sections 3.5 and 3.6, I describe two experiments I have carried out for investigating the relation between conversational interactivity and transmission delay. The results of these experiments are presented in Section 3.7. In Section 3.8, I summarize this chapter and draw conclusions from the results.

3.2 Related Work

This section presents related work concerning the analysis of conversations by means of conversational parameters and interactivity. I survey existing definitions of the term “interactivity”, and I cover related work on conversational parameters, the concept of turn-taking, and conversation scenarios.

3.2.1 Definitions of Interactivity

Surveying the literature, we find that interactivity is a concept used in a wide range of disciplines such as communication science, new media research, human-to-computer interaction, and web-design among others. In the following², I present an overview of different definitions of interactivity. I follow an explanation of the concept of interactivity given by Kioussis [67], who came up with a classification of interactivity as shown in Figure 3.1. He clusters the definitions of interactivity into three major groups: structure of technology, communication context and user perception.

All interactive services are based on some underlying technology, thus, I will first discuss definitions of interactivity which refer to the structure and capabilities

²Parts of this section have been published previously in [31] and have been revised.

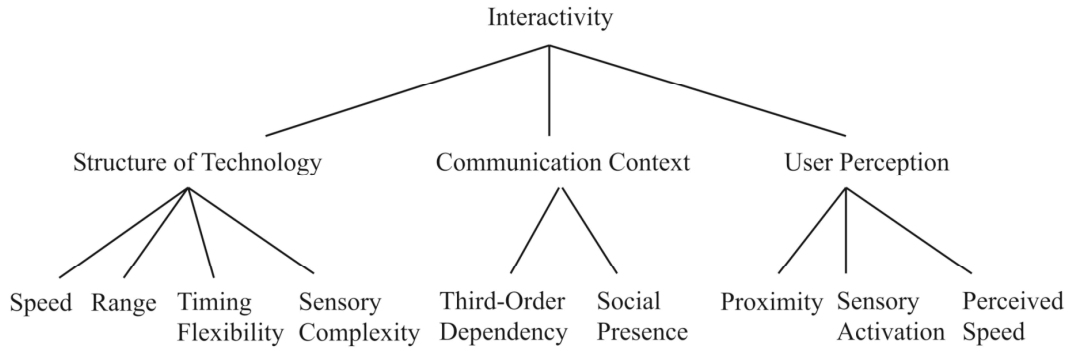


Figure 3.1: Categorization of interactivity (from [67]).

of that *technology*. Within the context of virtual reality, Steuer [102] defines interactivity as “the extent to which users can participate in modifying the form and content of a mediated environment in real time”. In the context of this definition, two essential factors contribute to interactivity: speed and range. *Speed*, or response time, refers to the rate at which input can be interpreted and realized within the mediated environment. As an example, the control unit of an interactive art installation takes some time to react to the data produced by a “data glove” which samples the hand movements of the performing user. *Range* refers to the number of attributes of the mediated environment that can be manipulated and by the amount of variation possible within each attribute. Examples for the range are spatial organization (where objects appear), and intensity, e.g., loudness of sounds, brightness of images, and intensity of smells. Downes and McMillan [15] identify the flexibility of message timing as a key dimension for both real-time communication and asynchronous communication such as email and newsgroups. *Timing flexibility* refers to the degree to which users can control the rate of information flow, e.g., a user can to great extent decide when to reply to an email. As an additional technical aspect, *sensory complexity*, the amount of devices employed by a system to activate the five human senses, contributes to the level of interactivity that the system can provide. For example, written text in a chat activates the visual sense, whereas a telephone conversation activates the acoustic sense.

In the group of definitions related to the *context of communication*, Rafaeli [85] gives a clear definition of interactivity towards *third-order dependency*. He defines interactivity as “an expression of the extent that in a given series of communication exchanges, any third (or later) transmission (or message) is related to the degree to which previous exchanges referred to even earlier transmissions” [85]. A message B as a response to the previous message A can be identified as re-action, while the following message C that is related to message B, and thus related to message A, creates inter-active communication. Third-order dependency can be quantified as

the percentage of third-order messages within a communication event.

As another context-related parameter, *social presence* is defined by Short et al. [98] as “the salience of the other in a mediated communication and the consequent salience of their interpersonal interactions”. Examples for social presence are the continuation of threads in email replies, e.g., “Subject: Re:”, and addressing other participants by name, e.g., “What is your opinion, Tom?”.

Regarding the *users’ perception*, three major dimensions are identified: proximity, sensory activation, and perceived speed [12, 14, 74]. *Proximity* represents the degree to which communication participants feel that they are “near” other participants when using the system. *Sensory activation* refers to the degree of subjects using their senses (sight, hearing, touch) during a communication event. Moreover, *perceived speed* determines how fast users thought that the system allowed the participants to react to each others transmissions.

Slightly more appropriate for our purposes, a recent study on the performance of default speech codecs carried out by 3GPP [1] reports on the use of interactivity in subjective quality tests. The test subjects were asked to judge the conversation when interacting with the conversation partner based on the occurrence of double talk (test persons talking simultaneously) and interruptions. As an example, “fair” interactivity was described as “sometimes, you were talking simultaneously, and you had to interrupt yourself”.

In this thesis, I aim at defining an instrumental metric for conversational interactivity that allows me to distinguish conversation scenarios at different conditions of delay based on the participants’ speech signals. None of the above mentioned definitions for interactivity seems viable for this purpose. Therefore, I create my own definition:

Conversational Interactivity is a single scalar measure based on the quantitative attributes of the participants’ spoken contributions.

Based upon this definition I will construct our instrumental metrics for conversational interactivity in Section 3.4

3.2.2 Conversational Parameters

As to be able to describe the conversational structure, I need to define its characteristic parameters. Brady [11] provided a detailed analysis of a set of conversation parameters which he extracted from the recordings of 16 conversations. He defined conversational events like talk spurt, mutual silence, double talk, and interruptions. The test persons who participated in the data collection were close friends who were asked to talk about anything they wished. Brady analyzed the speech data by using a threshold-based speech detection algorithm that filled pauses smaller than 200 ms

and skipped talk spurts shorter than 15 ms.

ITU-T Rec. P.59 [43] provides a standard method for artificial conversational speech pattern generation. The standard is based on a four-state model for two-way conversations and gives the corresponding temporal parameters. The parameters were obtained from a set of conversations in English, Italian, and Japanese. In the results section, I present the values of these parameters for comparison with our own results.

3.2.3 The Concept of Turn-Taking

The conversational parameters which Brady has defined (cf. previous section) are related to talk spurts, i.e., utterances. At a higher level of abstraction, I can determine whether speaker A or speaker B has the conversation floor. This kind of description leads us to the concept of turn-taking (cf. Sacks et al. [94]) which is used in “traditional” Conversation Analysis (CA, [28,81,104]). Besides the speaker turns, semi-verbal utterances like “uh-huh” or laughter are considered backchanneling events which maintain the rapport between speaker and listener and may encourage the speaker to proceed. While in CA, backchanneling events are not regarded as speaker turns, in our investigations I do not distinguish turns and backchannels because I cannot instrumentally identify backchannels on a purely speech signal based analysis of the conversation.

3.2.4 Conversation Scenarios

A fundamental element of the telephone conversation situation is the purpose of the call. Hence, in subjective tests, appropriate conversation tasks need to be used to provoke the participants to talk to each other. This section briefly sketches a number of scenarios that have been used in related studies.

In his analysis of on-off patterns of conversations, Brady [11] asked the test persons to simply talk about whatever they wanted (cf. Section 3.2.2). Richards [91] describes scenarios which have been used for conversational speech quality tests in the 1970s. These scenarios include the annotation of maps or random shapes. In 1991, Kitawaki [68] carried out subjective speech quality tests with focus on pure delay impairments. The six tasks he used stimulated the conversations in different ways and are listed in Table 3.1. Möller [72] (see also Wiegelmann [106]) presents the usage of a set of tasks denoted as Short Conversation Tests (SCT). SCTs represent real-life telephone scenarios like ordering a pizza or reserving a plane ticket, leading to natural, comparable and balanced conversations of a short duration of 2–3 minutes. These properties allow for efficient conversational speech quality testing. The SCT scenarios are now commonly used in conversational speech quality assessment (e.g., in [29]).

In this thesis, I use the term “scenario” as a generic term for a set of similar tasks

Task 1	Take turns reading random numbers aloud as quickly as possible.
Task 2	Take turns verifying random numbers as quickly as possible.
Task 3	Complete words with missing letters.
Task 4	Take turns verifying city names as quickly as possible.
Task 5	Determine the shape of a figure described verbally.
Task 6	Free conversation.

Table 3.1: The six conversation scenarios as used by Kitawaki [68].

which result in comparable conversational structure and conversational interactivity. In turn, I define a conversation task as one implementation of a given conversation scenario.

3.3 Parametric Conversation Analysis

In this section, I introduce the new framework of Parametric Conversation Analysis (P-CA) which formalizes the structure of conversations by means of parameters that can instrumentally be extracted from conversation recordings. The term “parametric” facilitates the distinction from traditional CA which mainly investigates semantic/pragmatic aspects of conversations. The P-CA of two-way conversations is based on the parameters of a 4-state conversation model and conversational events which are described in the following sections: After introducing the underlying conversation model in Section 3.3.1 and conversational events in Section 3.3.2, I illustrate the impact of transmission delay on the conversation parameters in Section 3.3.3.

3.3.1 Conversation Model

The two-way conversation model I use discriminates four different states, as shown in Figure 3.2 (cf. ITU-T Rec. P.59 [43]). States A and B represent the situations that either speaker A or speaker B is talking only. State M (“mutual silence”) denotes the case that nobody talks at all, and state D (“double talk”) reflects the situation that both speakers are talking simultaneously. Based on these four states, a conversation can be modeled as a stochastic (e.g., Markov) process (cf. [43, 93]) as illustrated in Figure 3.3. The transitions between states A and B and between states M and D are omitted because these events occur very rarely, but could easily be included. The Markov process is usually described by the transition probabilities between the states. In our investigations, however, I will analyze the sojourn times t_A, t_B, t_M, t_D which represent the mean durations that the conversation remains in the corresponding state, and the state probabilities p_A, p_B, p_M and p_D which indicate the probabilities of the conversation being in states A, B, M , and D , respectively.

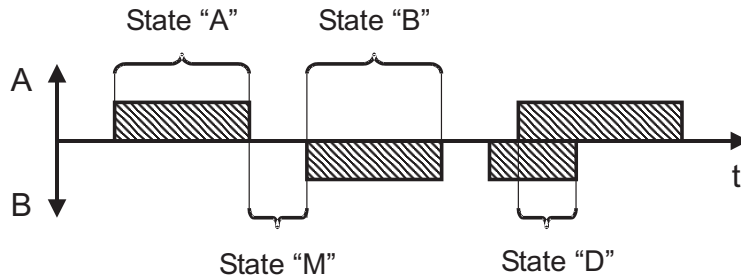


Figure 3.2: Division of the conversation structure into four states. The upper rectangles denote utterances stated by speaker A and the lower rectangles represent utterances stated by speaker B.

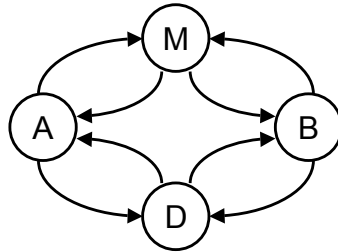


Figure 3.3: Modeling a conversation as a Markov process.

3.3.2 Conversational Events

In the following, I present the conversational events that provide more information about the characteristics of a conversation. As *speaker alternations* I consider either *speaker changes* which I define as state sequences in which two talk spurts are separated by mutual silence (A-M-B and B-M-A), or *interruptions*, i.e., sequences in which the speakers interrupt each other (A-D-B and B-D-A). The *Speaker Alternation Rate* (SAR) represents the number of speaker alternations per minute. The speaker alternation rate corresponds to a “turn rate” in traditional CA. Moreover, I define an *Interruption Rate* as the number of interruptions per minute. A *pause* is defined as a phase of mutual silence between the talk spurts of the same speaker (i.e., A-M-A and B-M-B), and *non-interruptive double talk* is defined as the event of double talk occurring without ending up in an interruption (i.e., A-D-A and B-D-B). These events are illustrated in Figure 3.4.

3.3.3 Impact of Transmission Delay on Conversational Structure

Throughout the remainder of this thesis, I focus on the communication between two remote user locations, mediated by a transmission system that introduces a considerable amount of transmission delay. The delayed transmission of utterances

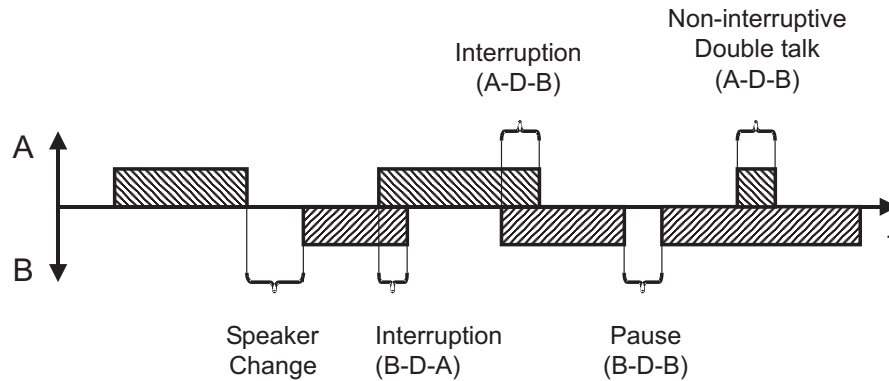


Figure 3.4: Illustration of conversational events.

results in different conversational patterns at the locations of speakers A and B. Thus, I distinguish the respective patterns at side A and at side B. Figure 3.5 depicts this issue.

The upper part of the figure shows the conversational pattern at speaker A's side, and the lower part of the figure depicts the pattern at speaker B's side. In between, the talk spurts of the speakers are delayed by the transmission system³. The first talk spurt of speaker A, i.e., spurt⁴ $A1_A$ is transmitted to side B, and after a phase of mutual silence, speaker B responds (spurt $B1_B$), resulting in a speaker change. However, when spurt $B1_B$ is received at speaker A's side, i.e., spurt $B1_A$, speaker A has already started to talk (spurt $A2_A$) and is eventually interrupted by B. Note that the time B took to respond to spurt $A1_B$ is increased at speaker A's side, i.e., the time period from the end of spurt $A1_A$ to the beginning of spurt $B1_A$ compared to the duration between the end of spurt $A1_B$ and the beginning of spurt $B1_B$. This increase equals the sum of the individual one-way absolute delays (which are equal in our example). At speaker B's side, the delayed spurt $A2_B$ results in an interruption causing speaker B to stop talking (end of spurt $B2_B$). After a while, speaker B interrupts speaker A (spurt transition $A2_B-B2_B$). Note that at speaker A's location, the delayed spurt $B2_A$ does not lead to an interruption. At this point, the spurts $B1_A$ and $B2_A$ are parted by a pause of speaker B.

For convenience, I distinguish two types of interruptions: in an *active interruption*, a participant interrupts the speaker who is currently talking. In contrary, a *passive interruption* denotes the event of being interrupted by another participant while

³Figure 3.5 illustrates the case that the delay from side A to side B equals the delay from side B to side A. In a real-world VoIP system, however, the respective one-way absolute delays may differ due to the fact that the speech packets may be routed through the Internet on different paths, and due to different playout buffer algorithms implemented in the users' terminals

⁴The talk spurts are labeled as follows. The first letter denotes the speaker who contributed the corresponding utterance, the digit represents the talk spurt number, and the subscripted letter identifies the side at which the talk spurt occurs.

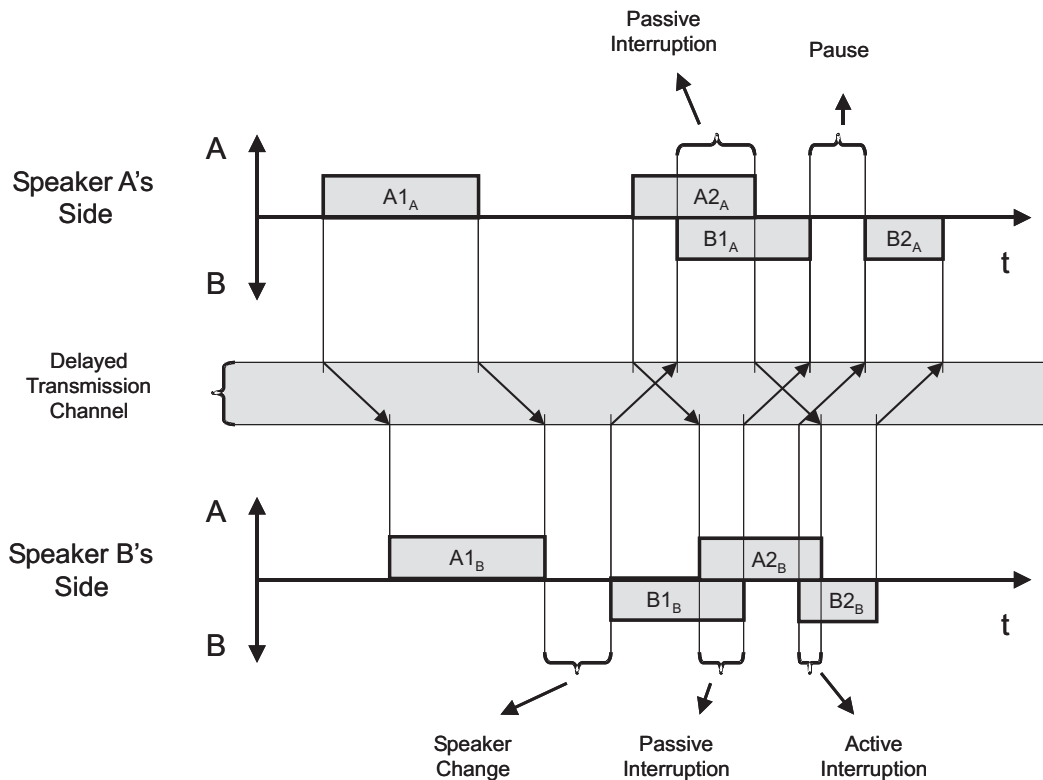


Figure 3.5: The transmission delay results in a considerable shift of the talk spurts.

talking myself. Both types of interruptions are illustrated in Figure 3.5.

3.4 Models for Conversational Interactivity

So far, I have explored parameters which describe a two-way conversation. In this section, I present three models for conversational interactivity: The speaker alternation rate, a conversational temperature model, and a model based on the entropy of speaker turns. Later in this chapter, I will apply these models to recorded conversations and compare their performance.

3.4.1 Speaker Alternation Rate

Taking my definition of interactivity (cf. Section 3.2.1) and the P-CA parameters into consideration, the simplest metric for conversational interactivity is the speaker alternation rate (SAR). As described in Section 3.3.2, the SAR represents the number of speaker alternations, i.e., A-M-B, B-M-A, A-D-B, and B-D-A, per minute (the patterns A-M-A, B-M-B, A-D-A, and B-D-B do not represent speaker alternations

but considered as pauses and non-interruptive double talk, as described in Section 3.3.2). A low SAR corresponds to low conversational interactivity and a high SAR corresponds to a highly interactive conversation. A major advantage of the speaker alternation rate is that, given the conversation pattern (talk spurts), it can simply be calculated by counting the speaker alternations and dividing them by the duration of the call.

3.4.2 Conversational Temperature

The interactivity metric presented in this section⁵ is based on the conversation model described in Section 3.3.1. For each state $I \in \{A, B, M, D\}$, let t_I be the average sojourn time spent in these states, with $t^* = \max \{t_A, t_B, t_M, t_D\}$ being their maximum. In this section, I derive a scalar parameter $\tau = \tau(t_A, t_B, t_M, t_D)$ as a function of these mean sojourn times, leading to a simple but efficient and intuitive one-dimensional metric for describing conversational interactivity.

While the speaker alternation rate (SAR, cf. Section 3.4.1) provides an explicit definition of a metric for conversational interactivity, in this section, I adopt the implicit approach presented in [88] by introducing three descriptive Desirable Properties which characterize central features of conversational interactivity.

Desirable Property I (Limiting Behavior):

$$\lim_{t^* \rightarrow \infty} \tau(t_A, t_B, t_M, t_D) = 0 \quad \text{and} \quad (3.1)$$

$$\lim_{t^* \rightarrow 0} \tau(t_A, t_B, t_M, t_D) = \infty. \quad (3.2)$$

Desirable Property I suggests that a conversation is *not* interactive at all if either A or B are speaking all the time, no one is speaking at all, or both speakers are simultaneously active all the time (cf. Equation 3.1). On the other hand, the case of high interactivity corresponds to state sojourn times being short as represented in Equation 3.2. Note that I do not consider the case of $\min(t_A, t_B) = 0$, resulting in $\tau = 0$, as a conversation because this case implies that one speaker is not talking at all.

Desirable Property II (Normalization):

The standard conversation has reference interactivity τ^{Ref} .

Desirable Property II scales our interactivity metric alongside an abstract “standard conversation” with sojourn times averaged over many different conversation samples. This step allows for comparability among different conversation scenarios.

⁵Parts of this section have previously been published in [35] and have been revised.

From ITU-T Rec. P.59 (cf. Section 3.2.2), I have derived the mean sojourn times based on a set conversations in English, Italian and Japanese which represent the “standard conversation”: $t_A^{Ref}=t_B^{Ref}=0.78$ s, $t_M^{Ref}=0.51$ s, $t_D^{Ref}=0.23$ s.

Desirable Property III (Monotonicity):

$$\frac{\partial \tau}{\partial t_I} < 0 \quad \forall I, t_I \in \mathfrak{R}^+ \quad (3.3)$$

Finally, Desirable Property III implicates monotonicity of τ in the sense that decreasing sojourn time in one of the states leads to an increase of the interactivity metric and vice versa.

In our daily language use, we sometimes describe conversations by means of “heat”, e.g., “hot discussions”. This leads us to a class of problems well-known from statistical thermodynamics [103]: Imagine a single particle moving within a quantum well bordered by potential walls in which the particle continuously tries to jump over one of the walls. An important parameter describing this system is its temperature T , and the success rate λ of the jumping particle depends on T according to

$$\lambda = \nu \cdot \exp\left(-\frac{\Delta E}{kT}\right). \quad (3.4)$$

Here, ΔE describes the height of the potential walls, ν is the oscillation frequency of the particle, and k is known as “Boltzmann’s constant”.

Now I interpret the 4-state conversation model as depicted in Figure 3.3 as a Continuous Time Markov Chain (CTMC) and imagine a “state token” hopping between the states. Then, the mean sojourn time of the state token in state I of the CTMC is exponentially distributed [93] with parameter

$$\lambda_I = \frac{1}{t_I}, \quad (3.5)$$

where λ_I represents the total transition rate out of state I . Applying the thermodynamic concept of the jumping particle to our conversation model, I obtain the following relation:

$$\lambda_I = \frac{1}{t_I} = \nu_I \cdot \exp\left(-\frac{\tau^{Ref}}{\tau}\right), \quad (3.6)$$

Comparing (3.6) and (3.4) suggests an interpretation of the interactivity metric τ in terms of a temperature, the so-called “conversational temperature” as proposed in [88]. τ represents the conversational temperature, and τ^{Ref} constitutes the conversational temperature of a standard conversation. From here, it is left to determine the parameter ν_I . From Desirable Property II and (3.6) I learn that a standard conversation, with sojourn time t_I^{Ref} in state I at a reference temperature τ^{Ref} , leads

to

$$\frac{1}{t_I^{Ref}} = \nu_I \cdot \exp\left(-\frac{\tau^{Ref}}{\tau^{Ref}}\right) = \frac{\nu_I}{e}. \quad (3.7)$$

Now, ν_I can easily be determined:

$$\nu_I = \frac{e}{t_I^{Ref}}. \quad (3.8)$$

Thus, the mean sojourn time in one of the four states is determined by

$$t_I = t_I^{Ref} \exp\left(\frac{\tau^{Ref}}{\tau} - 1\right). \quad (3.9)$$

Equation 3.9 describes $t_I(\tau)$ regarding a standard conversation, and is thus based on fixed relations among the individual t_I s. The relations among the mean sojourn times of conversation measured in the real world, however, are usually not in accordance with the standard conversation. Hence, I estimate the global, single scalar, conversational temperature τ of the conversation using least squares estimation resulting in the estimated temperature $\hat{\tau}$ as shown in Equation 3.10.

$$\hat{\tau} = \underset{\tau}{\operatorname{argmin}} \sum_I (t_I^{Ref} \cdot \exp\left(\frac{\tau^{Ref}}{\tau} - 1\right) - t_I)^2 \quad (3.10)$$

Finally, I have to quantify τ^{Ref} from Desirable Property II. For the sake of simplicity, in the remainder of the thesis, I choose the conversational temperature⁶ of a standard conversation to be “*room temperature*”, i.e., 21.5°.

The main benefit of the conversational temperature, as compared to the speaker alternation rate, is the use of a standard conversation to calibrate the four individual sojourn times.

3.4.3 Entropy Model

Up to now, I have focused on a two-way communication using the four-state model as given in Figure 3.3. This model is intuitive and simple to handle, however, I cannot use it for analyzing multi-party conversations incorporating more than two participants. Hence, I use a speaker *turn* model instead of a *talk spurts* model. Each time that one of N speakers starts to talk, she gets the floor. Her turn ends as soon as another speaker starts to talk. This principle is illustrated in Figure 3.6. Turns are assigned in two ways: either a pause (mutual silence) is assigned to the previous talk spurt or a turn ends as soon as another speaker interrupts.

⁶Matlab code for the calculation of the conversational temperature from the sojourn times of a given conversation is provided in the Appendix in Section C.1.

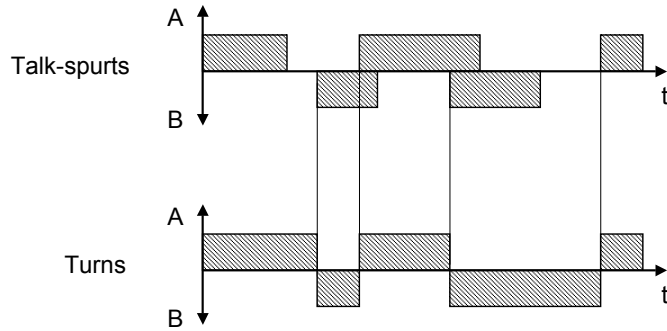


Figure 3.6: Assignment of turns from talk spurts. The turn assignment illustrated in the lower part of the figure has been used for my measurements.

The corresponding turn-model is depicted in Figure 3.7. The major differences to the 4-state model are that states M and D are omitted and that the turn-model is able to cope with conversations involving more than two speakers. The essential model parameters are the state probabilities p_k and the mean turn durations \bar{t}_k .

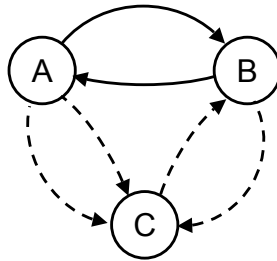


Figure 3.7: Turn model.

Based on this turn-model, I can derive a metric for interactivity that is motivated by an information theoretic approach. The entropy, as defined by Shannon [97]

$$H(x) = \sum_{k=1}^N (-p_k * \text{ld}(p_k)) \quad (3.11)$$

denotes the “uncertainty” or “non-predictability” of a random event x with states $1..N$. In the case of conversations, p_k denotes the state probability of speaker k speaking. I apply the concept of entropy to the context of conversational interactivity and define an entropy-rate τ_e as

$$\tau_e = \frac{1}{t_T} (-p_A * \text{ld}(p_A) - p_B * \text{ld}(p_B)), \quad (3.12)$$

where \bar{t}_T represents the mean overall turn duration of the conversation, and p_A and p_B denote the probabilities that, at any given moment in time, speakers A and B talk, respectively. For instance, $\tau_e = 1 \text{ bit/s}$ when both speakers take turns at equal rates assuming a mean overall turn duration of $\bar{t}_T = 1 \text{ s}$.

The calculation of the turn probabilities p_i for N speakers is presented in Equation 3.13, where $\bar{t}_{T,i}$ and $\bar{t}_{T,k}$ denote the mean turn duration for speaker i and k , respectively.

$$p_i = \frac{\bar{t}_{T,i}}{\sum_{k=1}^N \bar{t}_{T,k}}, \quad (3.13)$$

A major advantage of the entropy model is that it can easily be extended to conversations of $N > 2$ speakers as shown in Equation 3.14. However, in this thesis, I focus on two-way conversations⁷.

$$\tau_e = \frac{1}{\bar{t}_T} \sum_i^N -p_i * \text{ld}(p_i) \quad (3.14)$$

After presenting the related work, introducing the concept of P-CA and identifying three metrics for conversational interactivity, in the next section I will describe the experimental environments of the user tests I have carried out.

3.5 Experiment 1

3.5.1 Objective

Previous studies using the SCT scenarios (cf. Section 3.2.4) have shown that one-way transmission delays of up to 1 s only had a minor effect on the users' opinion (cf. [72, 83]). Thus, we⁸ have developed interactive Short Conversation Test scenarios (iSCT scenarios) at the Institute of Communication Acoustics (IKA) at Ruhr-University Bochum in Germany which were expected to result in more interactive conversations. The goals of Experiment 1 are two-fold: Firstly, we compare the SCT and iSCT scenario by analyzing their conversational interactivity (Experiment 1a). Secondly, we study the iSCT scenario in more detail at different conditions of transmission delay (Experiment 1b).

3.5.2 Measurement Setup and Test Procedure

This study is based on conversations recorded during a speech quality test carried out in office rooms at the IKA. The laboratory setup for these tests is depicted

⁷Matlab code for the calculation of the conversational entropy rate is provided in the Appendix in Section C.2.

⁸Note that parts of this work has jointly been carried out with my colleague Alexander Raake at the Institute of Communication Acoustics at Ruhr-University-Bochum.

in Figure 3.8. We have used a line simulation tool that emulates most of the impairments that occur in PSTN-, ISDN-, and VoIP-networks. The tool was developed at IKA based on the description of network parameters given in the E-model (see Section 3). On each participants side, i.e., the left and right end of the figure, a telephone handset has been used that provided was adjusted to the characteristics of an “Intermediate Reference System” (IRS, ITU-T Rec. P.48 [42]). The line simulation tool provided a symmetric setup in order to be able to set the same conditions for each participant. In Figure 3.8, the triangles represent filters or programmable attenuators, and the rectangles denote delay lines (for T , T_a and T_r), external codecs and the channel bandpass (BP) filter. The setup includes a path that simulates talker echo, represented by the factors Le (echo level) and T (echo delay time). The absolute delay is introduced in the Ta blocks. According to Rec. G.107 [61], Ta represents the “Absolute one-way delay in echo free connections”. As to compensate for the delay resulting from the VoIP interface, in our setup, the values of $Ta1$ and $Ta2$ were set 40 ms below the intended absolute delay. Throughout this study, the delay conditions were symmetric, i.e., $Ta1 = Ta2$, in order to obtain consistent results regarding the users’ quality ratings. However, as I have pointed out in Section 3.3.3 already, in a real-world VoIP system, the individual absolute delays of each direction may differ from each other due to different network and terminal conditions. For emulating VoIP transmission, a component was integrated in the simulation tool which is capable of G.711 and G.729 speech coding and dropping speech packets based on given packet loss traces

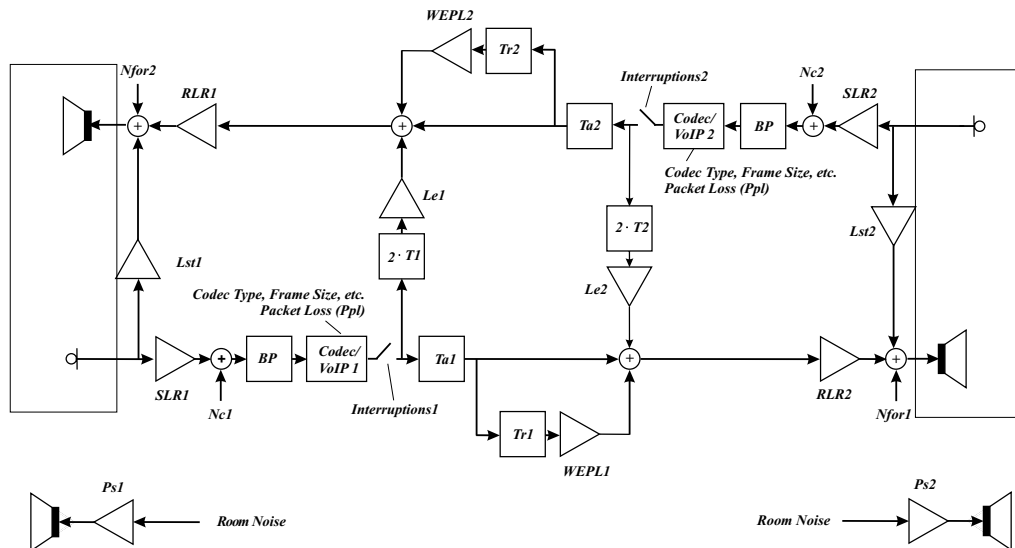


Figure 3.8: Measurement setup at the Institute of Communication Acoustics (IKA, from [59]).

in both directions (cf. “Codec/VoIP” block in Figure 3.8). For a more detailed description of the line simulation tool, I refer to [72, 87]. The parameters of the E-model and their default values are listed in Table D.1 of Appendix D.

The first part of my study (*Experiment 1a*) is based on recordings of conversational speech quality tests concerning the impact of noise and bursty packet losses. In this test, the subjects were asked to accomplish a set of SCT tasks. I have selected the first of 15 test conversations in which every pair of test persons accomplished the same task, i.e., ordering a pizza. This conversation was held under clean conditions (ITU-T G.711 codec, no noise, no packet losses, 66 ms of delay).

In previous studies, the delay has only slightly affected the perceived quality when using SCTs as conversation scenario. Therefore, we introduced a new scenario which is expected to yield more conversational interactivity, and thus increased impact of delay on the quality. The interactive Short Conversation Test (iSCT) scenario is based upon the rapid exchange of numerical and lexical data, such as weather data and email addresses. One of the data items was missing at each side requiring additional turn-taking for clarification (see the example of the iSCT scenario in Section B.3 of the Appendix, where items “Salzburg” and “Klagenfurt” were not available at both sides). We included this feature in order to prevent the users from applying a strategy that results in semi-duplex (“walkie-talkie”-like) conversation in which strict turn-taking is performed (only one speaker speaking at a time). To increase interactivity during the tests involving the iSCT scenarios, we aimed for lowering the conversation discipline by selecting pairs of subjects who knew each other well, by instructing them to address themselves using their first names, and to perform the tasks as quickly as possible. The iSCTs lead to comparable and balanced conversations of higher interactivity compared to the standard Short Conversation Test (SCT) scenarios [72]. An iSCT example can be found in the Appendix (Section B.3).

In the second part of our study (*Experiment 1b*), I focus on the impact of transmission delay on the conversational structure and on perceived speech quality⁹. In Experiment 1b, the conversational tests consisted of VoIP connections using the ITU-T G.729 codec with different bursty packet loss rates (0%, 3% 5% and 15%, packet size was 20 ms) combined with transmission delay of 60 ms, 400 ms, 600 ms and 1000 ms¹⁰. The test conditions are summarized in Table 3.2 and were randomized for each pair of subjects. The packet size was 20 ms and the internal packet loss concealment algorithm of the G.729 codec has been used. The general settings of the line simulation tool in Experiment 1 are given in Table 3.3. At the beginning of the test, the test persons were exposed to four different conditions

⁹Note: The impact of delay on speech quality will be accounted for in Chapter 4.

¹⁰The combined loss and delay impairment has been studied by Raake [84].

#	Codec	Ppl [%]	Ta [ms]
1	G.711	0	66
2	G.729	0	1000
3	G.729	0	60
4	G.729	3	60
5	G.729	5	60
6	G.729	15	60
7	G.729	0	400
8	G.729	3	400
9	G.729	5	400
10	G.729	15	400
11	G.729	0	600
12	G.729	3	600
13	G.729	5	600
14	G.729	15	600

Table 3.2: Test conditions of Experiment 1. Note that for the present study, I have only considered non-packet loss conditions (*Ppl*... Percentage packet loss).

(including 15 % packet loss and 600 ms delay) when reading a written dialogue taken from a book [70]. In my investigation, I restrict myself to scenarios without packet losses in order to explore the pure delay effect on interactivity. However, the entire number of test conditions are given in Table 3.2 for completeness sake.

As to be able to compare the SCT and iSCT scenarios at clean conditions (G.711 codec, no packet losses, 66 ms of delay), I analyzed the structure of one specific iSCT task (rapid exchange of weather data) with the SCT task described above.

In order to be able to explore the interactivity of the conversations held in Experiment 1, I have directly recorded the microphone signals of both speakers on a stereo file, and manually coded the talk spurts in order to reach high accuracy in the derived parameters. Since the microphone signal at each speakers side was recorded simultaneously, the talk spurts were shifted in time according to the absolute delay as to obtain the conversational patterns as perceived at the individual participant's side.

11 pairs of naïve¹¹ German speaking test persons were paid for taking part in this experiment (11 female, 11 male). The test persons were aged 18-30, the average age was 23.7. Note that the pairs of test persons knew each other.

¹¹Naïve test persons are subjects who have not attended a similar quality test before.

Codec	G.711 (20 ms) G.729 A (20 ms)	Ta	66 ms (G.711) 60 ms (G.729)
Ie	0 (G.711) 11 (G.729)	T	33 ms (G.711) 30 ms (G.729)
VAD	disabled	Tr	0 ms
SLR	13 dB	WEPL	110 dB
RLR	2 dB	TELR	65 dB
STMR	15 dB	Nc	-70 dBm0p
LSTR	16 dB	Nfor	-64 dBmp
Ds=Dr	1	Ps=Pr	35 dB(A)

Table 3.3: Default parameter settings of Experiment 1. A complete list of the parameters of the E-model and their default values as given in ITU-T Rec. G.107 [61] are listed in Table D.1 of Appendix D.

The results of this experiment are presented and discussed in Section 3.7.1.

3.6 Experiment 2

3.6.1 Objective

As I have presented in Section 3.2.4, a variety of scenarios has been used in past conversational quality tests. However, the conversational structure of a variety of individual scenarios has not been described and compared so far. The main goal of this study is to investigate the differences between the scenarios with regard to the conversational structure and interactivity at different delay conditions. For this purpose, I selected four conversation scenarios and carried out a user test for data collection¹².

3.6.2 Selection of Conversation Scenarios

Before describing the test scenarios I have chosen for our study, I distinguish the terms *scenario* and *task*. As already stated in Section 3.2.4, I use the term “scenario” as a generic term for a set of similar tasks which result in comparable conversational structures. In turn, I define a conversation task as one implementation of a given conversation scenario.

¹²Compared to Experiment 1, this study compares a larger number of test scenarios and incorporates test subjects of a larger range of age groups.

- *Random Number Verification (RNV)*. This scenario requires the rapid verification of a given set of random numbers. The test persons are asked to alternately verify the numbers either in rows or in columns. This type of scenario was taken and adapted from Kitawaki’s study [68]. It is expected to be highly interactive and to yield high impact of transmission delay on perceived quality. An RNV sample task is given in Section B.1.
- *Short Conversation Test (SCT)*. The SCT represents today’s standard scenario in conversational speech quality assessment [72] and is based on tasks like ordering a pizza or booking a hotel room. The SCTs result in natural, balanced conversations of about 2–3 minutes. Previous tests suggest that SCTs do not lead to sufficient conversational interactivity to generate significant impact of delay on perceptual quality. An SCT sample task is given in Section B.2.
- *Asymmetric Short Conversation Test (aSCT)*. This new scenario is a variation of the iSCT described in Section 3.5.2. Like the iSCTs, the aSCTs comprise tasks which require the rapid exchange of numerical or lexical data. In the aSCT tasks however, the called person is given all of the information, while the calling person needs to request it. Thus, the structure of the resulting conversations is expected to be asymmetric by means of the speech activity of the two participants. An aSCT sample task is given in Section B.4.
- *Free Conversation (FC)*. The free conversation scenario results in “everyday” conversations of about seven minutes based on given topics like the organization of a party for a friend. In this kind of scenario, the structure of the conversations is not strictly pre-determined by a given task, but rather driven by the conversation behavior of the test subjects. I consider free conversations the most realistic scenario in our setup. An FC sample task is given in Section B.5.

3.6.3 Measurement Setup and Test Procedure

I have carried out user tests at the facilities of the Telecommunications Research Center Vienna (ftw.) in order to collect speech material for the comparison of the scenarios. The measurement setup is illustrated in Figure 3.9. I set up a VoIP system in quiet ftw. office rooms, consisting of two PC-VoIP-clients which are connected to an IP-testbed-PC. The testbed-PC allows for very accurate emulation of voice-packet delay. The delay emulator runs on Real-Time Linux and has been developed at ftw. [37]. In addition to delay emulation, the testbed-PC facilitates the routing between the two clients. As the user-interface, I used the “Gnome-meeting” VoIP-software-client running on a Linux-PC.

I used headsets (Plantronics) as electro-acoustic interface which help avoiding acoustical echoes. The VoIP-client was set to use G.711 speech coding [41] and did not

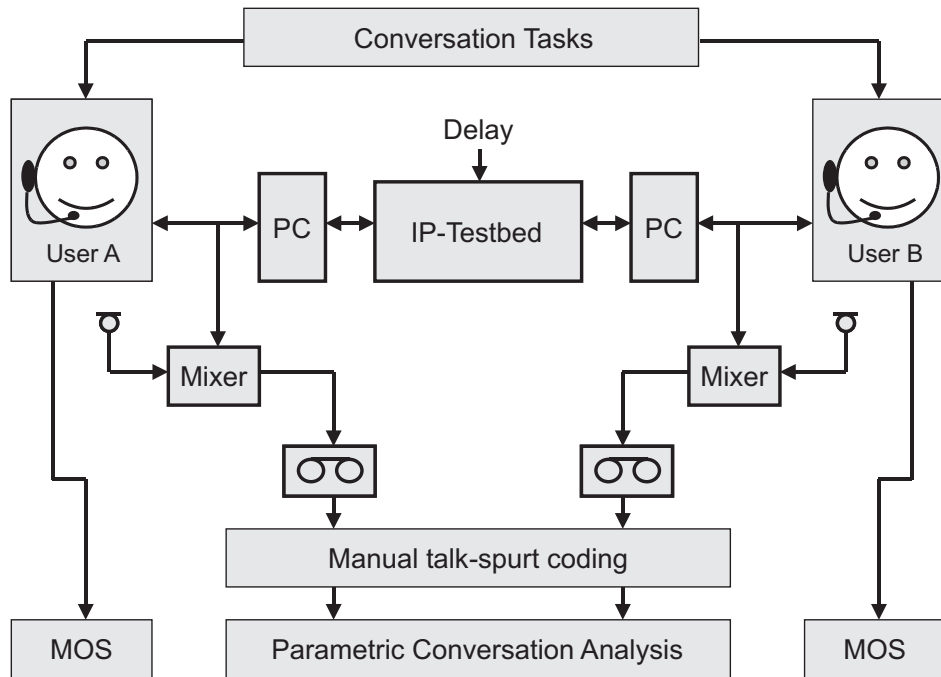


Figure 3.9: Test setup of Experiment 2.

provide IRS-filtering of the speech signals. As absolute delay conditions, I have chosen values of 200, 350, and 500 ms. The lower limit of 200 ms was pre-determined by the entire transmission chain and approximately equals the absolute delay of a mobile-to-mobile connection (GSM). The other conditions are chosen to just about meet (350 ms) and exceed (500 ms) the limit for maximum delay of 400 ms that is acceptable for users as given in ITU-T Recommendation G.114 [56]. The test conditions are illustrated in Table 3.4.

Each scenario was represented by three tasks. Within the RNV scenario I used three different lists of numbers and varied the way of verifying them, i.e., vertically vs. horizontally. The SCT scenario included tasks like ordering a pizza, booking a hotel room, or looking for a flat at a real estate agent. The aSCT scenario consisted of tasks like specifying data of furniture in a depot, codes of vehicles of a car rental company, and weather data. In the free conversation scenario, the test persons were instructed to organize a party as a surprise for a friend, talk about the latest vacation, or plan a bank robbery. I have combined each task with each delay condition resulting in a total of 12 conditions. For each pair of test subjects the conditions were randomized.

Before the actual testing, I have instructed the subjects about the purpose of the experiment, i.e., measuring the interactivity and quality of different scenarios using VoIP, both in written form and orally. The test persons were instructed about the test procedure and the rating scales in use (see also Section 4.4.2). Before the actual

#	Scenario	Ta [ms]
1	RNV	200
2	RNV	350
3	RNV	500
4	SCT	200
5	SCT	350
6	SCT	500
7	aSCT	200
8	aSCT	350
9	aSCT	500
10	FC	200
11	FC	350
12	FC	500

Table 3.4: Test conditions of Experiment 2. Ta represents the absolute delay.

test calls, the subjects made two test calls in order to get used to the system and test procedure. In total, the subjects made a total number of 14 telephone calls. After each call, they filled out a questionnaire regarding the overall quality (cf. the box “MOS” (Mean Opinion Score) in Figure 3.9), interactivity related questions and how realistic the task would be. Since the quality aspects are considered in Chapter 4, these questions are described in detail there. In the first two calls, an aSCT task at a delay of 200 ms and an RNV task at a delay of 500 ms, the subjects could get used to the setup and procedure. Then they took turns calling each other and fulfilling the given tasks.

At each participant’s side, the conversations were recorded on a PC. I recorded the loudspeaker signal of the headset and I used an external microphone (AKG C1000) in front of the subjects of which the signal was easier to capture than the signal of the headset microphone. These two signals were mixed to a stereo signal (mixer in the setup depicted in Figure 3.9). As a result, the conversations were stored as stereo-files in the WAV-format (16bit, 8kHz, left channel: microphone signal, right channel: headphones signal). Figure 3.10 presents how the conversation recordings were analyzed. I have manually extracted the talk spurt cues in order to obtain accurate cue-lists.

12 pairs of naïve, German speaking test subjects (15 female, 9 male) participated in the experiment who were each paid EUR 15,-. The imbalance between female and male subjects resulted from the fact that it was hard to find test persons at all. The subjects were between 18 and 61 years old (31.0 years in average). 9 subjects were younger than 22 years, 7 subjects were between 22 and 40 years old, and 8 subjects were older than 40. The entire experiment was held in German. The pairs

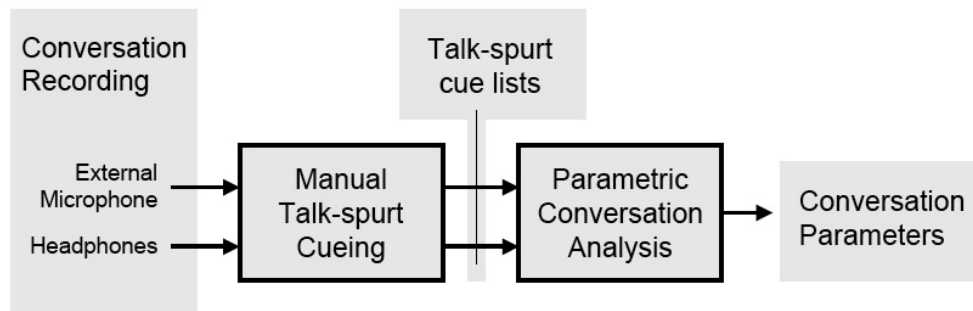


Figure 3.10: Experiment 2. Processing of the conversation recordings.

were of about the same age and knew each other well.

The results of this experiment are presented and discussed in Section 3.7.3.

3.7 Results and Discussion

This section presents the results of Experiment 1 and Experiment 2. I first compare the conversational parameters of SCT and iSCT tasks in the absence of transmission delay (Experiment 1a). Then, I focus on the impact of delay on the conversational parameters and interactivity of the iSCT tasks (Experiment 1b). Further on, I investigate the conversational parameters and interactivity of our selection of conversation tasks (Experiment 2). Note that the parameters derived from basic conversation parameters, such as the conversational temperature, were first calculated per conversation and then averaged. The parameter averages have been calculated based upon the conversation parameters on both sides.

3.7.1 Comparison SCT vs. iSCT (no delay)

Conversational Parameters

I compare the SCT and iSCT scenarios (10 conversations per scenario) and the “standard conversation” given in ITU-T Rec. P.59 [43] by exploring the corresponding state probabilities and sojourn times of the conversational model, as given in Figures 3.11 and 3.12 and Tables 3.5 and 3.6¹³. For the comparison of the SCT and iSCT scenarios, clean PCM encoded connections were analyzed. To increase comparability, one specific SCT scenario (pizza order) was compared to one specific iSCT

¹³Parts of the results of this section have been published in [34].

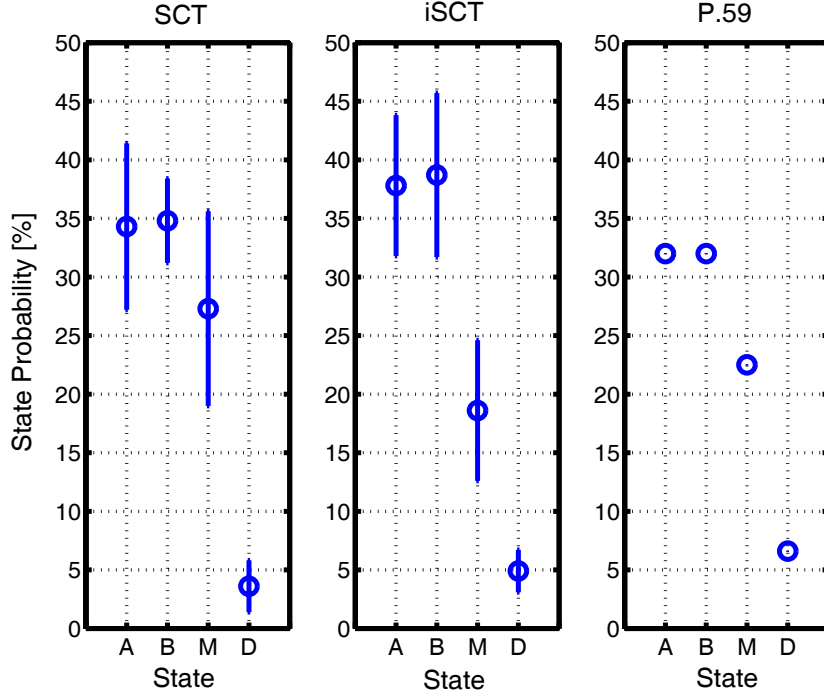


Figure 3.11: State Probabilities [%] of the SCT and iSCT scenarios and ITU-T Rec. P.59.

Scenario	State Probabilities [%]			
	A	B	M	D
SCT	34.3 (7.1)	34.8 (3.6)	27.3 (8.3)	3.6 (2.2)
iSCT	37.8 (6.0)	38.7 (7.0)	18.6 (6.0)	4.9 (1.8)
P.59	35.2	35.2	22.5	6.6

Table 3.5: State probabilities [%] for the SCT and iSCT scenarios and ITU-T Rec. P.59 [43].

scenario (exchange of weather data).

As the most obvious result, both the mean state probabilities and the mean sojourn times for mutual silence differ significantly. While the state probability of state *M* of the SCT scenario is higher than the value given in P.59, the iSCT scenario results in lower amount of mutual silence than the standard conversation. Both the (single-talk) speech activity of the SCT and iSCT scenarios is higher than given in P.59 with a trend of higher probabilities of states *A* and *B* in the iSCT scenario. The

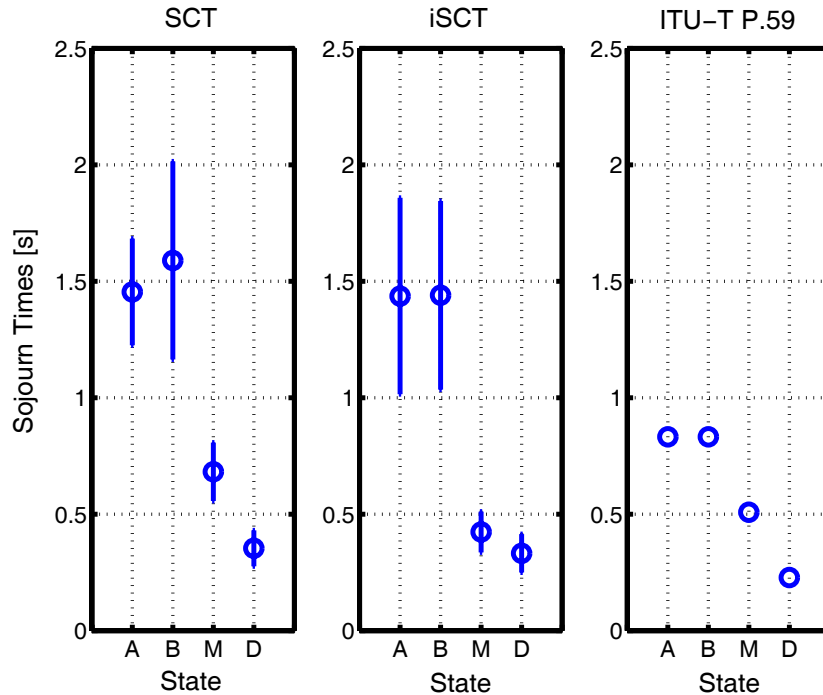


Figure 3.12: Sojourn Times [s] of the SCT and iSCT scenarios and ITU-T Rec. P.59.

Scenario	Sojourn Times [s]			
	A	B	M	D
SCT	1.45 (0.23)	1.59 (0.43)	0.68 (0.13)	0.35 (0.08)
iSCT	1.44 (0.42)	1.44 (0.40)	0.42 (0.09)	0.33 (0.08)
P.59	0.78	0.78	0.51	0.23

Table 3.6: Sojourn Times [s] for the SCT and iSCT scenarios and ITU-T Rec. P.59 [43].

iSCT scenario results in slightly more double talk than the SCT scenario, and the amount of double talk of both scenarios remains below the level of double talk in ITU-T Rec. P.59 [43].

Regarding the mean sojourn times of the three types of conversation, I observe that the single-talk phases of the SCT and iSCT scenarios are about twice as long as in the standard conversation. This difference may result from the fact that in our study the conversations were held in German, whereas the numbers given in Rec. P.59 are based on conversations which were held in English, Japanese, and Italian. Similar to the results of the state probabilities, the mutual silence phases

in the SCT scenario are longer than in P.59, while, on average, the iSCT scenario results in shorter sojourn times for state M than given in the standard conversation.

Table 3.7 presents a comparison of the mean interruption rates (Total Interruption Rate (TIR), Active Interruption Rate (AIR), Passive Interruption Rate (PIR)) of the SCT and iSCT scenarios. The iSCT scenario seems to provoke conversations in which the speakers tend to interrupt each other more often.

Task	TIR	AIR	PIR
SCT	2.15 (1.21)	2.33 (1.24)	1.97 (1.19)
iSCT	3.00 (1.32)	3.55 (1.19)	2.44 (1.23)

Table 3.7: Mean values and standard deviations of the Total Interruption Rate (TIR), the Active Interruption Rate (AIR), and the Passive Interruption Rate (PIR) for both SCTs.

Conversational Interactivity

Figure 3.13 and Table 3.8 present a comparison of the interactivity parameters of the SCT and iSCT scenarios. I observe that the iSCT scenario results in a significantly higher speaker alternation rate than the SCT scenario. Similarly, the iSCT task results in higher interactivity than the SCT scenario in terms of the entropy rate. In comparison, the conversational temperature indicates increased interactivity of the iSCT scenario to a lesser extent. The main reason for this behavior is that the sojourn times of both scenarios are very similar and mainly differ in the values for mutual silence (cf. Figure 3.12). However, the results for conversational interactivity underline the more interactive structure resulting from the iSCT scenario.

Task	SAR	Temperature	Entropy rate
SCT	19.54 (4.28)	13.60 (1.26)	0.40 (0.09)
iSCT	26.18 (5.64)	14.67 (1.84)	0.54 (0.10)

Table 3.8: Mean values and standard deviations of the conversational temperature, and the entropy-rate for the SCT and iSCT scenario.

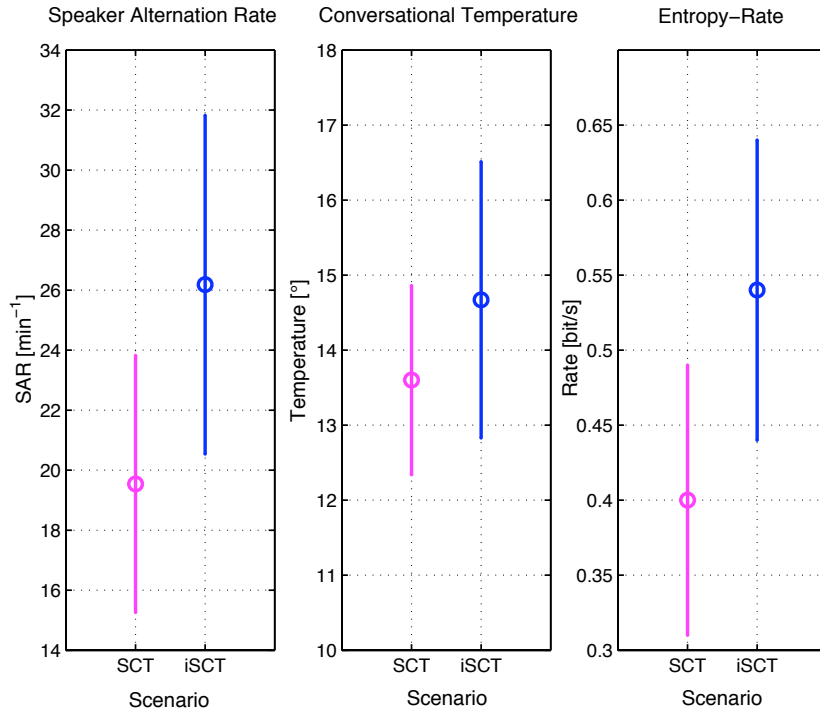


Figure 3.13: Comparison of the metrics for Conversational Interactivity for the SCT and iSCT scenarios.

3.7.2 The Effect of Delay on iSCTs

In this section, I study the effect of one-way transmission delay on the conversational parameters for ITU-T G.729 encoded connections at four different conditions: 60 ms, 400 ms, 600 ms, or 1000 ms¹⁴. In the following, I present the analysis of conversations performed by 7 pairs of test persons (8 female, 6 male) who knew each other.

Conversational Parameters

The evolution of the mean state probabilities and mean sojourn times over delay are illustrated in Figure 3.14. For the analysis, the parameters for speakers A and B were labeled as to assign higher single-talk speech activity among the test persons to speaker B. I observe that the single-talk speech activity (states *A* and *B*) decrease between 400 ms and 600 ms both by means of state probabilities and sojourn times. In turn, the amount of mutual silence (State *M*) increases at these delay conditions. This increase in mutual silence was not unexpected (cf. Section 3.3.3). However, the mean sojourn times of state *M* do not increase by the round-trip time, but by 71 ms between 60 and 400 ms of delay, by 160 ms between 400 and 600 ms of delay, and by

¹⁴Parts of the results presented in this section have been published in [35]

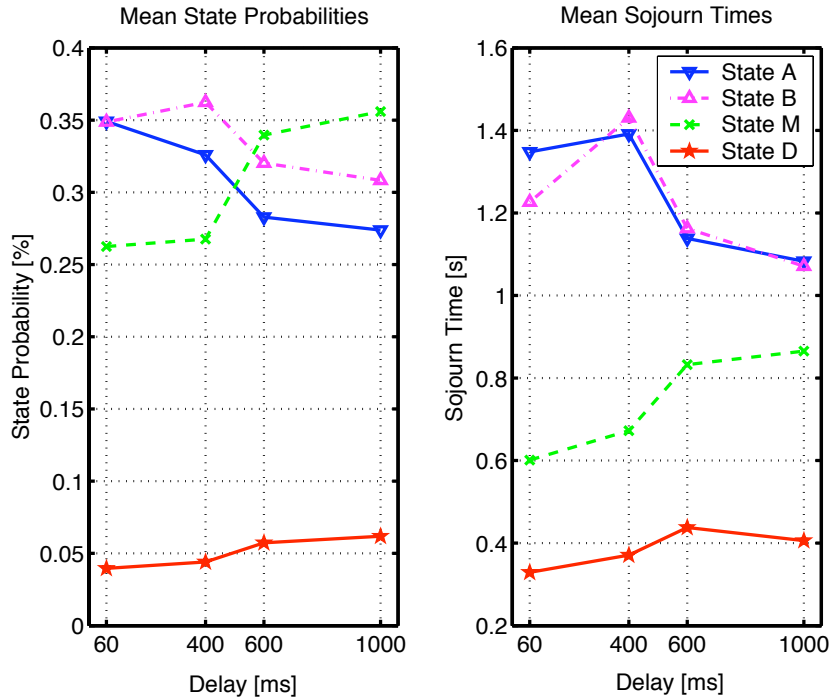


Figure 3.14: Mean state probabilities and mean sojourn times of the iSCT scenario versus delay.

33 ms between delays of 600 and 1000 ms (cf. green dashed line in the right graph of Figure 3.14). From the significant change in conversational parameters between 400 ms and 600 ms, I conclude that the test subjects adapted their conversational behavior at the higher delay conditions. Regarding the double talk performance, I can observe a slight increase of the mean state probabilities and mean sojourn times only.

Figure 3.15 presents the active and passive interruption rates defined in Section 3.3.2. The Active Interruption Rate (AIR) significantly decreases with delay from a mean of 2.73 min^{-1} (standard deviation 1.46 min^{-1}) at a delay of 60 ms to 1.86 min^{-1} (0.78 min^{-1}) at a delay of 400 ms. For delay values above 400 ms, no significant change of the AIR can be detected. These results suggest that at least some of the test persons adapt their conversational behavior at a delay 400 ms. An Analysis of Variances (ANOVA¹⁵, [10]) showed that the Passive Interruption Rate (PIR)

¹⁵Throughout the remainder of this thesis, I present the F and p values of the ANOVA. The F value is calculated dividing the between-groups mean square variance, e.g., across delay conditions, by within-groups mean square variance, e.g., within a particular delay condition. If F is greater than 1, then the between groups variation is larger than the variation within groups, and thus the grouping variable, e.g., the delay condition, shows an effect. The p value denotes the

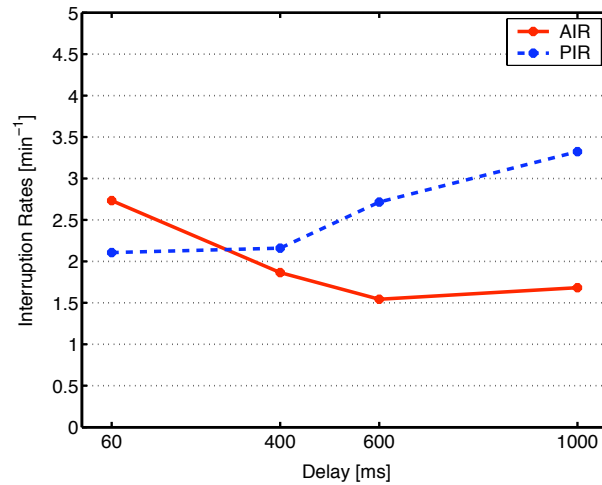


Figure 3.15: Average rates of active interruption (AIR) and passive interruption (PIR).

increases significantly with delay from 2.16 min⁻¹ (0.89) at a delay of 400 ms to 3.32 min⁻¹ (1.50) at a delay of 1000 ms ($F=8.61$, $p<0.05$). I expected this behavior in Section 3.3.2 by pointing out the shuffling of the talk spurts due to the time lag of the transmission.

The mean call durations I have observed are shown in Table 3.9. The call duration values appear to saturate at high delay values. The call durations are affected by both delay ($F=11.14$, $p<0.001$) and the subjects ($F=12.54$, $p<0.001$). I consider the call durations as reasonably long for the subjects to be able to get an appropriate impression of the quality of the connection in use.

Delay [ms]	Call Duration [s]
60	122.35 (40.22)
400	141.26 (44.84)
600	154.25 (41.09)
1000	155.43 (35.81)

Table 3.9: Mean iSCT call durations and standard deviations versus transmission delay.

probability of the “null hypothesis” which represents the case of no significant effect. If $p \leq 0.05$, an effect is considered significant.

Interactivity

In Figure 3.16, I illustrate the results for our measures of interactivity at the four delays under investigation. The speaker alternation rate exhibits a significant impact of delay ($F=6.32$, $p<0.01$), decreasing from 25.18 min^{-1} (5.99 min^{-1}) down to 20.25 min^{-1} (3.23 min^{-1}). From 400 ms delay, no meaningful variation of the speaker alternation rate is indicated. Both the conversational temperature and the entropy-rate tend to decrease between a delay of 60 ms and 400 ms. However, while the results for the conversational temperature show a trend towards an increase in temperature from a delay of 400 ms upwards ($F=3.37$, $p<0.05$), this evolution cannot be confirmed for the entropy rate. From these results, I conclude that iSCT conversations are most interactive at very low delay conditions.

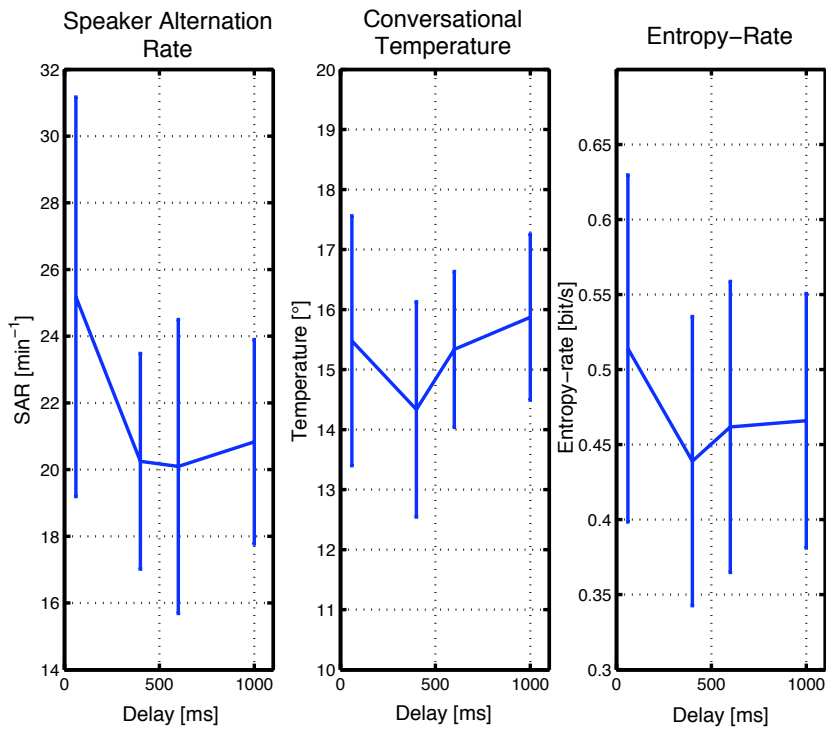


Figure 3.16: Conversational Interactivity metrics vs. Delay for the iSCT scenario.

3.7.3 Comparison of various Conversation Scenarios

In the following, I present the results of Experiment 2, i.e., the conversational parameters and conversational interactivity of the four scenarios I introduced in Section 3.6.2¹⁶.

¹⁶Parts of this section have been published in [30]

Conversation parameters

Based on the four-state conversation model presented in Section 3.3.1, I have derived the mean state probabilities and mean sojourn times for the individual conversation scenarios. Figure 3.17 presents the mean state probabilities at different amounts of delay. As in Figure 3.14, the parameters for speakers A and B were labeled as to assign higher single-talk speech activity among the test persons to speaker B. In the random number verification (RNV) scenario, the mean speech activity of both speaker A and speaker B significantly decreases from 28.7% (4.9% standard deviation) at a delay of 200 ms to 24.7% (4.4%) at a delay of 500 ms. The amount of mutual silence (M) significantly rises from 37.2% (8.0%) at 200 ms to 43.3% (6.8%) at 350 ms and 45.3% (8.4%) at 500 ms ($F=24.14$, $p<0.001$), while the amount of double talk (D) remains stable at slightly above 5% at all delay conditions. The mean

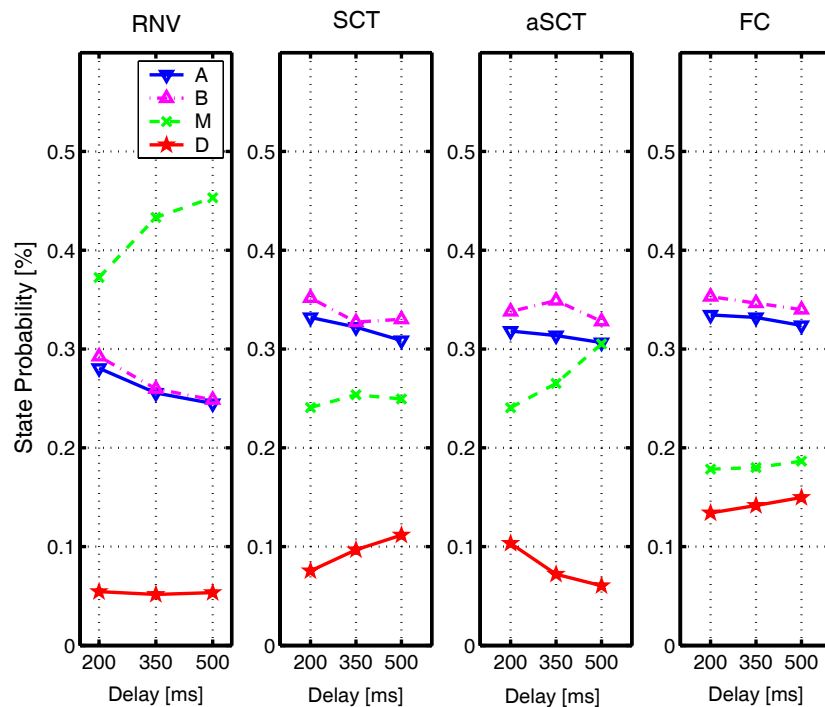


Figure 3.17: Mean state probabilities of four selected conversation scenarios.

state probabilities for states A and B of the SCT, aSCT and FC scenarios are about the same around 32% and do not significantly change with delay. While about 25% of an SCT conversation is filled with mutual silence, this amount is reduced to 18% in the free conversations (FC). The higher amount of mutual silence in SCTs compared to FCs may result from the structure of the scenario. The tasks are fulfilled within a free conversation, but the basic structure is given in the task specification (cf. Section B.2 in the appendix). In contrast, FCs result in a remarkably high amount of double

talk ranging between 13% and 15% (no significant increase). Note that as presented in Section 3.7.1, the double talk probability of a “standard conversation” according to ITU-T Rec. P.59 is 6.59%. The state probabilities of the aSCTs are characterized by a significant increase of mutual silence with delay ($F=10.9$, $p<0.001$), and a corresponding decrease of double talk which is not significant. As shown in the previous section, the amount of mutual silence also increased with delay for the iSCT scenario (cf. Figure 3.14), of which the aSCT scenario was derived from. Although this scenario was intended to provoke asymmetric conversations with regard to the speech activities of speakers A and B, our results show that the conversations were balanced. I have studied the differences between the speech activity of the calling party (who requires information) and the speech activity of the participant who was called and who was expected to talk more. The results of this analysis are shown in Figure 3.18. Surprisingly, I cannot detect any significant difference between the amount of active speech for this scenario. The same holds true for the respective sojourn times.

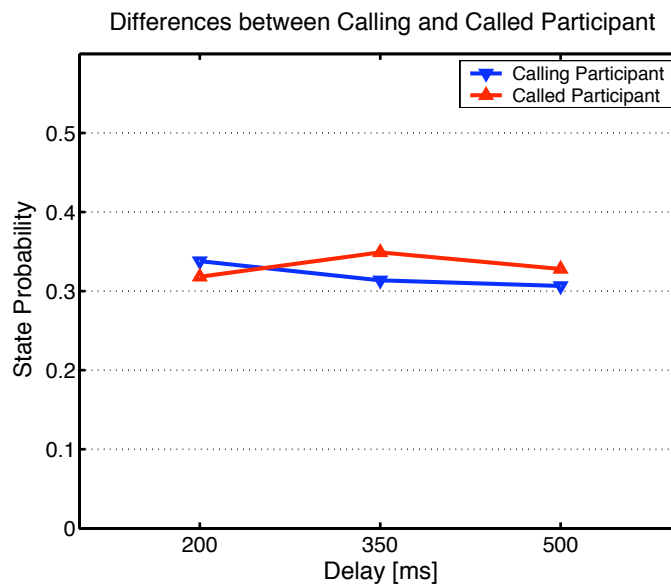


Figure 3.18: Comparison of mean speech activities of the calling participant and the called participant for the asymmetric Short Conversation Test (aSCT) scenario.

Figure 3.19 presents the mean sojourn times of the individual scenarios at different amounts of transmission delay. I can observe similar sojourn times of single talk (A, B) for the SCT, aSCT and FC scenarios. These sojourn times are reduced by half for the RNV scenario. Thus, the talk spurts are significantly shorter for the RNVs. Considering the RNVs, the sojourn times for mutual silence increase from 0.52 s (0.11 s) at a delay of 200 ms to 0.79 s (0.12 s) at 500 ms at a highly significant level

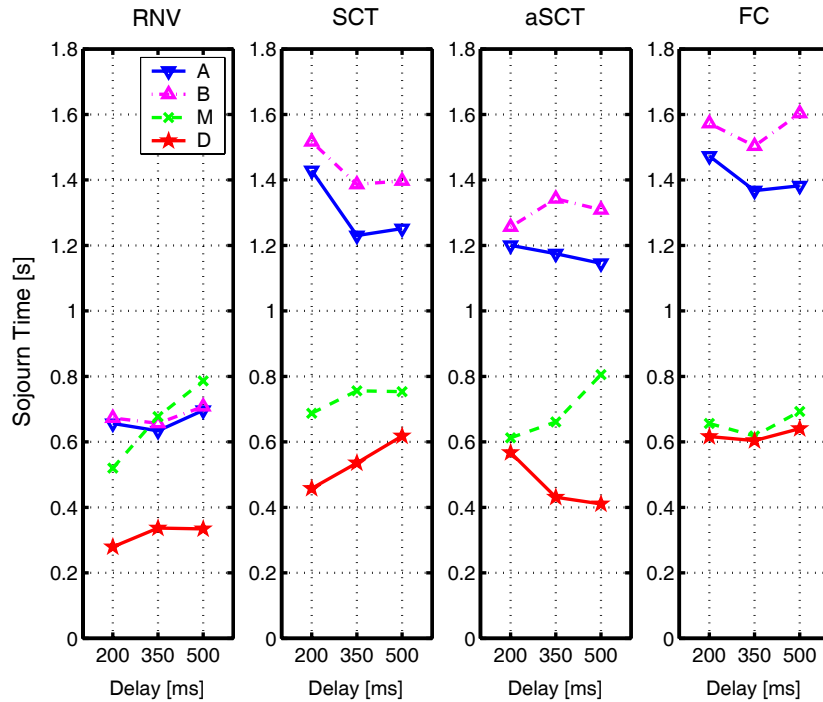


Figure 3.19: Mean sojourn times of four selected conversation scenarios.

($F=92.83$, $p<0.001$). This increase in mutual silence results from the very frequent speaker alternations between the speakers. Note that the difference in mutual silence (270 ms) approximately tends towards the increase in delay (300 ms). In SCTs, the mean amount of double talk time significantly increases from 0.46 s (0.09 s) at a delay 200 ms to 0.62 s (0.26 s) at 500 ms of delay time. In the FC scenario, no influence of delay on the mean sojourn times for mutual silence and double talk is detectable. Regarding the mean sojourn times of the aSCT scenario, only the impact of delay on mutual silence is significant ($F=12.26$, $p<0.001$). While at a delay of 200 ms, the sojourn times of mutual silence and double talk of the aSCTs are similar to those of the FC scenario, the amount of mutual silence increases with delay as mentioned above. The increase in mutual silence may result from the given tasks leading to more structured conversations compared to the free conversations due to the items to be read in the aSCT scenario. The decrease of delay between 200 ms and 350 ms is not significant.

Figure 3.20 presents the results for the active interruption rate (AIR) and the passive interruption rate (PIR). In none of the scenarios the delay seems to significantly influence the interruption rates. As the Figure shows, the result are similar for both rates.

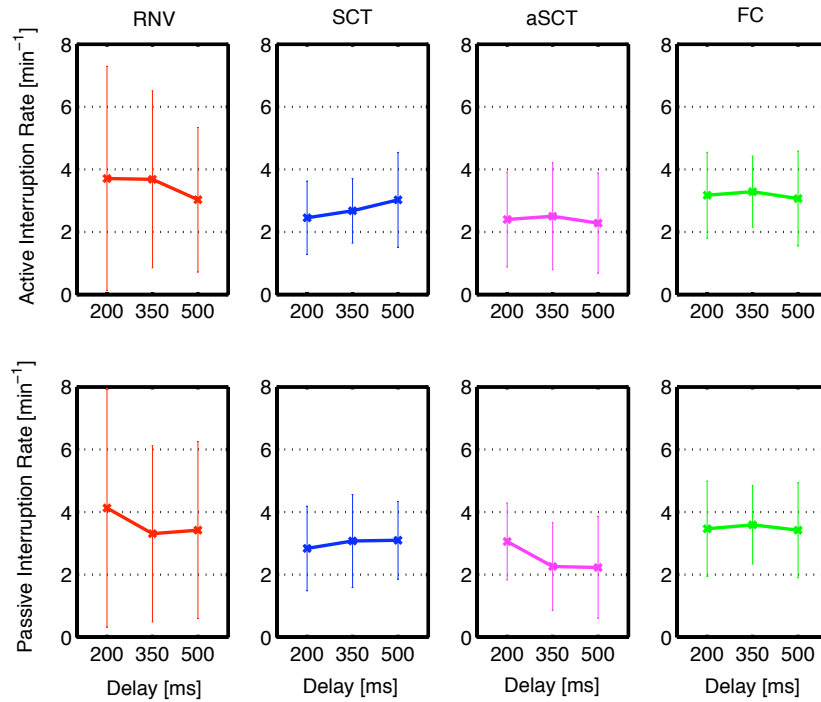


Figure 3.20: Active Interruption Rates (AIR) and Passive Interruption Rates (PIR) for the four conversation scenarios.

The call durations of the individual scenarios at different amounts of transmission delay as presented in Figure 3.21 give an idea about the total lengths of the conversations. While the call durations of the SCT, aSCT, and FC scenarios remain about equal over all delay conditions, the durations for the RNV scenario significantly increases from 102 s (21 s) at a delay of 200 ms to 126 s (22 s) at a delay of 500 ms ($F=9.55$, $p=0.0003$). An SCT task requires an average call duration of 196 s to be fulfilled, an aSCT task lasts about 171 s, and the free conversations take 447 s to be carried out. For the free conversations, the test persons were asked to meet a target duration of seven minutes, i.e., 420 seconds. They were not interrupted by the conductor of the test.

Interactivity

In this section, I present the results of an analysis of the recorded conversations with regard to the measures for conversational interactivity: speaker alternation rate, conversational temperature, and entropy-rate as introduced in Section 3.4. Figure 3.22 shows a comparison of the metrics using the four scenarios at different amounts of delay.

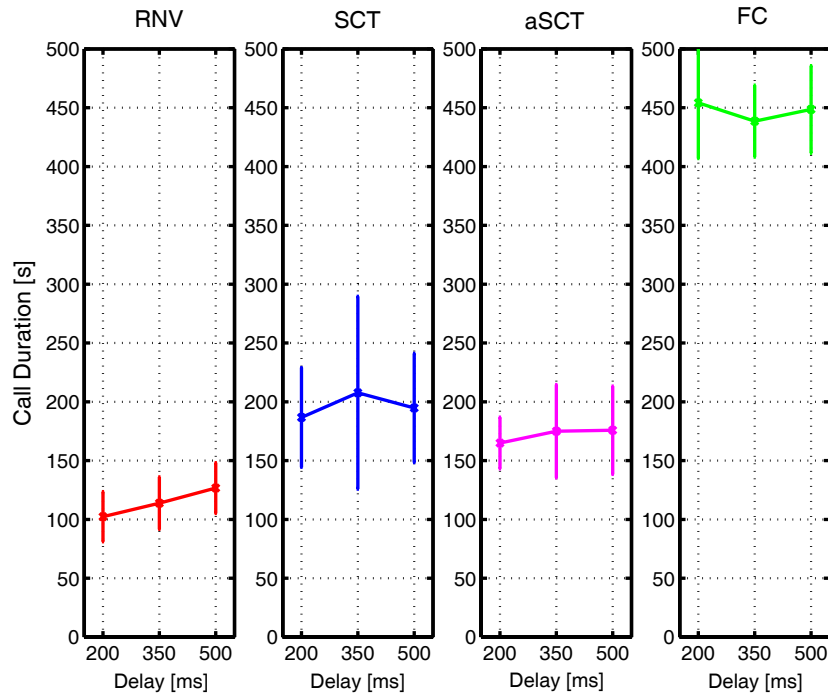


Figure 3.21: The call durations for the four scenarios.

Our first observation is that all three metrics lead to very similar results regarding the conversational interactivity. As expected, the interactivity rapidly decreases with delay in the RNV scenario. The speaker alternation rate decreases from 44.1 min^{-1} (9.8 min^{-1}) at a delay of 200 ms to 34.3 min^{-1} (6.2 min^{-1}) at a delay of 500 ms (effect of delay: $F=33.1$, $p<0.001$). The conversational temperature diminishes from 25.3° (3.8°) at 200 ms to 21.7° (2.8°) at 500 ms, and the entropy-rate declines from 0.85 bit/s (0.14 bit/s) at a delay of 200 ms to 0.66 bit/s (0.12 bit/s) at a delay of 500 ms.

In comparison, the conversational interactivity of the other three scenarios remains about constant over delay conditions. From the values of the speaker alternation rate and the conversational temperature, I may presume that among the three free-conversation like scenarios, the FC tends to be least interactive. In contrary, in terms of the entropy-rate metric, I may conclude that there is not even a trend concerning a ranking of the three scenarios with regard to conversational interactivity.

Karis [65] has studied the turn taking parameters of conversations that lasted 10 minutes over absolute (one-way) delay conditions of 0 ms, 300 ms and 600 ms. Therefore, the numbers of backchannels, speaker turns, and interruptions which were coded. Table 3.10 presents the results of that study. While the number of turns and backchannels do not significantly change with delay, rising delay increased the

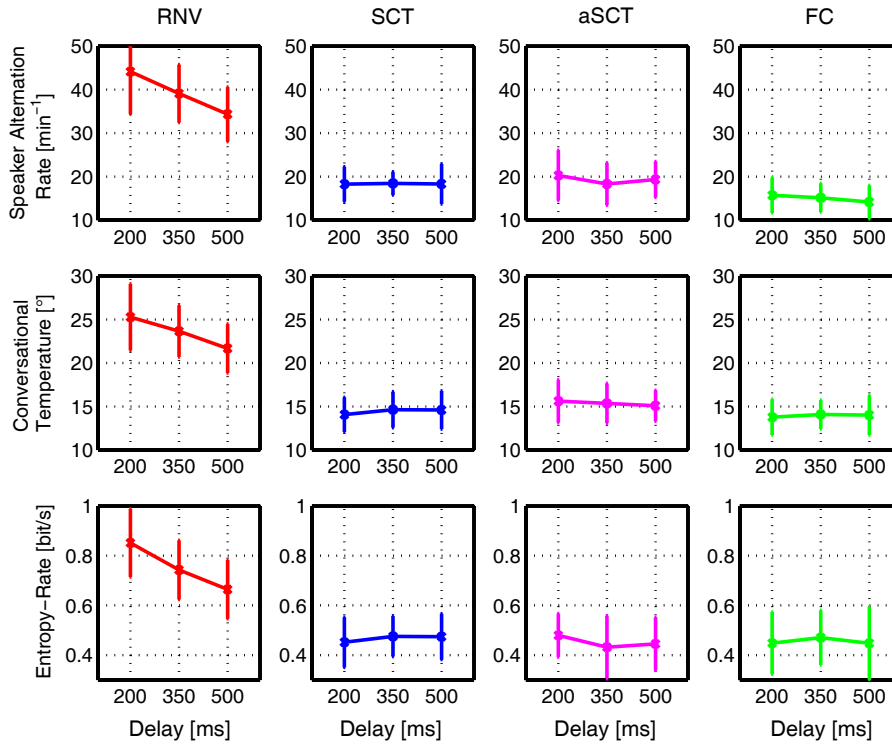


Figure 3.22: Measures for Conversational Interactivity: Mean values of speaker alternation rate, conversational temperature, and entropy rate, and their respective standard deviations.

frequency with which people interrupted each other. Note that the number of turns per minute, slightly decreasing from 15.1 min^{-1} at zero delay to 14.1 min^{-1} at a delay of 600 ms are in accordance with the values of the speaker alternation rates of the FC scenario as shown in Figure 3.22. The speaker alternation rates of the FC scenario are 15.7 min^{-1} at a delay of 200 ms, 15.1 min^{-1} at 350 ms, and 14.1 min^{-1} at a delay of 500 ms.

One-way delay [ms]	0	300	600
No. of Turns	151.3 (23.7)	144.8 (28.5)	141.3 (25.2)
No. of Backchannels	15.5 (10.0)	10.7 (7.4)	17.7 (9.1)
No. of Interruptions	23.5 (8.8)	39.2 (19.1)	47.3 (11.7)

Table 3.10: Effects of delay on turn taking parameters (means and standard deviations of conversations that lasted 10 minutes, from [65]).

3.8 Summary

In this chapter, I have investigated the influence of conversation context on the conversational structure and conversational interactivity by analyzing a number of conversation scenarios. Our aim was to distinguish the scenarios in terms of their interactivity and explore the evolution of the conversational parameters for increasing absolute delay. For the analysis of the structure of conversations, I have introduced the concept of parametric conversation analysis (P-CA) that defines conversational parameters and events. Since a thorough review of existing definitions of interactivity did not lead to a satisfying proposal for a quantitative metric, I developed three measures for conversational interactivity. The Speaker Alternation Rate (SAR) represents the number of speaker alternations per minute. The conversational temperature metric is calculated from the sojourn times of the conversation model. Finally, the entropy rate is based on a speaker turn model and corresponds to the uncertainty about who of the participants is talking.

In two separate user tests, I have collected speech data that was recorded during conversations based on a variety of scenarios. Comparing the interactivity of the scenarios in use, the Random Number Verification (RNV) scenario appeared to be the most interactive scenario. With increasing delay, the interactivity of the RNV scenario decreases due to the strict conversational structure. The free conversation-like scenarios, i.e., Short Conversation Test (SCT), asymmetric Short Conversation Test (aSCT), and Free Conversation (FC) resulted in about the same interactivity. On the one hand, this is a negative result, as it shows that the goal of creating a variety of scenarios exhibiting clearly distinct characteristics could not be reached. On the other hand, this is a positive result as it shows that the conversational interactivity parameters are quite robust for all kinds of scenarios. This allows to gain efficiency in testing by always considering only the simplest scenario such as the SCT and using the Speaker Alternation Rate (SAR) as a simple and efficient metric providing a meaningful representation of interactivity.

In the next chapter, I explore the influence of both the delay and the conversation scenario on the perceived speech quality.

4 Impact of Transmission Delay on Perceptual Speech Quality

4.1 Introduction

The introduction of packet-based speech transmission technology such as VoIP results in a considerable amount of absolute delay compared to circuit-switched technology. The International Telecommunications Union (ITU-T) recommends strict limits regarding the one-way delay, i.e., 150 ms for good speech quality and 400 ms for acceptable quality. Above 400 ms, the speech quality is supposed to be unacceptable for the users. In VoIP, different delay sources like the packet queueing in routers and playout-buffering may sum up to delay values that exceed the limits given above. As will be presented in Section 4.2, recent studies have shown that users hardly notice pure delay impairment. Most of these studies have used test scenarios that result in low levels of conversational interactivity and may not lead to situations in which the test subjects perceive the delay.

In this chapter, I investigate the quality impairment of delay using a variety of test scenarios. Since I focus on the pure delay impairment in an echo free environment, the delay is not “audible” in the sense that it can be heard as an echo-time. This has important consequences for quality perception and speech quality assessment: Test subjects may not perceive the impairment at all because they simply do not hear it. Instead, the quality perception may be determined by the *conversation context* as defined in Section 3.1, and thus by a set of parameters which are not related to network quality-of-service (QoS). This set of alternative parameters include the conversation situation, e.g., the purpose of a call, human factors, e.g., the experience of the user, and the structure/interactivity of a conversation. The interactivity of a conversation results from the given task and the realization of the task in an actual conversation which in turn depends on the personality of the users. All these parameters may affect the users perception of latency as an impairment.

This study will focus on the influence of the test scenario on the perception of delay impairment. For this purpose, I use the scenarios that have been introduced in Chapter 3. From these scenarios the Random Number Verification (RNV) scenario resulted in significantly higher conversational interactivity than free conversation-like scenarios (cf. Section 3.7.3), and is thus expected to cause more

quality degradation at delay conditions above the recommended limit of 150 ms one-way delay.

This chapter is structured as follows. In Section 4.2, I present related work with regard to the influence of transmission delay on perceived speech quality. Section 4.3 and 4.4 describe the subjective speech quality tests I have carried out. The results of our tests are presented and discussed in Section 4.5. Finally, I conclude this chapter in Section 4.6.

4.2 Related Work

This section provides an overview of studies in which the impact of transmission delay on the perceptual speech quality has been measured.

Kitawaki [68] has carried out a study in which he tested the detectability of transmission delay and its impact on the speech quality. He has used six different conversation scenarios ranging from taking turns reading random numbers aloud as quickly as possible, to having a free conversation as described in Section 3.2.4. The test subjects were divided into four groups. The first group consisted of four trained female experts. The second group consisted of 20-30 untrained employees of the laboratory. As the third group, 32-44 untrained businessmen, housewives and students took part in the study. Before the experiment, the test subjects experienced delay effects on communication quality for about 30 minutes. The results of Kitawaki's study showed that delay detectability highly depends on the experience of the user and on the conversation scenario. While the experts' detectability threshold was found in the range of 100-700 ms round-trip-delay, untrained subjects detected latency in the range of 350-1100 ms depending on the task (cf. Section 3.2.4). Figure 4.1 presents the results of Kitawaki's tests. The subjects rated the quality on a five-point scale: Excellent (5), Good (4), Fair (3), Poor (2), and Unsatisfactory (1)¹. Conversations based on task 1 (quick random number reading) and task 2 (quick verification of random numbers) resulted in worst quality ratings at increasing delay conditions.

Karis [65] has tested the impact of delay on the speech quality in a mobile telecommunication environment. In his experiment, he used a scenario in which the test persons had to match their halves within a matrix of postcards. The time given for completing the tasks was limited to 10 minutes. Six pairs of subjects who did not know each other fulfilled the tasks over echo-free connections at 0, 300, and 600 ms

¹The quality scale that was used in Kitawaki's study ranged from 0 (unsatisfactory) to 4 (Excellent) and was transformed to a scale ranging from 1 to 5 for consistency.

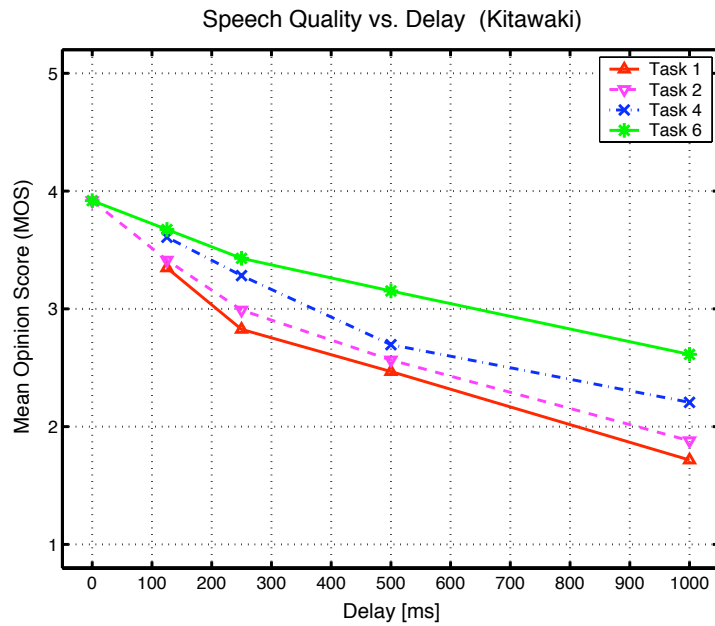


Figure 4.1: Kitawaki: Quality as a function of delay (from [68]). “Task 1” consisted of reading random numbers aloud as quickly as possible, in “Task 2”, the test persons verified random numbers as quickly as possible, “Task 4” consisted of the quick verification of city names, and “Task 6” resulted in a free conversation.

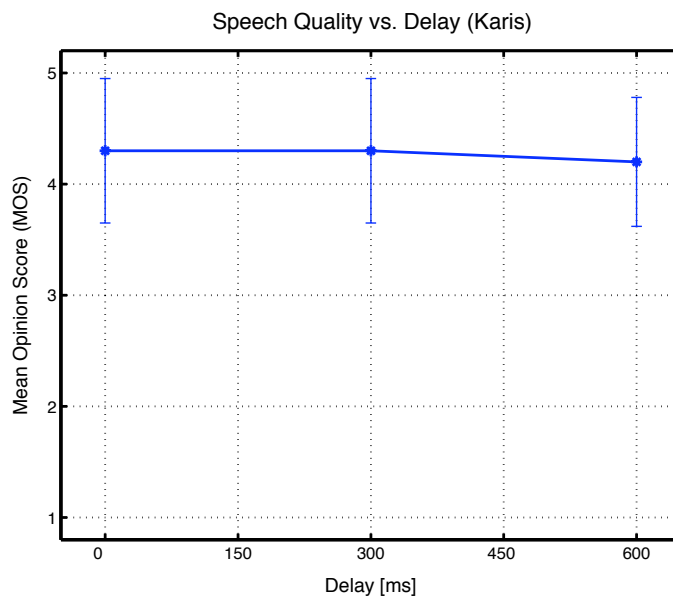


Figure 4.2: Speech quality as a function of delay (from [65]).

one-way delay. After each conversation, the subjects rated the speech quality on a 5-point absolute category rating scale and the listening effort on another 5-point scale (5-“complete relaxation”). In addition, the task performance was tested in terms of errors and incomplete cells. The results concerning the speech quality are depicted in Figure 4.2. The ratings show that the participants did not assign the impact of delay to the quality rating.

Raake has studied the quality impairment caused by combined degradations [83]. In a series of subjective quality tests he combined packet loss with noise, delay, and echo. In the packet loss/delay study, a G.729 codec was used at packet loss conditions of 0, 3, 5, 15% and delay conditions of 200, 400, and 600 ms. 22 naïve subjects² participated in the test. The perceived speech quality was assessed using a 5-point ACR-scale and a CR-10 (Category Rating) degradation scale [8]. Figure 4.3 depicts the results for the R ratings of the delay study in comparison to the E-model’s quality prediction at the respective conditions. The figure shows that large values of delay hardly affect the quality perception, and that the E-model underestimates the quality at these conditions. However, the degradation due to packet losses is more obvious (“audible”) to the test persons.

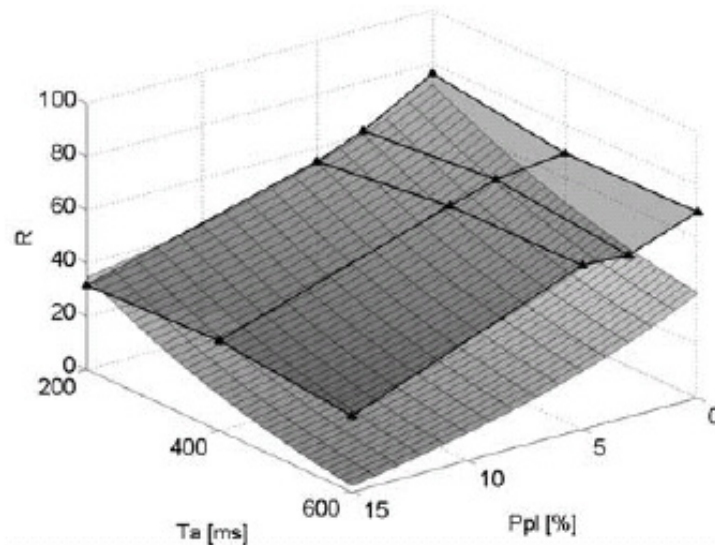


Figure 4.3: Quality as a function of delay and random packet loss (from [83]). R denotes the rating factor which corresponds to the predicted speech quality. Ta denotes the absolute delay, and Ppl represents the packet loss percentage. The upper plane illustrates the results from subjective conversational quality tests, and the lower plane depicts the speech quality ratings predicted by the E-model.

²The term *naïve* refers to the fact that the subjects have not attended a similar quality test before.

Guéguin et al. carried out subjective speech quality tests investigating the relationship between listening, talking and conversational quality in delay and echo situations [29,62]. The part concerning conversational quality consisted of 16 pairs of test subjects accomplishing French SCT tasks at four delay conditions (0, 200, 400, 600 ms). At one participant's side, two levels of echo were introduced (no echo, 25 dB electrical echo level attenuation). The subjects used ISDN handsets and rated the speech quality on a 5-point ACR-scale [49] and an echo annoyance scale [52]. The results (cf. Figure 4.4) show that if there is no echo, the delay hardly impairs the overall quality.

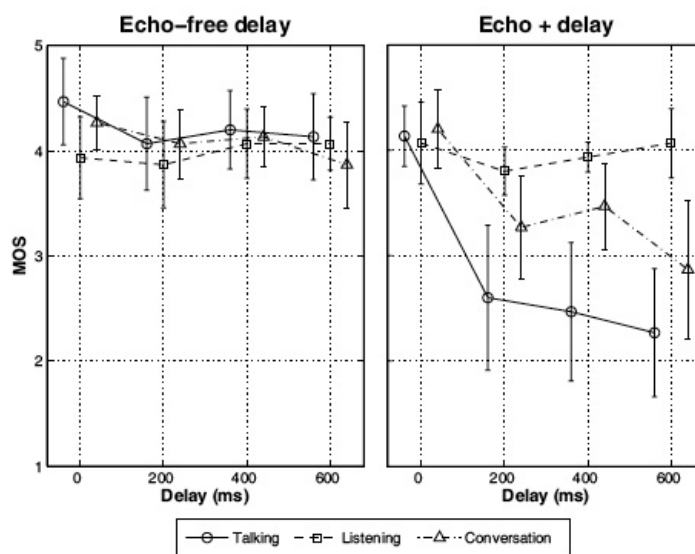


Figure 4.4: Quality as a function of delay with and without echo (from [62]). The x-axis denotes the delay, and the y-axis denotes the Mean Opinion Score (MOS) of the perceived speech quality.

A major methodological difference between these studies is the type of conversation tasks that have been used for the tests. As an example, the rapid exchange of random numbers as used by Kitawaki results in a completely different conversational structure and interactivity than tasks like reserving a plane ticket as used by Raake.

A comparison of the majority of the results presented in this section with the recommendations/requirements stated by the standardization (cf. previous section) shows that the limit of 150 ms one-way delay given by the ITU-T [56] is very conservative. In the following, I aim at studying the perceptual quality as a function of delay by employing conversation scenarios that result in different levels of interactivity (cf. Chapter 3).

In the following, I will first present the objectives and measurement setups of the subjective quality experiments. Then, I present the results of both experiments in a separate section (cf. Section 4.5).

4.3 Experiment 1

4.3.1 Objective

As a continuation of Experiment 1b of Chapter 3, in this experiment I investigate the perceived quality as a function of the transmission delay for the iSCT scenario³. The perceptual quality ratings are then put into relation with the conversational interactivity measures presented in the previous chapter.

4.3.2 Measurement Setup and Test Procedure

This study is based on the measurement setup and test procedure described in Section 3.5.2. Test subjects carried out iSCT scenario tasks at four pure delay conditions (60, 400, 600, and 1000 ms) using the ITU-T G.729 codec. In the current experiment, the perceived speech quality has been measured by means of two scales: the absolute (overall) quality was rated on a 5-point (ACR) scale [49] as presented in Section 1.4 and a degradation category rating scale (CR-10, [8, 72]). The CR-10 allows the subjects to rate the *perceived impairment* of the connection on a scale between 0 and 10, where 0 denotes that the user has not perceived any impairment at all, whereas, e.g., 2, 5 and 10 correspond to “weak”, “strong”, and “extremely strong” impairment, respectively⁴. The CR-10 scale is illustrated in Figure 4.5.

The results of Experiment 1 are presented and discussed in Section 4.5.1.

4.4 Experiment 2

4.4.1 Objective

In this experiment, I analyze the quality ratings resulting from different scenarios at different delay conditions. I put the perceived quality into relation to the conversational parameters and conversational interactivity. Moreover I analyze the perceived interactivity, and realism of the test tasks.

³Note that this work has jointly been carried out with my colleague Alexander Raake at the Institute of Communication Acoustics at Ruhr-University-Bochum.

⁴The CR-10 scale is copyrighted by Gunnar Borg [8]

0	nothing at all
0.5	extremely weak (just noticeable)
1	very weak
2	weak
3	moderate
4	
5	strong
6	
7	very strong
8	
9	
10	extremely strong (almost max)
•	maximal

Figure 4.5: CR-10 category rating scale [8]. This scale provides a measure about the perceived degradation.

4.4.2 Measurement setup and test procedure

The VoIP simulation setup used in this experiment is described in Section 3.6.3 of Chapter 3. The conversations were held using the G.711 codec over connections at transmission delays of 200, 350, and 500 ms. In order to measure the user's opinion about a given connection, the test subjects were required to fill out a questionnaire after each conversation. The post-conversation questionnaires consisted of questions about the overall quality, the perceived speaker alternation rate, the realism of the task, and the perceived conversation flow.

The overall quality was determined by the question, "How do you rate the quality of the connection you were just using?". The subjects rated their opinion on the 5-point ACR scale. As shown in Figure 4.6, a continuous scale was presented in order to obtain low standard deviations.

Note that I did not instruct the test persons about the detailed use of the quality scale in terms of reference conditions which result in a particular quality rating because I wanted to avoid directing the subjects towards a particular way of rating. Delay is a special impairment in the sense that the measurement of its impact on quality would probably be disturbed by the fact that the subjects know about its existence. As soon as the test persons know they need to rate the influence of the latency, they might focus their attention on this particular impairment during the

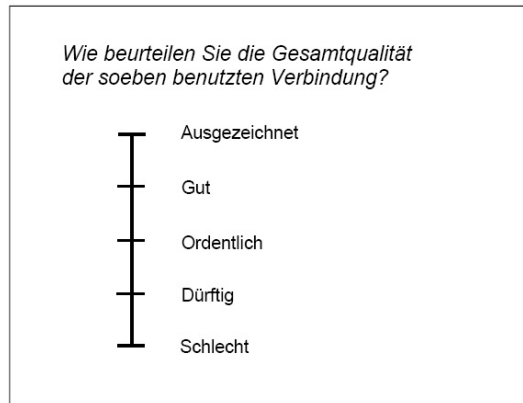


Figure 4.6: Absolute Category Rating Scale for the overall quality.

measurement phase, i.e., during the conversations. If the quality is degraded by, e.g., packet losses, the impairment is obvious. Situations in which the subjects cannot comprehend what was said can be used as examples for bad quality. In our kinds of measurements, the test persons might not even notice that delay occurred on the line unless they are told before. Kitawaki [68] has trained his test subjects and obtained results showing high impact of delay, while in almost all of the other related studies, delay did not have such an impact (cf. Section 4.2). The measurement of pure delay impairment substantially differs from the measurement of audible distortions of the speech signal such as packet losses. Considering a real-life situation, a user does not know about the delay condition that she may experience during a call, except for situations in which she makes a call being aware of increased delays, e.g., a long-distance call.

As I aimed at also measuring perceived interactivity, I continued the questionnaire by asking, “How fast were the speaker alternations of the conversation you’ve just had?”. I provided a continuous rating scale that ranged from “seldom” to “frequent”. From the results presented in Figure 3.22, I would expect that the RNV scenario would result in high values of the perceived speaker alternation rate, the values for the SCT and aSCT scenarios would range in about the middle of the scale, and FC conversations might result in rather low perceived speaker alternation rates.

The test procedure consisted of four different types of conversation scenarios. Test persons may perceive the realism of each individual task of a scenario in a different way. Hence, I asked the subjects, “How relevant was the previous task for your everyday life?”. The respective continuous scale ranged from “unrealistic” to “realistic” and is illustrated in Figure 4.8. Considering the scenarios in use, I would expect that the SCT and FC tasks are rated as most realistic (except for the bank robbery task) and the realism of the RNV tasks is rated low due to their artificial nature.

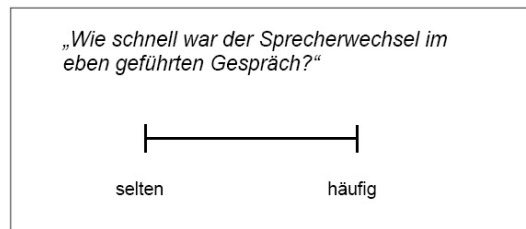


Figure 4.7: Scale for Perceived Speaker Alternation Rate.

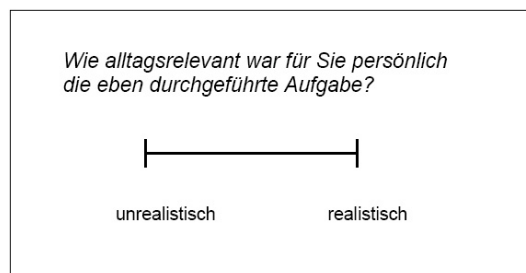


Figure 4.8: Scale for Perceived Task Realism.

Finally, I tested another concept that was expected to reflect perceived interactivity: “Perceived Conversation Flow” of the conversation. The question, “How fluid was the conversation?”, was to be rated on a continuous scale between “tough” and “fluent”.

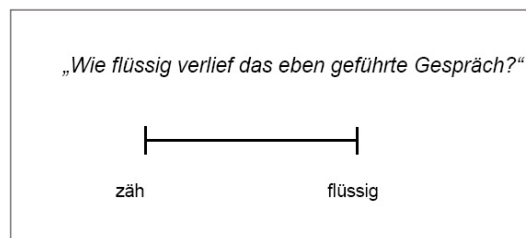


Figure 4.9: Scale for Perceived Conversation Flow.

The subjects were asked to rate the respective attribute by making a cross on the continuous scale. For my analysis of the results, I have measured the ratings using a ruler with an accuracy of 0.5 mm. The results of the perceived interactivity ratings and task realism were then transformed to a scale from 0 to 100.

The results of Experiment 2 are presented and discussed in Section 4.5.2.

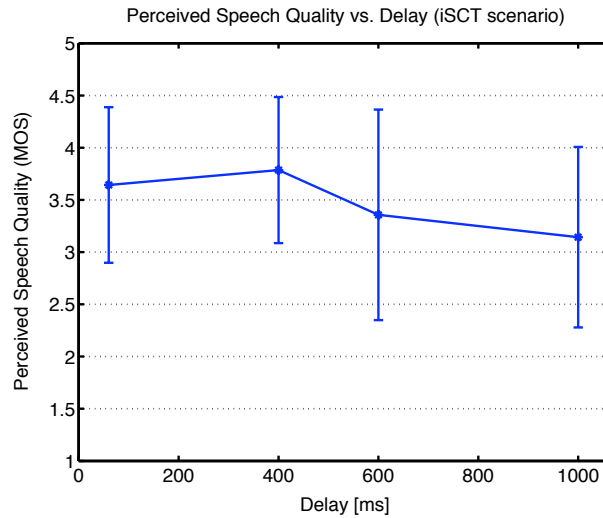


Figure 4.10: Perceived speech quality (MOS) using the iSCT scenario.

4.5 Results and Discussion

4.5.1 Quality Impairment using the iSCT scenario

In this section, I present the results of Experiment 1 with regard to the perceived quality⁵.

Perceived Speech Quality

Figure 4.10 presents the evolution of the perceived quality (MOS) and Figure 4.11 illustrates the CR-10 values with respect to increasing delay, where each parameter has been averaged over all conversations held by seven pairs of subjects (After screening the results, quality ratings that differed from the mean by more than 2 MOS-points and highly inconsistent ratings across all conditions were removed). Both the MOS and CR-10 ratings indicate only a slight decrease in perceptual quality at very high delay.

In the following, the MOS values are given as mean and standard deviation (in parentheses). At a delay of 60 ms, the speech quality ratings resulted in a MOS of 3.64 (0.74). As shown in the graph, the MOS increases to 3.79 (0.70) at a delay time of 400 ms. However, this increase in quality is not significant. At higher delays, the MOS decreases to 3.36 (1.01) at 600 ms and 3.14 (0.86) at 1000 ms.

Regarding the average values of the CR-10 ratings, I can observe a slight but steady increase of perceived impairment for increasing transmission delay. While a delay of

⁵Parts of these results have been published in [35] and in [82]

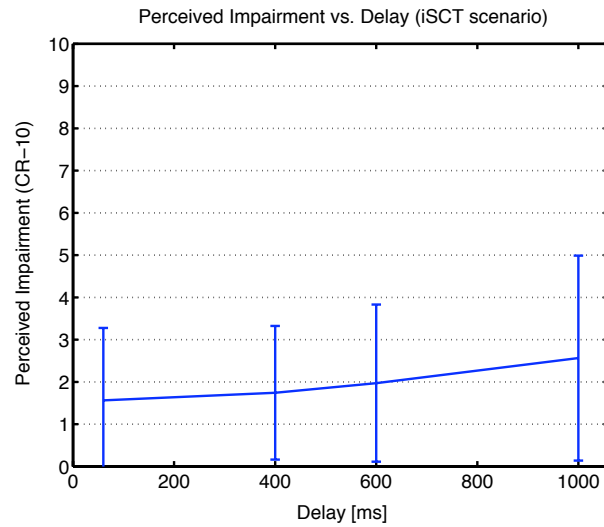


Figure 4.11: Experiment 1b: Perceived impairment (CR-10) for the iSCT scenario. CR-10 scores of 1, 2 and 3 denote “very weak”, “weak” and “moderate” quality impairment, respectively.

60 ms results in an average CR-10 value of 1.56 (1.71), the increases to 1.74 (1.58) at 400 ms, to 1.97 (1.86) at 600 ms, and to 2.56 (2.42) at 1000 ms are not significant.

4.5.2 Influence of Conversation Scenarios

In this section, I present the results of Experiment 2, i.e., the mean user ratings with regard to speech quality, “perceived interactivity” in terms of perceived speaker alternation rates and perceived conversation flow, and the realism of the conversation tasks.

Perceived Speech Quality

Figure 4.12 presents the average perceived speech quality of the individual scenarios at different amounts of transmission delay based on the ratings of $N=15$ test persons (After screening the results, quality ratings that differed from the mean by more than 2 MOS-points and highly inconsistent ratings across all conditions were removed). All conditions, i.e., four scenarios (RNV, SCT, aSCT and FC) and three delay conditions (200, 350 and 500 ms) resulted in a mean opinion score (MOS) of around 4.5. Following the results of an ANOVA, the delay had no effect on quality on any of the scenarios. Contrary to my expectations that the delay highly influences the perceived quality ratings in the RNV scenario, I did not observe any significant quality impairment caused by delay in our setup. Although the RNV scenario is highly interactive (as shown in Section 3.7.3 of Chapter 3), its pre-determined

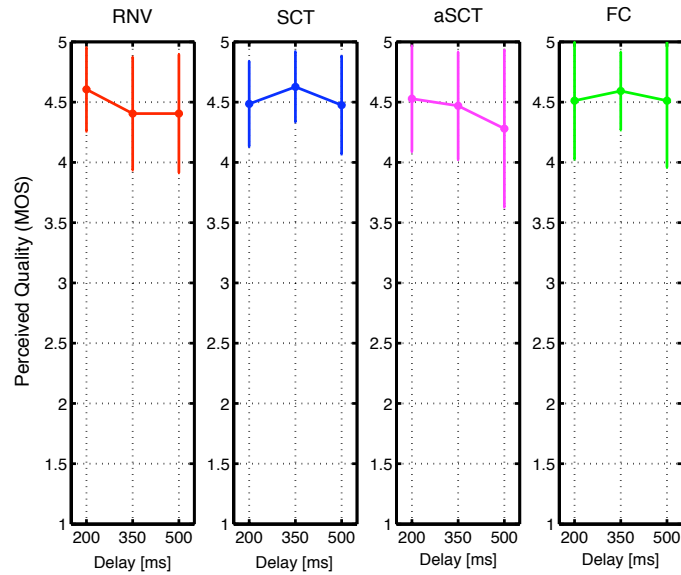


Figure 4.12: Mean perceived speech quality ratings vs. delay for four scenarios: Random Number Verification (RNV), Short Conversation Test (SCT), asymmetric Short Conversation Test (aSCT), and Free Conversation (FC).

structure may cause the subjects not to perceive higher delay conditions as disturbing. In the scenario, the random numbers are given and need to be verified in a given order. Moreover, the turn-taking process is controlled in terms of the subjects knowing that the conversation partner is supposed to reply to a given statement, i.e., the actual number to verify. As shown in Figure 3.17, the RNV scenario exhibits low amounts of double talk. Instead, the amount of mutual silence increases. Thus, the test persons seem to adapt to the new line conditions, tolerate increased response times, and do not experience serious amounts of double talk that might lead to reduced quality ratings.

The high impact of delay on quality as reported by Kitawaki [68] seems to be driven by the fact that the test subjects were exposed to high delay conditions in a training phase before the actual test. In my study, the subjects' delay sensitivity was not trained and the subjects were never told that the experiments have anything to do with delay. Maybe asking for the “perceived speech quality” even distracts further from the delay issue because the term “speech quality” may mostly be associated with impairments like noise or echo. My results do not indicate severe quality degradation at delays exceeding the limit of a one-way delay of 400 ms for acceptable quality as given by the standardization.

Perceived Interactivity

Figure 4.13 presents the Perceived Speaker Alternation Rates (PSAR) of the individual scenarios at different amounts of transmission delay on a scale from 0 to 100. “0” denotes seldom PSAR, and “100” indicates frequent PSAR. The SCT, aSCT and FC scenarios result in about the same amount of PSAR. As expected, the ratings for the RNV scenario higher (around a score of 85), but not by a significant amount.

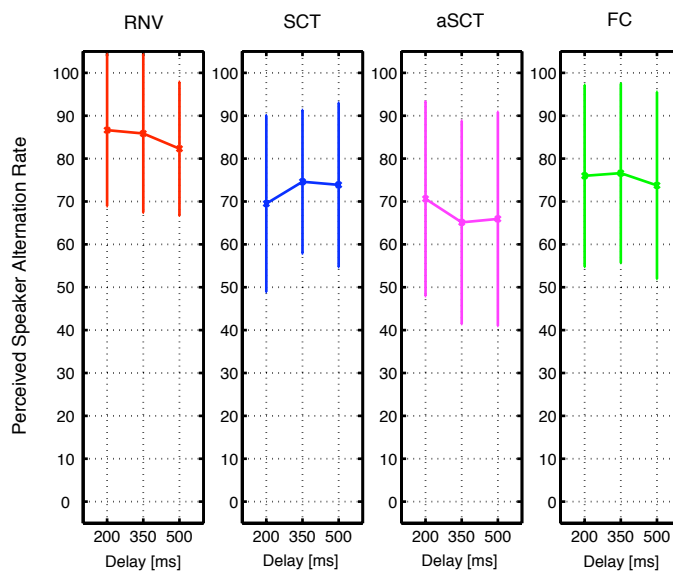


Figure 4.13: Mean perceived speaker alternation rates (PSAR) for the RNV, SCT, aSCT and FC scenarios. “0” denotes seldom PSAR, and “100” indicates frequent PSAR.

As another potential measure for perceived interactivity, I analyze the Perceived Conversation FLow (PCFL) of the individual scenarios at different amounts of transmission delay on a scale from 0 to 100. The PCFL is illustrated in Figure 4.14. All four scenarios resulted in about the same ratings around 85. The PCFL values of the RNV and aSCT scenarios tend to decrease at a delay of 500 ms. For the aSCT scenario, this decrease is confirmed by an ANOVA ($F=4.5$, $p<0.05$).

I conclude that the continuous scales for perceived interactivity (PSAR and PCFL) should be divided into a number of categories and corresponding attributes which can easily be associated with respective conversational situations by the test subjects.

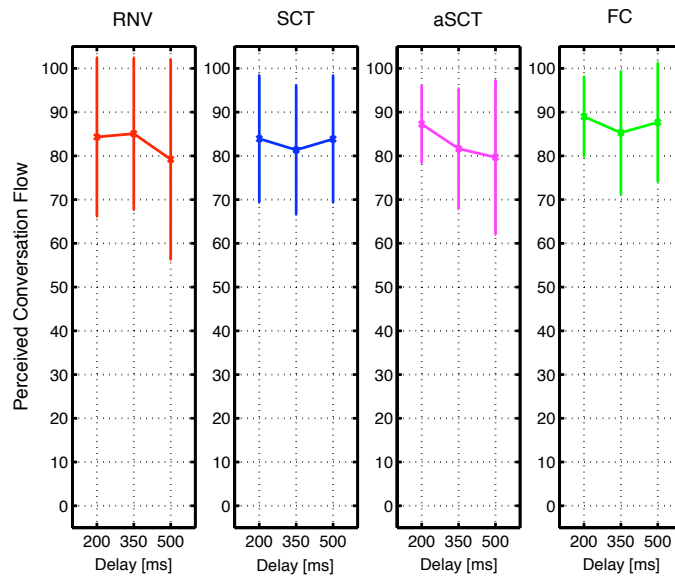


Figure 4.14: Mean perceived conversation flow for the RNV, SCT, aSCT and FC scenarios. A score of “0” denotes a tough conversation, whereas a score of “100” indicates that a fluent conversation was possible.

Task Realism

As a final result, Figure 4.15 presents the mean ratings regarding the task realism of the individual tasks per scenario. Free conversation tasks of organizing a party and talking about the latest vacation were highly realistic to the test persons, whereas, the bank robbery was rated least realistic with a score of 8.4 (11.1 standard deviation). The SCT scenario was also assigned high relevance (a score of about 83) followed by the aSCT scenario (mean score about 68 with the weather data task as an outlier at a score of 47). The RNV scenario was rated least realistic at a score of 50 points. I conclude that my results confirm the applicability of the SCT tasks for speech quality tests in terms of providing realistic situations.

4.6 Summary

In this chapter, I have studied the impact of absolute delay on the perceived speech quality by using different conversation scenarios. I have analyzed the results obtained in two conversational speech quality tests based on a total of five scenarios as described in chapter 3. In the first test, I have used the interactive Short Conversation Test (iSCT) scenario that was designed to result in higher conversational interactivity than a usual Short Conversation Test (SCT). The test persons were exposed to absolute delay times of 60, 400, 600, and 1000 ms.

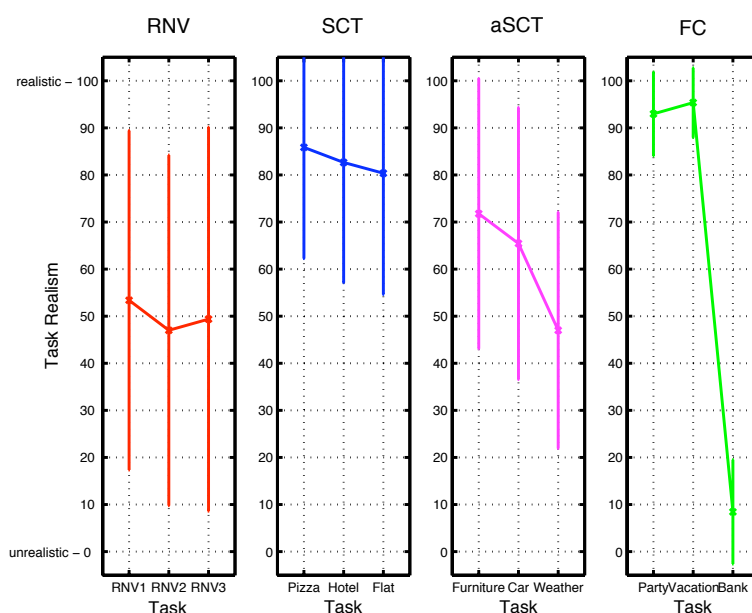


Figure 4.15: Mean ratings for task realism of the individual tasks of each of the four scenarios. A score of “0” denotes that the task is regarded as unrealistic, and a score of 100 indicates that the task is considered realistic.

In spite of the long delay times, the speech quality was hardly degraded for the iSCT scenario. In a second subjective test, the test persons had to accomplish four different scenarios (Random Number Verification, Short Conversation Test, asymmetric Short Conversation Test, and Free Conversation) at three different conditions of absolute delay (200, 350, and 500 ms). In this test the subjects represented a wide range of ages. As an approach for the measurement of perceived interactivity, questions about perceived speaker alternation rate and perceived conversation flow were included. The most remarkable result of the second test is that even the use of highly interactive random number verification scenario did not lead to increased quality degradation caused by delay. No significant influence of absolute delay on the perceived quality was found in any of the scenarios. I have investigated a first approach of measuring “perceived interactivity” in terms of perceived speaker alternation rate and perceived conversation flow. Regarding the perceived interactivity, the RNV scenario resulted in slightly higher perceived speaker alternation rates, and the aSCT scenario conversations were rated to be less fluent at a delay of 500 ms. The free conversation scenarios (except for the bank robbery) were rated most realistic followed by the SCTs. Finally, taking the conversational parameters and the task realism into account, I conclude that the SCT scenario represents everyday conversations well.

5 Conclusions and Outlook

5.1 Conclusions

The quality of packet-based telephony is influenced by two major factors that may degrade the perceived speech quality: packet losses and absolute delay. In this thesis I have approached these impairments in three ways. Firstly, I have investigated the possibility of improving the perceived quality of connections that suffer from bit errors by saving damaged speech packets from being discarded. Secondly, I have introduced measures of conversational interactivity which allow a distinction of scenarios that are used in conversational speech quality tests. Thirdly, I have carried out user tests regarding the quality impairment caused by delay. Those tests were based on scenarios that result in different levels of interactivity.

My first study dealt with the transport of speech packets over erroneous transmission channels. Modern speech codecs provide the possibility to distinguish between speech data that is perceptually important and data that is less important. Based on a modified transport protocol, i.e., UDP-Lite, I have simulated the incorporation of damaged speech bits into the speech decoding process. To this end, I considered three different scenarios: The first scenario represented the usual RTP/UDP/IP transport of speech frames without any possibility to save corrupted data. In the second scenario, I tolerated damaged speech bits that were not perceptually important, but used the packet loss concealment algorithm in the case that important bits were damaged. Finally, I incorporated all corrupted data bits into the speech decoding. In addition, I simulated the use of robust header compression, which significantly reduced the protocol overhead. As speech material, I have used high-quality speech samples containing German sentences. The speech quality of the resulting degraded speech samples was measured by using PESQ, an instrumental speech quality estimation algorithm. The results have shown that incorporating all damaged speech data improves the speech quality. The quality improvement gets even more obvious when robust header compression is used, since a lot of packets can be saved from being discarded. Those results were confirmed by applying TOSQA as an additional speech quality assessment tool. Comparing these two algorithms for instrumental quality estimation, the use of PESQ results in higher quality ratings for male speech samples, whereas TOSQA yields balanced quality ratings for both genders.

In Chapter 3, I addressed the conversational interactivity of telephone conversations. Based on the assumption that interactive scenarios result in higher impact of delay impairment on speech quality than less interactive scenarios, I aimed at defining a metric that allows for the distinction of individual scenarios. To this end, I have first defined conversational interactivity as “a single scalar measure based on the quantitative attributes of the participants’ spoken contributions”. Then, I introduced the framework of Parametric Conversation Analysis (P-CA) comprising a 4-state conversation model, its respective parameters, and conversational events such as speaker alternations and interruptions. In the next step, I developed three metrics for conversational interactivity: Firstly, the speaker alternation rate which represents the speaker alternations per minute. As the second metric, the conversational temperature reflects how long a conversation remains in one of the model’s states. Thirdly, the entropy rate which is based on a simplified model requiring one state per speaker, and can thus be extended to conversations with more than two participants. In two subjective speech quality tests based on a variety of conversation scenarios, the conversations were recorded and their structure was analyzed using the P-CA. Comparing the interactivity of the individual scenarios, only the Random Number Verification scenario resulted in significantly increased interactivity which significantly decreased with delay. Since the structure of the other scenarios was not designed as rigidly as the structure of the number verification scenario, their interactivity remained about the same, both across the scenarios and delay conditions. Regarding the question of how to characterize the interactivity of conversations, I conclude that the speaker alternation rate represents a simple and efficient measure for conversational interactivity.

As the third main topic of this thesis I investigated the impact of delay impairment on perceptual speech quality in Chapter 4. In two subjective speech quality tests participants accomplished various tasks of five different conversation scenarios of which one results in highly interactive conversations compared to the others. In general, the quality ratings showed that delay hardly influences the perceived quality for any of the studied scenarios. Particularly, in contradiction to my expectation, the highly interactive scenario did not yield a significant decrease of quality at one-way delay times up to 500 ms. As expected, the users rated the free conversation as most relevant for their everyday life, followed by the short conversation test scenario. Regarding the relation between delay, conversational interactivity and perceptual speech quality, I conclude that up to an absolute delay of 500 ms, delay does not seem to impair the perceived quality, even in highly interactive situations. These results suggest that users adapt to conditions of higher latency, and thus do not consider such conditions as bad.

5.2 Outlook

Based on the results presented in the related work as well as the results given in this thesis, I conclude that the methodology for measuring the quality degradation caused by delay can be improved. Firstly, a comparison of the ratings of trained and non-trained test persons would clarify the effects of how trained test persons react to higher delay conditions both by means of quality perception and conversational interactivity. With increased usage of delayed telephone connections, trained users can be expected to start getting bored by high latency and give worse quality ratings. In real life, however, situations may occur in which the user does not know about the line conditions in advance, and thus is not prepared to detect increased delay. A study about training effects would help understanding this issue.

The fact that users seem to tolerate absolute delays up to 500 ms has consequences concerning the packet-loss vs. delay trade-off. On the one hand, late packets may easily be incorporated into the decoding and playout process. On the other hand, packet loss repair mechanisms which introduce some latency like, e.g., forward-error-correction (FEC) or the repeated transmission of packets, may be applied to error-prone links without degrading the overall quality. This leads us to the question of how users distinguish and weigh the listening-only quality elements such as, e.g., packet loss, and the interaction-related quality components such as absolute delay.

The conversational temperature and entropy rate metrics give rise to further work on the development of the underlying functions and to the performance of user tests. Regarding the conversational temperature metric, the replacement of the exponential function by alternative functions may lead to more flexible modeling of the conversational interactivity. Furthermore, the underlying sojourn times of the states may require individual weighting in order to optimize the metric's performance. Considering the entropy metric, an elaboration is indicated on modeling multi-party conversations, for which the entropy rate is especially applicable. An associated study of multi-party telephone conferences requires the distinction of active and passive participants who are involved into the discussion at different kinds of intention, i.e., the purpose and motivation of the participation, and different levels of attention.

Another aspect of conversational interactivity in terms of quantitative attributes of the participants' spoken contributions is the mean number of active interruptions that occur during the accomplishment of a task. Interruptions generate situations in which the response time of the interrupted call participants and/or the duration of the double talk phases during the interruption may influence the quality ratings of the test subjects. In a test situation, the interruptions should occur incidentally as

to keep the natural flow of a conversation. Thus, the design of a scenario that intrinsically provokes interruptions may be of great value for further investigations on the relation between transmission delay, conversational structure/interactivity and QoE.

In this thesis, I have approached conversational interactivity as a parameter that can instrumentally be measured from conversation recordings. However, the attention and involvement of the participants depends to a great extent on the semantics and pragmatics of the conversation, i.e., to the meaning of what was said. Within an ongoing call, in some situations absolute delay may disturb the users because they need to interrupt the other participant. In other stages of a conversation the participants simply do not care about the consequences of interruptions because they are in a joyful mood and make jokes. Therefore, distinguishing a manageable set of parameters that characterize different phases of a conversation, and generating such situations by designing accordant conversation tasks (see above) may be an interesting interdisciplinary research topic. Obviously, taking the conversation contents into account requires a great deal of human intervention, i.e., coding and transcription, and may not be accomplished in an instrumental way in the near future.

A Acronyms

3GPP 3rd Generation Partnership Project

3SQM Single Sided Speech Quality Measure

ACR Absolute Category Rating

AMR Adaptive Multi-Rate

ANOVA Analysis of Variance

aSCT Asymmetric Short Conversation Test

BER Bit Error Rate

CCR Comparison Category Rating

DCR Degradation Category Rating

DSL Digital Subscriber Line

DTX Discontinuous Transmission

FC Free Conversation

FEC Forward Error Correction

GSM Global System for Mobile communication

IETF Internet Engineering Task Force

IKA Institute of Communication Acoustics

IP Internet Protocol

INMD In-Service, Non-Intrusive Measurement Device

IRS Intermediate Reference System

iSCT Interactive Short Conversation Test

ISDN Integrated Services Digital Network

ITU International Telecommunications Union

LBR Low Bit Redundancy

MOS Mean Opinion Score

P-CA Parametric Conversation Analysis

PCM Pulse Code Modulation

PESQ Perceptual Evaluation of Speech Quality

PLC Packet Loss Concealment

POTS Plain Old Telephone Service

PSTN Public Switched Telephone Network

ROHC RObust Header Compression

RS Reed-Solomon

RTP Real-time Transport Protocol

QoS Quality of Service

RNV Random Number Verification

SCT Short Conversation Test

TOC Table Of Contents

TOSQA Telecommunication Objective Speech Quality Assessment

VoIP Voice over Internet Protocol

WLAN Wireless Local Area Network

UDP User Datagram Protocol

UEP Unequal Error Protection

UMTS Universal Mobile Telecommunications System

VAD Voice Activity Detection

VoIP Voice over IP

WiMAX Worldwide interoperability for Microwave Access

WLAN Wireless Local Area Network

B Scenarios

B.1 Random Number Verification

Example for a Random Number Verification (RNV) task (calling party)¹:



Aufgabe: Überprüfung von Zahlen



Ihr Gesprächspartner hat auch so eine Liste. Manche Zahlen stimmen nicht mit denen Ihres Gesprächspartners überein. **Finden Sie die falschen Zahlen so schnell wie möglich** indem Sie sie abwechselnd zeilenweise lesen. Bestätigen Sie die Richtigkeit mit „JA“ oder „NEIN“ und streichen Sie die falschen Zahlen durch. Sie lesen dabei die roten Zahlen, Ihr Gesprächspartner die blauen.

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

¹Instructions: “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 red numbers and your conversation partner will read the blue ones”.

Example for a Random Number Verification (RNV) task (called party)²:



Aufgabe: Überprüfung von Zahlen



Ihr Gesprächspartner hat auch so eine Liste. Manche Zahlen stimmen nicht mit denen Ihres Gesprächspartners überein. **Finden Sie die falschen Zahlen so schnell wie möglich** indem Sie sie abwechselnd zeilenweise lesen. Bestätigen Sie die Richtigkeit mit „JA“ oder „NEIN“ und streichen Sie die falschen Zahlen durch. Sie lesen dabei die blauen Zahlen, Ihr Gesprächspartner die roten.

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

²Instructions: “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 blue numbers and your conversation partner will read the red ones”.

B.2 Short Conversation Test

Example for a Short Conversation Test (SCT) task (calling party)³:



Aufgabe: Pizzabestellung



Bestellen Sie eine große Pizza für zwei Personen bei der Pizzeria Don Pedro.



Die Pizza soll vegetarisch sein.



Belag : _____

Preis : _____ €



Lieferung an : Blütengasse 12/7,
1030 Wien,
☎: 347 34 20



Wie lange dauert es bis die Pizza geliefert wird?

³Task: Pizza order

Instruction: "Order a large pizza for two persons at Pizzeria Don Pedro."

Requirement: "The pizza shall be vegetarian."

Items to be provided: Delivery address and telephone number.

Additional question: "How long does it take until the pizza is delivered?".

B.3 Interactive Short Conversation Test

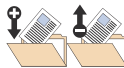
Example for an interactive Short Conversation Test (iSCT) task (calling party)⁵:



Meteorologisches Institut FTW



Tauschen Sie mit Ihrem Gesprächspartner die fehlenden Informationen aus. Tun Sie dies für eine Stadt nach der anderen. [In der ersten Zeile finden Sie ein Beispiel.]



Ort	Meteorologische Daten			
	Gestern		Heute	
	Temperatur	Luftfeuchtigkeit	Temperatur	Luftfeuchtigkeit
Beispiel	15,3°C	53%	16,5° C	63%
Linz	15,2°C	78%		
Graz	16,9°C	65%		
Salzburg	20,4°C	55%		
Innsbruck	14,8°C	84%		
Bregenz	16,2°C	77%		

⁵Meteorologic Institute FTW.

Instructions: “Exchange the missing information with your conversation partner for one city after the other. [An example is shown in the first line of the table.]”.

Provided information: Yesterday’s meteorological data of different cities, i.e., temperature, humidity.

Missing items: Data for today.

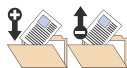
Example for an interactive Short Conversation Test (iSCT) task (called party)⁶:



Meteorologisches Zentrum Annaberg



Tauschen Sie mit Ihrem Gesprächspartner die fehlenden Informationen aus. Tun Sie dies für eine Stadt nach der anderen.
[In der ersten Zeile finden Sie ein Beispiel.]



Ort	Meteorologische Daten			
	Gestern		Heute	
	Temperatur	Luftfeuchtigkeit	Temperatur	Luftfeuchtigkeit
Beispiel	15,3°C	53%	16,5° C	63%
Linz			18,2°C	75%
Graz			17,1°C	61%
Klagenfurt			22,2°C	60%
Innsbruck			15,8°C	81%
Bregenz			16,6°C	74%

⁶Meteorologic Center Annaberg.

Instructions: “Exchange the missing Information with your conversation partner for city line after the other. [An example is shown in the first line of the table.]”.

Provided information: Today’s meteorological data of different cities, i.e., temperature, humidity.

Missing items: Data for yesterday.

B.4 Asymmetric Short Conversation Test

Example for an asymmetric Short Conversation Test (aSCT) task (calling party)⁷:

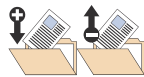


Aufgabe: Möbellagerinformationen

Name: *FTW Möbel-Lagerhaltung*



Bitte Sie Ihren Gesprächspartner von der *FTW Möbel-Verkaufsabteilung* um die Bestell-Nummern und Lagerplätze der unten angeführten Artikel. Tragen Sie die entsprechenden Daten in die Tabelle ein. [In der ersten Zeile finden Sie ein Beispiel.]



Artikel	Bestell-Nr	Lagerplatz
Regal Beispiel	AB-514-78-44	R12 P23 S01
Drehstuhl Datatri		
Kleiderschrank Knoten		
Küche Binnär		
Unterschrank Judipi		
Computertisch Maltimideea		

⁷Task: Furniture store information, Name: FTW Furniture Store

Instructions: “Ask your conversation partner at the FTW Furniture Sales Department for the order numbers and storage positions of the listed items. Fill out the table with the respective information. [An example is shown in the first line of the table.]”

Example for an asymmetric Short Conversation Test (aSCT) task (called party)⁸:



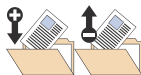
Aufgabe: Möbellagerinformationen

Name: *FTW Möbel-Verkaufsabteilung*



Sie werden von Ihrem Gesprächspartner von der *FTW Möbel-Lagerhaltung* um die Bestell-Nummern und die Lagerplätze einiger Artikel gebeten.

[In der ersten Zeile finden Sie ein Beispiel.]



Artikel	Bestell-Nr	Lagerplatz
Regal Byspiel	AB-514-78-44	R12 P23 S01
Drehstuhl Datatri	BP-145-27-84	R01 P04 S12
Kleiderschrank Knoten	AS-157-82-25	R09 P12 S02
Schreibtisch Peipeer	BS-541-18-87	R05 P01 S03
Unterschrank Judipi	KF-354-38-45	R05 P02 S01
Computertisch Maltimideea	BC-854-25-52	R02 P04 S05

⁸Task: Furniture store information, Name: FTW Furniture Sales Department

Instructions: “You will be asked by your conversation partner at the FTW Furniture Store to provide the order numbers and storage for the items listed in the table. [An example is shown in the first line of the table.]”

B.5 Free Conversation

Example for a Free Conversation (FC) task (calling and called party) ⁹:



Aufgabe: Party organisieren



Organisieren Sie mit Ihrem Gesprächspartner eine Überraschungs-Geburtstagsfeier für einen Freund/eine Freundin. Zur Party sollen ca. 30 Personen geladen werden. Sie haben ca. 7 Minuten Zeit.

Notizen:



Aufgabe: Party organisieren



Organisieren Sie mit Ihrem Gesprächspartner eine Überraschungs-Geburtstagsfeier für einen Freund/eine Freundin. Zur Party sollen ca. 30 Personen geladen werden. Sie haben ca. 7 Minuten Zeit.

Notizen:

⁹Instructions: “Organize a birthday party as a surprise for a friend. About 30 persons shall be invited to the party. Take about 7 minutes to fulfill this task.”

C Algorithms

This chapter provides Matlab-code for the calculation of the conversational temperature and the entropy rate as presented in sections 3.4.2 and 3.4.3, respectively.

C.1 Conversational Temperature

```
function [temp] = state2temp(durA, durB, durM, durD)

% =====
% Estimation of Conversational Temperature
% from sojourn times measured in real conversations
% 2006 by Florian Hammer
% =====

% Sojourn times of a "norm" conversation
% taken from ITU-T Rec. P.59
defA = 0.78;
defB = 0.78;
defM = 0.51;
defD = 0.23;

% norm-temperature = room-temperature
def_temp = 21.5;

% calculate temperatures for each state
% and apply least-squares fitting
tt=1;
for t=0.1:0.1:100
    ss1 = defA*exp((def_temp/t)-1)-durA;
    ss2 = defB*exp((def_temp/t)-1)-durB;
    ss3 = defM*exp((def_temp/t)-1)-durM;
    ss4 = defD*exp((def_temp/t)-1)-durD;
    hm(tt) = ss1*ss1 + ss2*ss2 + ss3*ss3 + ss4*ss4;
    tt=tt+1;
end;
```

```
mm = min (hm);
ind = find(hm==mm);
temp = ind(1,1)/10;
```

C.2 Entropy Rate

```
function entropyrate=entropy(mtimeA,mtimeB,fs);
% =====
% Calculation of the Conversational Entropy-rate
% based on the sojourn times of a two state speaker-turn-model
% 2006 by Florian Hammer
% Input: mtimeA...mean sojourn time of talker A speaking
%        mtimeB...mean sojourn time of talker B speaking
% =====
stime=sum([mtimeA,mtimeB]);
mtime=stime/2;

pA=mtimeA/stime; % probability that A talks
pB=mtimeB/stime; % probability that B talks

tavg=mtime/fs; % average sojourn time

% calculate the entropy-rate
entropyrate=1/tavg*(-pA*log2(pA)-pB*log2(pB));
```

D E-model Parameters

Table D.1 presents the parameters of the E-model and their default values following ITU-T Rec. G.107 [61].

Parameter	Abbr.	Unit	Default value
Sending Loudness Rating	SLR	dB	+8
Receiving Loudness Rating	RLR	dB	+2
Side Tone Masking Rating	STMR	dB	15
Listener Sidetone Rating	LSTR	dB	18
D-Factor of Telephone, Send Side	Ds	-	3
D-Factor of Telephone, Receive Side	Dr	-	3
Talker Echo Loudness Rating	TELR	dB	65
Weighted Echo Path Loss	WEPL	dB	110
Mean One-Way Delay of the Echo Path	T	ms	0
Round-Trip Delay in a 4-Wire Loop	Tr	ms	0
Absolute Delay in echo-free Connections	Ta	ms	0
Number of Quantization Distortion Units	qdu	-	1
Equipment impairment factor	Ie	-	0
Packet-loss Robustness Factor	Bpl	-	1
Random Packet-loss Probability	Ppl	%	0
Burst Ratio	BurstR	-	1
Circuit Noise referred to 0 dBr-point	Nc	dBm0p	-70
Noise Floor at Receive Side	Nfor	dBmp	-64
Room Noise at the Send Side	Ps	dB(A)	35
Room Noise at the Receive Side	Pr	dB(A)	35
Advantage Factor	A	-	0

Table D.1: Default values for the E-model parameters.

Bibliography

- [1] 3GPP. 3rd Generation Partnership Project; Technical specification group services and system aspects; Packet switched conversational multimedia applications; Performance characterisation of default codecs (Release 6). *3GPP TR 26.935 v6.0.0*, June 2004.
- [2] Third Generation Partnership Project (3GPP). <http://www.3gpp.org/>.
- [3] John W. Allnatt. Subjective rating and apparent magnitude. *Int. J. Man Machine Studies*, 7:801–816, 1975.
- [4] Ronald Appel and John G. Beerends. On the quality of hearing one’s own voice. *Journal Audio Eng. Soc.*, 50(4):237–248, April 2002.
- [5] Jens Berger. *Instrumentelle Verfahren zur Sprachqualitätsschätzung - Modelle auditiver Tests*. PhD thesis, CAU Kiel, 1998.
- [6] M. P. Althoff Bernard H. Walke, P. Seidenberg. *UMTS: The Fundamentals*. John Wiley & Sons, Chichester, UK, 2003.
- [7] Richard E. Blahut. *Theory and Practice of Error Control Codes*. Addison-Wesley, NY, 1983.
- [8] Gunnar Borg. *Borg’s Perceived Exertion and Pain Scales*. Human Kinetics, Champaign, IL, 1998.
- [9] Carsten Bormann et al. Robust header compression (ROHC): Framework and four profiles: RTP, UDP, ESP, and uncompressed. *Request for Comments (Standards Track) RFC 3095, Internet Engineering Task Force*, July 2001.
- [10] Jürgen Bortz. *Statistik für Sozialwissenschaftler*. Springer, Berlin, 1999.
- [11] Paul T. Brady. A statistical analysis of on-off patterns in 16 conversations. *Bell System Technical Journal*, 47(1):73–91, January 1968.
- [12] Rudolf C. Bretz and Michael Schmidbauer. *Media for interactive communication*. Sage Publications, Beverly Hills, CA, 1983.

- [13] Walter Y. Chen. *DSL: Simulation Techniques and Standards Development for Digital Subscriber Line Systems*. Macmillan Technical Publishing, Indianapolis, IN, 1998.
- [14] James W. Chesebro and Donald G. Bonsall. *Computer-Mediated Communication*. University of Alabama Press, Tuscaloosa, AL, 1989.
- [15] Edward J. Downes and Sally J. McMillan. Defining interactivity: A qualitative identification of key dimensions. *New Media & Society*, 2(2):157–179, 2000.
- [16] E. Ekudden, R. Hagen, I. Johansson, and J. Svedberg. The adaptive multi-rate speech coder. In *IEEE Speech Coding Workshop*, pages 117–119, Porvoo, Finland, June 1999.
- [17] Mohamed A. El-Gendy, Abhijit Bose, and Kang G. Shin. Evolution of the Internet QoS and support for soft real-time applications. *Proc. of the IEEE*, 91(7):1086–1104, July 2003.
- [18] European Telecommunications Standards Institute. Transmission and multiplexing (TM); Considerations on transmission delay and transmission delay values for components on connections supporting speech communication over evolving digital networks. *ETR 275*, April 1996.
- [19] European Telecommunications Standards Institute. AT&T labs AMR characterization phase final report. *ETSI SMG11#11, Tdoc 193/99*, May 1999.
- [20] European Telecommunications Standards Institute. 3rd Generation Partnership Project; Technical specification group services and system aspects; Speech codec speech processing functions; AMR wideband speech codec; General description (release 5). *ETSI TS 126 171 v5.0.0*, August 2002.
- [21] European Telecommunications Standards Institute. Digital cellular telecommunications system (Phase 2+); GSM enhanced full rate speech processing functions: General description (3GPP TS 46.051 version 5.0.0 Release 5). *ETSI TS 146 051 v5.0.0*, June 2002.
- [22] European Telecommunications Standards Institute. Speech processing, transmission and quality aspects (STQ); Specification and measurement of speech transmission quality; Part 1: Introduction to objective comparison measurement methods for one-way speech quality across networks. *ETSI EG 201 377-1 v1.2.1*, December 2002.
- [23] European Telecommunications Standards Institute. Universal mobile telecommunications system (UMTS); AMR speech codec; Error concealment of lost frames (3GPP TS 26.091 version 5.0.0 Release 5). *ETSI TS 126 091 v5.0.0*, June 2002.

-
- [24] European Telecommunications Standards Institute. Universal mobile telecommunications system (UMTS); Mandatory speech codec speech processing functions; Adaptive Multi-Rate (AMR) speech codec frame structure (3GPP TS 26.101 version 5.0.0 Release 5). *ETSI TS 126 101 v5.0.0*, June 2002.
- [25] European Telecommunications Standards Institute. Universal mobile telecommunications system (UMTS); AMR speech codec; General description (3GPP TS 26.071 version 6.0.0 Release 6). *ETSI TS 126 071 v6.0.0*, December 2004.
- [26] Victor Firoiu, Jean-Yves Le Boudec, Don Towsley, and Zhi-Li Zhang. Theories and models for internet quality of service. *Proc. of the IEEE*, 90(9):1565–1591, September 2002.
- [27] Bavarian Archive for Speech Signals (BAS). Phondat 1 corpus. <http://www.bas.uni-muenchen.de/Bas/BasPD1eng.html>.
- [28] C. Goodwin and J. Heritage. Conversation analysis. *Annual Review of Anthropology*, 19:283–307, 1990.
- [29] Marie Guéguin, Valérie Gautier-Turbin, Laëtitia Gros, Vincent Barriac, Régine Le Bouquin-Jeannés, and Gérard Faucon. Study of the relationship between subjective conversational quality, and talking, listening and interaction qualities: towards an objective model of the conversational quality. In *Measurement of Speech and Audio Quality in Networks, MESAQIN 2005*, 2005.
- [30] Florian Hammer. Wie interaktiv sind Telefongespräche? In *32. Deutsche Jahrestagung für Akustik - DAGA 06*, Braunschweig, Germany, March 2006.
- [31] Florian Hammer and Peter Reichl. How to measure interactivity in telecommunications. In *Proc. 44th FITCE Congress 2005*, pages 187–191, Vienna, Austria, September 2005.
- [32] Florian Hammer, Peter Reichl, Tomas Nordström, and Gernot Kubin. Corrupted speech data considered useful. In *Proc. First ISCA International Tutorial and Research Workshop on Auditory Quality of Systems*, pages 51–54, Herne, Germany, April 2003.
- [33] Florian Hammer, Peter Reichl, Tomas Nordström, and Gernot Kubin. Corrupted speech data considered useful: Improving perceived speech quality of VoIP over error-prone channels. *Acta Acustica united with Acustica*, 90(6):1052–1060, Nov/Dec 2004.
- [34] Florian Hammer, Peter Reichl, and Alexander Raake. Elements of interactivity in telephone conversations. In *8th International Conference on Spoken Language Processing (ICSLP/INTERSPEECH 2004)*, pages 1741–1744, Jeju Island, Korea, October 2004.

- [35] Florian Hammer, Peter Reichl, and Alexander Raake. The well-tempered conversation: Interactivity, delay and perceptual VoIP quality. In *Proc. IEEE Int. Conf. Communications*, Seoul, Korea, May 2005.
- [36] Florian Hammer, Peter Reichl, and Thomas Ziegler. Where packet traces meet speech samples: An instrumental approach to perceptual QoS evaluation of VoIP. In *IEEE International Workshop on Quality of Service IWQOS 2004*, pages 273–280, Montreal, Canada, June 2004.
- [37] Eduard Hasenleithner, Thomas Ziegler, and Peter Krüger. A performance evaluation of software tools for delay emulation. In *IEEE International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS)*, pages 904–913, Philadelphia, PA, July 2005.
- [38] Markus Hauenstein. *Psychoakustisch motivierte Maße zur instrumentellen Sprachgütebeurteilung*. PhD thesis, Christian-Albrechts-Universität Kiel, 1997.
- [39] Olivier Hersent, David Gurle, and Jean-Pierre Petit. *IP Telephony: Packet-Based Multimedia Communications Systems*. Addison-Wesley, London, 1999.
- [40] Institute of Electrical and Electronics Engineers. IEEE standard for local and metropolitan area networks - Part 16: Air interface for fixed broadband wireless access systems. *IEEE Standard 802.16-2004*, October 2004.
- [41] International Telecommunication Union. Pulse code modulation (PCM) of voice frequencies. *ITU-T Recommendation G.711*, November 1988.
- [42] International Telecommunication Union. Specification for an intermediate reference system. *ITU-T Recommendation P.48*, November 1988.
- [43] International Telecommunication Union. Artificial conversational speech. *ITU-T Recommendation P.59*, March 1993.
- [44] International Telecommunication Union. Models for predicting transmission quality from objective measurements. *ITU-T Series P, Supplement 3*, 1993.
- [45] International Telecommunication Union. Objective measurement of active speech level. *ITU-T Recommendation P.56*, March 1993.
- [46] International Telecommunication Union. Terms and definitions related to quality of service and network performance including dependibility. *ITU-T Recommendation E.800*, August 1994.
- [47] International Telecommunication Union. Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP). *ITU-T Recommendation G.729*, March 1996.

- [48] International Telecommunication Union. Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s. *ITU-T Recommendation G.723.1*, March 1996.
- [49] International Telecommunication Union. Methods for subjective determination of transmission quality. *ITU-T Recommendation P.800*, August 1996.
- [50] International Telecommunication Union. Subjective performance assessment of telephone-band and wideband digital codecs. *ITU-T Recommendation P.830*, August 1996.
- [51] International Telecommunication Union. TOSQA - Telecommunication objective speech quality assessment. *ITU-T COM 12-34*, December 1997.
- [52] International Telecommunication Union. Subjective performance evaluation of network echo cancellers. *ITU-T Recommendation P.831*, December 1998.
- [53] International Telecommunication Union. Definition of categories of speech transmission quality. *ITU-T Recommendation G.109*, September 1999.
- [54] International Telecommunication Union. Perceptual evaluation of speech quality (PESQ) , an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *ITU-T Recommendation P.862*, February 2001.
- [55] International Telecommunication Union. Asymmetrical digital subscriber line (ADSL) transceivers - 2 (ADSL2). *ITU-T Recommendation G.992.3*, November 2002.
- [56] International Telecommunication Union. One-way transmission time. *ITU-T Recommendation G.114*, May 2003.
- [57] International Telecommunication Union. Packet-based multimedia communications systems. *ITU-T Recommendation H.323*, July 2003.
- [58] International Telecommunication Union. Definition of quality of experience (qoe). *ITU-T SG12 D.197 (P. Coverdale)*, March 2004.
- [59] International Telecommunication Union. E-model: Additivity of burst loss impairment with other impairment types. *Source: Ruhr-University Bochum (A. Raake), ITU-T Delayed Contribution 221*, March 2004.
- [60] International Telecommunication Union. Single-ended method for objective speech quality assessment in narrow-band telephony applications. *ITU-T Recommendation P.563*, May 2004.

- [61] International Telecommunication Union. The E-model, a computational model for use in transmission planning. *ITU-T Recommendation G.107*, March 2005.
- [62] International Telecommunication Union. Report on a new subjective test on the relationships between listening, talking and conversational qualities when facing delay and echo. *Source: France Télécom R&D (M. Guéguin), ITU-T Delayed Contribution 45*, January 2005.
- [63] Ute Jekosch. *Voice and Speech Quality Perception – Assessment and Evaluation*. Springer, Berlin, 2005.
- [64] Wenyu Jiang and Henning Schulzrinne. Comparison and optimization of packet loss repair methods on VoIP perceived quality under bursty loss. In *Proc. Int. Workshop Network and Operating Systems Support for Digital Audio and Video NOSSDAV*, pages 73–81, Miami Beach, FL, May 2002.
- [65] Demetrios Karis. Evaluating transmission quality in mobile telecommunication systems using conversation tests. In *Human Factors Society 35th Annual Meeting*, volume 1, pages 217–221, Santa Monica, CA, 1991.
- [66] Lars-Åke Larzon, Mikael Degermark, Stephen Pink, Lars-Erik Jonsson, and Godred Fairhurst. The lightweight user datagram protocol (UDP-Lite). *Request for Comments (Standards Track) RFC 3828, Internet Engineering Task Force*, July 2004.
- [67] Spiro Kioussis. Interactivity: a concept explication. *New Media Society*, 4:355–383, September 2002.
- [68] Nobuhiko Kitawaki and Kenzo Itoh. Pure delay effects on speech quality in telecommunications. *IEEE J. Sel. Areas Comm.*, 9(4):586–593, May 1991.
- [69] Thomas J. Kostas et al. Real-time voice over packet-switched networks. *IEEE Network*, 12(1):18–27, Jan./Feb. 1998.
- [70] Lorient. *Das Frühstücksei*. Diogenes, Zürich, 2003.
- [71] Mathworks. Matlab reference guide. The MathWorks, Inc., Natick, MA., 1998.
- [72] Sebastian Möller. *Assessment and Prediction of Speech Quality in Telecommunications*. Kluwer Academic Publishers, Boston, MA, 2000.
- [73] Sue B. Moon, Jim Kurose, and Don Towsley. Packet audio playout delay adjustment: Performance bounds and algorithms. *ACM/Springer Multimedia Systems*, 6:17–28, January 1998.

- [74] Gail E. Myers and Michele T. Myers. *The Dynamics of Human Communication: A Laboratory Approach*. McGraw-Hill, New York, NY, 1991.
- [75] Frank Ohrtman. *WiMAX Handbook*. McGraw-Hill Professional, New York, NY, 2005.
- [76] Ghyslain Pelletier. Robust header compression (rohc): Profiles for user datagram protocol (udp) lite. *Request for Comments (Standards Track) RFC 4019, Internet Engineering Task Force*, April 2005.
- [77] Colin Perkins, Orion Hodson, and Vicky Hardman. A survey of packet loss recovery techniques for streaming audio. *IEEE Network*, 12(5):40–48, Sept./Oct. 1998.
- [78] Jon Postel. User datagram protocol. *RFC 768*, August 1980.
- [79] Jon Postel. Internet protocol. *RFC 791*, September 1981.
- [80] Jon Postel. Transmission control protocol. *RFC 793*, September 1981.
- [81] G. Psathas. *Conversation Analysis: The Study of Talk-in-Interaction*. Sage, London, 1995.
- [82] Alexander Raake. *Assessment and Parametric Modelling of Speech Quality in Voice-over-IP Networks*. PhD thesis, Ruhr-University Bochum, 2004.
- [83] Alexander Raake. Predicting speech quality under random packet loss: Individual impairment and additivity with other network impairments. *ACUSTICA/Acta Acustica*, 90(6):1061–1083, Nov/Dec 2004.
- [84] Alexander Raake. *Speech Quality of VoIP: Assessment and Prediction*. John Wiley & Sons, Chichester, UK, 2006.
- [85] Sheizaf Rafaeli. Interactivity: From new media to communication. In *Sage Annual Review of Communication Research: Advancing Communication Science*, volume 16, pages 110–134. Sage, Beverly Hills, CA., 1988.
- [86] Ramachandran Ramjee, Jim Kurose, Don Towsley, and Henning Schulzrinne. Adaptive playout mechanisms for packetized audio applications in wide-area networks. In *Proc. IEEE INFOCOM*, volume 2, pages 680–688, 1994.
- [87] Sebastian Rehmann, Alexander Raake, and Sebastian Möller. Parametric simulation of impairments caused by telephone and voice over IP network transmission. In *Proc. EAA 2002 – Forum Acusticum*, volume 1, Sevilla, Spain, September 2002.

- [88] Peter Reichl and Florian Hammer. Hot discussion or frosty dialogue? Towards a temperature metric for conversational interactivity. In *8th International Conference on Spoken Language Processing (ICSLP/INTERSPEECH 2004)*, pages 317–320, Jeju Island, Korea, October 2004.
- [89] Peter Reichl, Gernot Kubin, and Florian Hammer. A general temperature metric framework for conversational interactivity. In *Proc. 10th International Conference on Speech and Computer (SPECOM 2005)*, Patras, Greece, October 2005.
- [90] Søren Andersen, Alan Duric, Henrik Astrom, Roar Hagen, W. Bastiaan Kleijn, and Jan Linden. Internet low bit rate codec (iLBC). *Request for Comments (Standards Track) RFC 3951, Internet Engineering Task Force*, December 2004.
- [91] D. L. Richards. *Telecommunication by Speech*. Butterworths, London, 1973.
- [92] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler. SIP: Session initiation protocol. *Request for Comments (Standards Track) RFC 1883, Internet Engineering Task Force*, June 2002.
- [93] Sheldon M. Ross. *Stochastic Processes*. Wiley, New York, NY, 1996.
- [94] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, 1974.
- [95] Henning Schulzrinne. Converging on internet telephony. *IEEE Internet Computing*, 3(3):40–43, May/June 1999.
- [96] Henning Schulzrinne, Stephen L. Casner, Ron Frederick, and Van Jacobson. RTP: A transport protocol for real-time applications. *Request for Comments (Standards Track) RFC 3550, Internet Engineering Task Force*, July 2003.
- [97] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3/4):379–423, 623–656, 1948.
- [98] J. Short, E. Williams, and B. Christie. *The Social Psychology of Telecommunications*. Wiley, London, 1976.
- [99] Amoolya Singh, Almudena Konrad, and Anthony D. Joseph. Performance evaluation of UDP lite for cellular video. In *Proc. Int. Workshop Network and Operating Systems Support for Digital Audio and Video NOSSDAV*, pages 117–124, Port Jefferson, NY, June 2001.

- [100] Johan Sjöberg, Magnus Westerlund, Ari Lakaniemi, and Qiaobing Xie. Real-time transport protocol (RTP) payload format and file storage format for the adaptive multi-rate (AMR) and adaptive multi-rate wideband (AMR-wb) audio codecs. *Request for Comments (Standards Track) RFC 3267, Internet Engineering Task Force*, June 2002.
- [101] Skype. <http://www.skype.org>.
- [102] Jonathan S. Steuer. Defining virtual reality: Dimensions determining telepresence. *Journal of Communication*, 42(4):7393, 1992.
- [103] Keith Stowe. *Introduction to Statistical Mechanics and Thermodynamics*. Wiley, New York, NY, 1983.
- [104] Paul ten Have. *Doing Conversation Analysis: A Practical Guide*. Sage, London, 1999.
- [105] Peter Vary and Rainer Martin. *Digital Speech Transmission Enhancement, Coding and Error Concealment*. John Wiley & Sons, Chichester, UK, 2006.
- [106] Stephan Wiegelmann, Sebastian Möller, and Ute Jekosch. Scenarios for economic conversation tests in telephone speech quality assessment. In *Joint Meeting ASA/EAA/DEGA, Forum Acusticum 1999, Acta acust. 85 Suppl. 1, 48*, Berlin, Germany, 1999.

Biography

Florian Hammer was born in Enns, Austria, in 1974. He received his Dipl.Ing. degree in Electrical Engineering in 2001, and his Dr.techn. degree (with distinction) in 2006, both from the University of Technology at Graz, Austria. During his diploma thesis studies, he worked at the Center for Research in Electronic Art Technology (CREATE), Univ. of California at Santa Barbara. In 2001, Mr. Hammer joined the Telecommunications Research Center Vienna where he focussed on the perceptual speech quality of Voice-over-IP systems. His research interests include the perceptual quality of multimedia communication systems, methodologies for subjective quality measurement and the interactivity of conversations. In his spare time, Florian Hammer is a Singer-Songwriter playing the guitar and the piano. His live-performances include vocal and guitar improvisations using electronics.