

PhD Thesis

QoE Analysis for Interactive Internet Applications in the Presence of Delay

Sebastian Egger

Signal Processing and Speech Communication Laboratory
Graz University of Technology



Supervisor: Univ.-Prof. Dipl.-Math. Dr. Peter Reichl

Examiners:

Univ.-Prof. Dipl.-Ing. Dr.techn. Gernot Kubin, TU Graz

Prof. Dr.-Ing. Sebastian Möller, TU Berlin

Univ. Prof. Dipl.-Math. Dr. Peter Reichl, TU Graz

Graz, June 2014

This work has been conducted at the Competence Center FTW Forschungszentrum Telekommunikation Wien GmbH, which is funded within the program COMET - Competence Centers for Excellent Technologies by BMVIT, BMWA, and the City of Vienna. The COMET program is managed by the FFG.

Supervisor at FTW: Peter Reichl and Raimund Schatz

This thesis has been prepared using \LaTeX .

Kurzfassung

Quality-of-experience (QoE) von über TCP/IP Netze übermittelten, interaktiven Applikationen gewinnt zunehmend an Bedeutung. Die QoE solcher Applikationen wird hauptsächlich von TCP/IP Netzen inhärenten, übermittlungsbedingten Verzögerungen bestimmt. In diesem Kontext identifiziert die vorliegende Arbeit den Anfrage - Antwort Zyklus als Gemeinsamkeit interaktiver Internet Applikationen und zeigt dass übertragungsbedingte Verzögerungen diesen Zyklus erheblich stören. Anhand zweier prototypischer Anwendungen, interaktive Internet Telefonie und browserbasierte Applikationen, wird der Einfluss von Verzögerungen auf die QoE analysiert. Eine Analyse der Oberflächenstruktur von Konversationen im Kontext interaktiver Internet Telefonie identifiziert kommunikative Probleme und Veränderungen in der Gesprächsstruktur die durch die Verzögerungen ausgelöst werden. Basierend auf diesen Ergebnissen werden zwei Konversationsmetriken abgeleitet, die den Einfluss von Verzögerungen auf vermittelte Interaktionen abbilden. Diese beiden Metriken werden als zusätzliche Vorhersageparameter in eine erweiterte Version des E-Models integriert und es wird gezeigt, dass diese zusätzlichen Parameter die Vorhersagegenauigkeit des E-Models erheblich verbessern.

Für browserbasierte Applikationen wird eine neue subjektive Test Methodologie entwickelt, die einen realistischen Ablauf von Web-Browsing Sessions garantiert und eine Atmosphäre interaktiver flow-experience kreiert. Daten zweier Laborstudien and eines Feldversuchs zeigen, dass diese Testmethodologie im Stande ist in zwei verschiedenen Test Kontexten zuverlässige und konsistente Testergebnisse zu liefern. Bezüglich des Zusammenhangs von Wartezeit und QoE im Kontext von browserbasierten Applikationen wird die WQL Hypothese postuliert: "Der Zusammenhang von **W**artezeit und resultierender **Q**oE ist **L**ogarithmisch." Für File Downloads und einfaches Web Browsing kann die WQL Hypothese anhand von Daten dreier Studien verifiziert werden, für den Anwendungsfall komplexen Web Browsers muss sie aber verworfen werden. Eine nachfolgende Analyse identifiziert praktische Probleme welche die Anwendung der WQL Hypothese auf komplexes Web Browsing erschweren. Diese Analyse zeigt auch, dass subjektiv empfundene Ladezeiten von Web Seiten durch den Interaktionsprozess beeinflusst werden und als zusätzlicher Eingangsparameter für QoE Modelle verwendet werden können. Abschließend wird ein Qualitätsperzeptionsmodell entworfen, dass Aspekte von Interaktionsqualität in den Formationsprozess von Qualität inkludiert und Benutzer (Re-)Aktionen anhand bestimmter Eingangssignale als aktive Ausgangssignale beschreibt.

Abstract

Quality-of-experience (QoE) of interactive applications transmitted over TCP/IP networks has recently gained considerable attention, and is mainly influenced by transmission delays due to TCP/IP's retransmission characteristic. This thesis shows that interactive Internet applications share the commonality of a recurring request-response cycle that is highly vulnerable to such transmission delays. In the context of two prototypical services, interactive Internet telephony and browser based applications, the impact of transmission delays on QoE is analysed. In terms of interactive Internet telephony, a surface structure analysis of delay impaired voice calls reveals several changes in conversation behaviour caused by the delay. From this analysis, two conversational metrics are derived that capture the influence of delay on human-to-human conversations. Using these metrics as additional input parameters, an update to the E-Model is proposed that enhances prediction performance considerably. For browser based applications, a novel subjective testing methodology is presented that establishes a realistic flow-experience in the resulting web browsing sessions. Data from two lab studies and a field trial proves the ability of this test methodology to provide reliable and consistent results across different contexts. In terms of the relationship between waiting time and QoE for browser based applications, this thesis postulates the WQL hypothesis: the relationship between "Waiting time and resulting QoE is Logarithmic". With the acquired data from the three studies, the WQL is verified for file downloads and simple web browsing. Contrary, in the context of complex web browsing the WQL has to be rejected. A following analysis reveals several challenges and practical issues that complicate the use of the WQL for this service. Additionally, it identifies the subjectively perceived page-load-time as an interaction based measure of waiting time and promising input parameter for novel QoE models. Finally, a human perception model, that considers interaction quality performance aspects in the quality formation process, and that explains (re-)actions to (conversational) input signals in the form of active output signals is derived.

Acknowledgment

Writing a PhD thesis in an applied research environment, like the one at the Telecommunications Research Center Vienna (ftw.), poses particular challenges and is a demanding endeavour. Thus, I am grateful that a number of people have supported me in this endeavour over the last couple of years.

A majority of this work has been conducted within the ACE 2.0 and ACE 3 projects at FTW. A1 Telekom Austria AG, Telekom Austria Group AG, Vodafone Group Services Limited and Vodafone Germany mad these projects happen by bringing in their real world problems in order to identify research questions with practical relevance and by contributing considerable resources to the projects. Thanks to their valuable input this thesis evolved far beyond a purely academic exercise and provides actionable guidelines that can be used for the optimization and parametrization of existing mobile and fixed line data networks.

On a personal level, I would like to thank my supervisor Peter Reichl who encouraged me to follow my initial interest in combining knowledge from Sociology, Psychology and the Technical sciences and applying inter-disciplinary concepts to quality related deficiencies in telecommunications. In addition to plentiful discussions on the objectives of this thesis in an early stage, the invitation to his SISCOM research chair at Université Européenne Rennes and the inspiring conversations we had there helped a lot to focus on the important contributions of this work.

In my daily work at FTW, the team of the three ACE projects was of great help in acquiring all the data I have used throughout this thesis. In particular, Ronny Fischer, Kathrin Masuch and Andreas Sackl were extremely supportive and flexible in preparing, last minute problem fixing and executing numerous user studies within these projects.

Thanks to Raimund Schatz' efforts and patience my scientific writing skills have improved to a level of (somehow) clearly and distinctly readable English. He was also of great help in several discussions regarding the framework of this thesis and supportive in promoting my research interests in application oriented projects.

Further, I would like to thank Gernot Kubin and Sebastian Möller for their

critical questions in high level discussions regarding the research goals of this work and their efforts in evaluating this thesis. Thanks to Sebastian Möller's support I have been able to actively participate in the QoE related research community on several occasions.

In terms of data analysis and modelling Tobias Hossfeld was a great mentor regarding methods and approaches for digging in the data pile and identifying relationships between variables that were not always obvious.

While my research stay at TU Berlin several discussions with Alexander Raake, on the psychological and perceptual principles involved in QoE perception, helped me a lot in classifying my research contributions in the bigger picture of QoE formation and modelling.

Although professional knowledge and advices are important for a successful PhD completion, the importance of mind distraction can not be ranked high enough for successfully mastering this endeavour. Therefore, I want to thank Matthias Baldauf, Andreas Berger, Pedro Casas, Marcin Davies, Roland Tresch and Danilo Valerio for making it possible to maintain the balance between serious work, relaxing discussions and fun activities in a stressful environment.

Most of all, I would like to thank my parents and Lisa for their love, support and patience during the last years.

Graz, June 2014

Sebastian Egger

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Graz, June 20th 2014

(Signature)

Contents

1	Introduction	1
1.1	Motivation and Background	1
1.1.1	The ACE Project Series	4
1.1.2	Research Stays at International Research Institutions	5
1.2	Scientific Contribution	5
1.3	Outline of the Thesis	10
2	QoE for Interactive Internet Applications: Background	13
2.1	The Concept of QoE	14
2.2	Definition of Interactive Internet Applications and Selection of Analysed Services	17
2.2.1	Characteristics of Interactive Applications	17
2.2.2	Interactivity and Delay Impairments	22
2.3	QoE Assessment Methodologies	27
2.3.1	Subjective Speech QoE Assessment Methodologies	30
2.3.2	Subjective Web QoE Assessment Methodologies	33
2.4	Challenges and Requirements for QoE Assessment Methodologies for Interactive Internet Applications	35
3	Internet Telephony	37
3.1	Background: Perceived Quality, Conversational Analytics and Communication Theory	38
3.1.1	Impact of Delay on Perceived Quality	40
3.1.2	Conversation Tests and Delay Impaired Conversations	45
3.1.3	Communication Theory based Considerations	51
3.1.4	New Conversational Metrics (<i>UIR</i> , I^3R)	53
3.2	Subjective Experiments	56
3.2.1	Technical Setup	56

3.2.2	Tasks and Test Procedure	57
3.2.3	Result Analysis	58
3.3	Interruption Metrics, Delay and QoE	67
3.4	Conclusion and Lessons Learned	73
3.4.1	Appropriateness of New Interruption Metrics to Capture the Delay Impact on Conversations	73
3.4.2	Updated Delay Thresholds	74
3.4.3	New Interruption Metrics as Model Factors	75
4	Browser Based Applications over HTTP	76
4.1	Related Work on QoE for Browser Based Applications	77
4.2	Subjective Experiments	80
4.2.1	Novel Subjective Testing Methodology and Related Tasks . . .	80
4.2.2	Test Content	82
4.2.3	Test Facilities and Study Setup	83
4.2.4	Verification of the Novel Subjective Testing Methodology . . .	87
4.3	Modelling QoE for Browser Based Applications by Identifying Fun- damental Relationships between QoE and QoS	90
4.3.1	Logarithmic Relationships – The Law of Weber-Fechner	90
4.3.2	Exponential Relationships – The IQX Hypothesis.	91
4.3.3	QoS equals Time for Browser Based Applications	92
4.3.4	Time Perception in Psychology	93
4.3.5	Fundamental Relationships in Human Time Perception	96
4.4	Verifying the WQL Hypothesis in Browser Based Applications	96
4.4.1	File Downloads and Simple Web Browsing	97
4.4.2	Data from Related Work	100
4.4.3	Complex Web Browsing	102
4.4.4	Perceptual Challenges and Practical Issues for the Application of the WQL to Complex Web Browsing	103
4.5	Conclusion and Lessons Learned	111
4.5.1	Novel QoE Assessment Methodology for Web Browsing	111
4.5.2	QoE Modelling for Browser Based Applications	112
4.5.3	Challenges and Practical Issues for Modelling QoE Based on Waiting Times for Complex web browsing	112

5	QoE Perception and Formation for Interactive Internet Applications	114
5.1	Quality Formation Process for Static Media Experiences	115
5.2	Interaction Performance Aspects	117
5.3	Perception Model for Interacting Entities	119
5.4	Quality Formation Process for Interactive Media Experiences	122
6	Conclusions and Future Work	125

List of Figures

1.1	Overview of publications by the author	9
1.2	Structure of the Thesis	10
2.1	QoE influence factors belonging to context, human user, and the technical system itself.	16
2.2	Illustration of a series of requests and responses throughout an interaction that constitute the request-response pattern	20
2.3	Enlarged illustration of an interactivity constituting request-response pattern	21
2.4	Recommended delay categories for different applications from [116] .	24
2.5	Overview of the human internal timing systems	25
2.6	Existing subjective QoE assessment methodologies standardised by ITU-T, taken from [124].	29
2.7	a) Waiting times related to request-response patterns in web browsing [126] and b) a web session as series of page views with different waiting times.	34
3.1	Categorisation of related work on the influence of delay in telecommunication systems along analysis methods and year of publication .	39
3.2	QoE (MOS) vs. transmission delay from related work for SCT and free conversation scenarios of comparable conversational interactivity	42
3.3	QoE (MOS) vs. transmission delay from related work for RNV and comparison task scenarios of comparable conversational interactivity .	43
3.4	Unintended Interruptions Rate UIR	54
3.5	Interruptive (and) Intented Interruptions Rate I^3R	55
3.6	Testbed at FTW's i:Lab	57

3.7	Subjective quality ratings (MOS) for different transmission delays as acquired in Study 1 and Study 2 for (a) and normalized to the ratings of SCT ₁ for (b).	60
3.8	Speaker alternation rate (SAR) vs. one-way delay	62
3.9	Mutual silence (MS) and double talk (DT) vs. one-way delay	63
3.10	Active (AIR) and passive interruption rate (PIR) vs. one-way delay	64
3.11	Interruptive (and) intended interruption rate (I ³ R) and unintended interruption rate (UIR) vs. One-Way Delay	65
3.12	The ratio between interruptive (and) intended interruption rate (I ³ R) and unintended interruption rate (UIR) in [%].	66
3.13	Idd values predicted by ITU-T Rec. G.107 versus Idd' values	69
3.14	Idd' predictions from the extended E-model	71
3.15	Idd' values calculated with the E-model modification from Equation (3.6)	72
4.1	Technical setup of the two lab studies (study A and B).	84
4.2	Technical setup of the field trial (study C).	86
4.3	Comparing rating data from lab and field environments for highly interactive web browsing ((a) across all websites) and file downloads (b).	88
4.4	Perceived duration vs. objective duration from [191].	94
4.5	Download of files of various sizes obtained in three subjective user studies conducted in 2009 (study A), 2010 (study C) and in 2011 (study B), respectively (DL task).	98
4.6	User satisfaction for various constant page-load-times (PLT task).	100
4.7	Results from [169] with logarithmic fittings applied	101
4.8	Web browsing with downlink bandwidth limitation instead of instrumented constant page-load-times.	103
4.9	The cumulative distribution function of application-level page-load-times over one browsing session	105
4.10	Perceptual events in a web page view cycle from the end-user point of view. The lower timeline (blue) displays related technical events on application or network level.	108
4.11	Perceived subjective vs. application-level PLT for different pages.	110
5.1	Quality formation process as depicted in	115

5.2	Taxonomy of influence factors, interaction performance aspects and quality features	118
5.3	Perception model that allows to capture interaction performance aspects from an interaction between two or more entities	121
5.4	Integration of the proposed perception model for interacting entities into the quality formation process	123

Chapter 1

Introduction

1.1 Motivation and Background

One everlasting problem since the formation of the Internet has been the steady increase of traffic volumes. While in the late 1990's and early 2000's capacities in the wired last mile networks have been the bottleneck which has been eradicated by high xDSL, cable and fiber-to-the-home penetration (in the western world), the recent bottleneck is mobile broadband access which is growing at large due to the fast spread of mobile computing on-the-go, smartphones and tablets [73]. This growth in sheer (mobile) number of devices additionally introduces a large amount of small data transmissions at the edge of the network which poses a threat towards latency and mandated Quality-of-Service for mobile devices [129]. These technical challenges become especially eminent in the context of interactive web applications and file downloads, where high latency and long waiting times directly translate into user annoyance and churn.

From a network provider's point of view this leads to a highly demanding situation. On the one hand, traffic volume in their mobile data networks is growing at large and therefore routing challenges due to latency issues introduced by largely growing smartphone and tablet traffic get apparent again, which call for investments in high performance networks. On the other hand economical constraints, stemming from the highly competitive market with decreasing average revenue per user (ARPU), are tight in order to stay price competitive. Consequently, operators have to trade off between investing in their network infrastructure at minimal costs and in the same moment ensuring sufficiently performing network quality for satisfying their customer base.

In this context Quality-of-Experience (QoE) is currently receiving an immense increase in interest both from an academic and industrial perspective as a new attempt to describe the qualitative performance of communication systems and applications not only in terms of traditional Quality-of-Service (QoS), but to link it as closely as possible to the subjective perception of the end user. This user centricity is also reflected in the current working definition of QoE from [1] which reads as follows: *Quality of Experience is the degree of delight or annoyance of the user of an application or service. It results from the fulfilment of his or her expectations with respect to the utility and / or enjoyment of the application or service in the light of the user's personality and current state.* Bridging this user centric concept with the technical network perspective from operators leads to the following question: *Which network quality (QoS) is sufficient to ensure decent QoE?* For answering this question subjective tests where users experience different quality conditions and subsequently report on their associated experiences are the key. In terms of QoE research and related subjective assessment methodologies the QoE domain has so far been dominated by multimedia services. However, these results and methods are mainly targeted towards 'static' media experience and signal fidelity (cf. [114,122,125,131]), hence they do not properly address the growing area of interactive Internet applications. As a result of this focus on multimedia services, attention in QoE research was focused on media fidelity as key determinant of QoE and 'new' impairments of delays and waiting times and their relation towards QoE have received far less attention. This represents a major issue as the growing number of interactive Internet applications are especially prone to delays and waiting times as they deteriorate their interactive nature which is in turn negatively assessed by the end user of such applications. From a technical point this set of problems will be further accentuated: (1) By the increasing use of TCP as transport protocol and the corresponding translation of packet losses and re-orderings into delays and waiting times on the application level. (2) Additional latency issues introduced by the large traffic growth in mobile data networks [129] will impair the networked services.

In order to properly assess and model QoE for these interactive services the two most important challenges amongst others are: First, the QoE concept and the related QoE formation process itself has to be broadened to also consider interactive services, and second novel assessment methodologies addressing the special requirements of interactive QoE tests have to be established. Therefore, current QoE formation processes, as described e.g. in [1], have to be extended such that interactions between two or more entities, in the context of interactive media experiences, can be

properly analysed and incorporated. Such experiences are characterised by a (time) series of input signals and additional interaction processes which are deteriorated by temporal impairments. However, the interaction process and its deterioration can be a novel source of information and therefore serve as an additional input into the quality formation process. Together with the series of media fidelity input signals such an extended QoE formation process outputs then a holistic QoE for interactive applications. Furthermore, for gathering QoE results and data for QoE modelling of interactive applications novel assessment methodologies are needed. Such methodologies have to establish realistic interaction processes for the targeted applications. Only if appropriate interaction processes are established the delay induced degradations of the interactive nature can be sensed, and consequently evaluated by the test subjects, leading to externally valid results for this application category.

In the context of this thesis, the above discussed research challenges have been addressed in several research cooperations: with industry partners in the context of the application oriented ACE project series, and in the context of academic research through scientific cooperations with academic partners. These cooperations, which form the background of this thesis, are described in the following sections.

1.1.1 The ACE Project Series

The majority of the work contained in this thesis has been conducted throughout three projects: ACE, ACE 2.0 and ACE 3.0. Together they form the ACE project series at the telecommunication research center Vienna (FTW)¹. These application oriented projects have been conducted within the COMET framework, which is funded by BMVIT, BMWA, and the City of Vienna. In terms of the research methodology, the ACE projects utilise a strictly user-centric cross-layer approach towards QoE by taking into account relevant influence factors on network, application and user-level. In this context, the author was responsible for voice related work packages which provided data and analysis results for Study 1, described and discussed in Chapter 3. Furthermore, the authors work on QoE evaluation in browser based applications within these projects has led to the development of the subjective testing methodology introduced in Chapter 4. Based on this methodology, numerous laboratory studies, as well as one field trial on browser based applications have provided the quantitative data used for QoE analysis and modelling in Chapter 4.

¹The projects within this series share the following common aims:

Understanding, measuring and managing quality in communication networks has become a vibrant area of applied research. The key reason is that improving service quality directly supports carriers in winning and keeping customers and reduce churn. However, since customers are also the ultimate judges of service quality, it is vital for the industry to move beyond traditional QoS and adopt a more holistic understanding of quality as perceived by end-users. Such a shift towards Quality of Experience (QoE) raises fundamental questions relating to which QoS parameters are truly relevant to users of a given service class, how these parameters can be measured and which quality levels actually define a satisfying user experience.

The ACE project series aims to realise this paradigm shift by investigating the link between technical network parameters and the customer's Quality of Experience (QoE) in the context of mobile and fixed broadband. The ACE projects consist of user-centric as well as network measurement centric activities which together constitute an integrated approach towards QoE assessment and measurement for convergent networks and services.

In particular, the currently running project ACE 3 focuses on the following aspects of broadband QoE: high-speed/LTE scenarios, impact of user terminals (smart devices), convergence, QoE for Web and Cloud services, VoIP and video (adaptive streaming, IPTV) quality.

Its predecessor, ACE 2.0 addressed the following aspects of mobile broadband QoE: impact of user terminals (smart devices), different demographics, convergence, QoE for data services (Web 2.0, file downloads, progressive downloads) as well as the relationship between Customer Experience and QoE.

This description has been acquired from <https://ace.ftw.at/about> on 05-01-2014

1.1.2 Research Stays at International Research Institutions

SISCOM International Research Chair: Parts of the work on the WQL hypothesis have been conducted in the context of a research stay of the author as a visiting researcher at the SISCOM International Research Chair (Prof. Peter Reichl), Université Européenne de Bretagne, Rennes, from July to September 2011.

Telekom Innovation Laboratories Berlin: Refinements of the perceptual model used in Chapter 5 were inspired by several discussions with Alexander Raake, while the author was working as an external researcher at the AIPA group within Telekom Innovation Laboratories Berlin from August to December 2013.

The data of the conversational Study 2 and the respective analysis discussed in Section 3.2 are the result of a close collaboration between FTW and Telekom Innovation Laboratories initiated by the author and ongoing since 2009 with mutual short term visits.

1.2 Scientific Contribution

The scientific objective of this thesis is fourfold: First, a review of related work towards QoE assessment methodologies for interactive Internet applications is used to determine shortcomings of current assessment methodologies. From these shortcomings, requirements for the special needs of QoE assessment methodologies for interactive Internet applications are derived. Second, the interactive target application of Internet telephony is analysed from a communication theoretical viewpoint. The main focus is put on the pragmatic dimension of human-to-human interaction and its alteration due to transmission delays. Results of this analysis are merged with identified shortcomings from the first step and used to derive certain interaction metrics which are captured through subjective experiments and analysed towards their relation to conversational QoE for Internet telephony. Third, as a second application web browsing and file downloads are analysed with respect to waiting times and their relation towards QoE. In order to achieve this, a novel test procedure for browser based interactive applications is derived and used for gathering a comprehensive dataset for analysis. Furthermore, this data is then used to identify fundamental relationships between waiting times and QoE for these applications. Fourth, existing QoE formation models targeted towards static media experiences have to be extended, such that they incorporate interaction performance aspects

and account for recurring (inter)actions of the subjects. Together these objectives will provide answers to the following research question(s):

What impairments on the interactional structure of interactive Internet applications are caused by the transmission delay and what is the impact of transmission delay and these impairments on the QoE of such applications?

- RQ1** Which common interaction patterns can be identified in the interaction structures of different interactive Internet applications?
- RQ2** What is the impact of one-way delay on the pragmatic dimension of human-to-human mediated interaction, and can this impact be quantified in related interactional metrics?
- RQ3** Which interactional metrics can be used to enhance prediction performance of QoE models for human-to-human mediated interaction and how does such a model look like?
- RQ4** What are requirements for subjective testing methodologies that produce reliable and consistent QoE scores for interactive browser based applications?
- RQ5** Can fundamental relationships of human time perception be utilised to model the relationship between waiting time and QoE in browser based applications?
- RQ6** How does a human quality perception model look like that incorporates interactional metrics into the quality formation process and that is able to explain the formation of (inter)actions between interacting entities?

In order to answer these research questions an initial analysis of commonalities for interactive applications reveals request-response (=interaction act) patterns as an interaction structure common to interactive Internet applications in general. This is followed by an analysis how delay and waiting times respectively, impair these interaction patterns. Considering the impairments introduced and their impact on the interaction process itself interaction performance aspects are identified as important factors for capturing the influence of delay impairments on the interaction process. Based on this finding, existing QoE assessment methodologies targeting interactive Internet applications are reviewed regarding their consideration of interaction performance aspects. Furthermore it is revealed which of these aspects are

not sufficiently covered and which requirements have to be met to properly take them into account in interactive QoE assessment.

Regarding Internet Telephony existing QoE assessment methodologies analyse conversational QoE and statistics of the conversational surface structure. However, they fall short in analysing interaction performance aspects and how these aspects are altered by delay impairments. To this end communication theoretic considerations regarding the pragmatic dimension of human-to-human communications are introduced and discussed. Thereby different types of interruptions are identified as key interaction performance metrics, which are severely impacted by transmission delays. Through subjective conversational tests and data gathered with these tests an analysis of conventional surface parameters is achieved and new interactional metrics are presented. This analysis reveals the impact of delay on these metrics and additionally shows how these metrics are related to conversational QoE. Furthermore, these results are used to propose an update to the existing E-model, that incorporates interactional metrics and enhances prediction performance of the model.

For web browsing and file downloads waiting times are the key influence factor regarding interactive QoE of these applications. As for other QoE domains the existence of fundamental relationships is present in related work a (successful) attempt is made to identify a fundamental relationship between waiting time and interactive QoE which is then proposed as the WQL hypothesis (The relationship between **Waiting time** and its **QoE** evaluation on a linear ACR scale is **Logarithmic**). In order to prove this hypothesis reliable QoE data needs to be acquired. As appropriate test methodologies for web browsing are not available, a respective test procedure and tasks that allow to acquire reproducible QoE ratings for this service type are derived and verified. This methodology is then used to acquire QoE data in three subjective studies. Based on these data the WQL hypothesis can be verified for file downloads and simple web browsing tasks. For highly interactive web browsing the hypothesis has to be rejected for numerous reasons which are analysed in detail.

In terms of QoE formation within human subjects, existing models mainly address static media signals (in terms of interaction between two or more signal exchanging entities). Thereby, they fall short in incorporating interaction performance metrics, and they do not consider the inter-relation between (input-)signals and following (re-)actions of the users, which influences QoE as well. An extension of current models of the QoE formation process exemplifies how such interactional aspects can be gracefully included in existing models.

Scientific Publications of the Author and their Contribution within this Thesis

This sections gives an overview of the scientific publications, tutorials and standardisation related work the author of this thesis actively contributed to. The most important publications are summarised according to their contribution to the related chapters as follows.

In the domain of conversational quality and interactive communications the concepts, the work as presented in Chapter 3 is based on, have been published in [2] where the analysis of conversational surface structure was introduced together with the unintended interruption ratio (UIR) and [3] where the dataset was presented and the interruptive (and) intended interruptions rate (I³R) measure was introduced. In addition, the updated version of the E-model as proposed in [4] has been included in a new version of ITU-T Rec. G.107 [123].

Chapter 4 is based on result published in [5] where the application of the WQL to QoE in telecommunication systems was introduced, [6] and [7] described the data acquired in lab and field studies and presented related results and [8] applied the WQL to temporal stimuli in the context of browser based applications. Furthermore, the work on perceptual influence factors in the context of web browsing has been brought to ITU-T Study Group 12 and has been issued as ITU-T Rec. G.1031 [127], and the test methodology for browser based applications has led to the release of ITU-T Rec. P.1501 [128].

The work on QoE foundations and according quality formation used to derive the model of an interactive quality formation process in Chapter 5 was published in [1] where the quality formation process for static media signals that served as basis for the interactive process described in this chapter was derived, [9] refined the foregoing model of the quality formation process and described in detail the perceptual model, which was used for the proposed perception model that explains the formation of interactional quality features and incorporates (inter-)action between interactants.

Figure 1.1 provides an overview of all scientific publications in peer-reviewed conference proceedings, journals and books as well as standardisation contributions where the author was actively involved while working on this thesis. The research studies and methodological contributions are classified according to the major application classes on the x-axis and the related QoE research categories on the y-axis. Application classes are sub-divided into human-to-human interactions, human-to-machine interaction and media delivery and QoE research categories are

Media Fidelity			[40] [41]
	[39] [14]	[28]	[5] ⁴ [21] ^{3,4}
	[20] ³	[44]	[30] [22] [42] [45]
Assesmat Methodologies	[19] ³	[50] ⁴ [15] ⁴ [12] ⁴	[37] [23] [48]
	[49] ^{3,4}	[51] ⁴ [11] ⁴	[32] [33]
		[10] ⁴ [17] ⁴ [18] ⁴ [13] ⁴ [7] ⁴	[34] [24] [38]
Temporal Impairments	[43] ³ [16] ³	[36] [29] ⁴	[31]
	[47] ³ [4] ³	[6] ⁴	[25]
	[3] ³ [16] ³ [2] ³	[26] [8] ⁴	[35] [27] [46]
	Human-to-Human mediated Communication	Human-to-Machine Communication	Media Delivery

Figure 1.1: An overview of peer-reviewed scientific publications, held tutorials and standardisation contributions from the author. Their respective contribution in this thesis is indicated with the notion $[x]^y$ indicating that the scientific publication $[x]$ is discussed in Chapter y .

sub-classified into temporal impairments, contributions towards assessment methodologies and media fidelity. As can be seen, the major focus of the contribution to the scientific community has been in the categories of assessment methodologies and the analysis of temporal impairments. Regarding their contribution in this thesis they are marked with the the notion $[x]^y$ indicating that the scientific publication $[x]$ is discussed in Chapter y and are additionally marked in red for Internet telephony related contributions in Chapter 3 and blue for contributions towards browser based applications in Chapter 4.

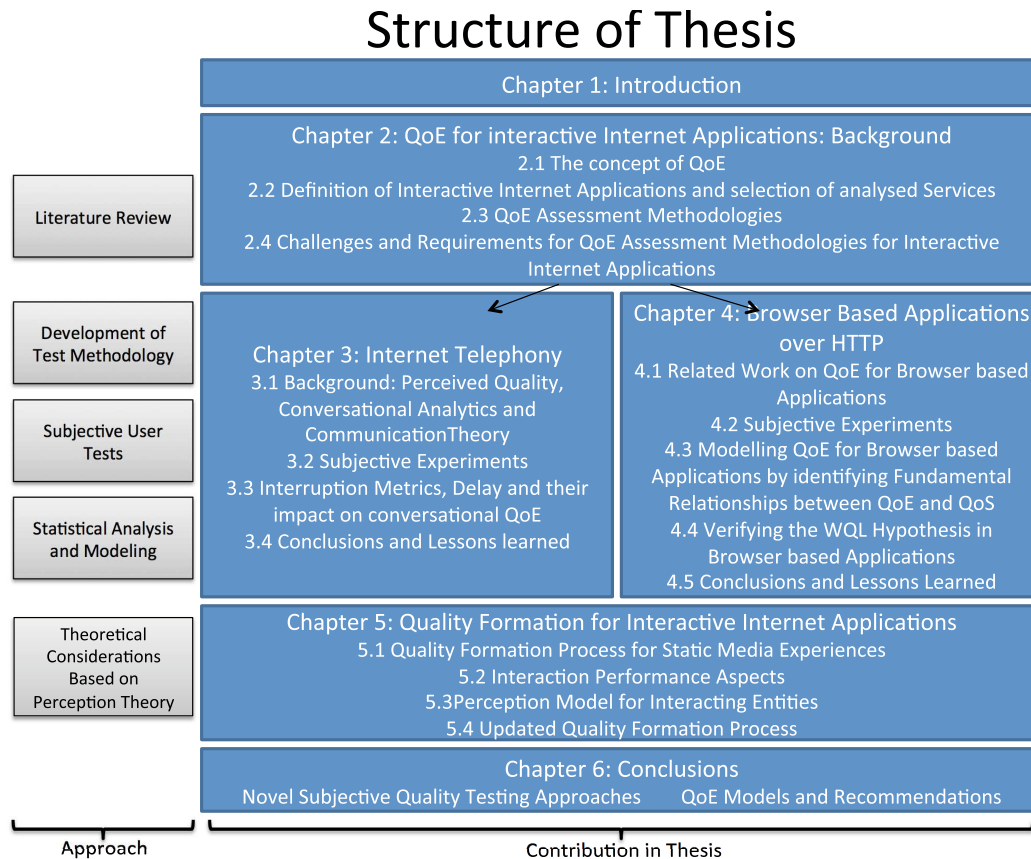


Figure 1.2: Structure of the Thesis

1.3 Outline of the Thesis

The organisation of the thesis is depicted in Figure 1.2. After motivating the focus of the thesis and describing the scientific contribution of the thesis in the current chapter, the second chapter reviews the QoE concept with respect to interactive applications and identifies (interactional) commonalities of interactive applications as the request-response cycle. Furthermore it discusses how these interactions are impaired by transmission delays in communication networks. By analysing related work from psychology on human time perception it is shown that two different human timing systems do process transmission delays or waiting times depending on the delay range they fall into. According to these delay ranges, the application of Internet telephony is selected to study the impact of delay on QoE for delays < 1 s, and browser based applications are selected to study the impact of longer delays, as they are typically impaired by waiting times > 1 s. Furthermore, it discusses existing QoE assessment methods and concludes which challenges and requirements

for QoE assessment have to be considered to properly assess interactive applications and to measure interaction related aspects.

Following, in Chapter 3 conversation analytic approaches towards capturing interactivity and relating it to QoE are reviewed, and blind spots and shortcomings are identified. Based on that, new conversational metrics that are able to identify the influence exerted by delay on human-to-human conversations are proposed. In order to apply these new metrics as additional input parameters for QoE prediction models, subjective tests are conducted which do capture conversational interactivity measures for these modelling purposes. The result data is first analysed with respect to interactional changes related to transmission delay, which reveal interactional problems that are not reflected in the conversational quality ratings. In a second step, this data is used to derive a conversational prediction model that includes interactional metrics as additional input. A following comparison with a state-of-the-art model shows that the prediction performance can be increased by considering interactional metrics as additional input parameter.

The second prototypical application is addressed in Chapter 4 where a novel QoE assessment methodology is introduced that considers the interactive nature of browser based applications properly. By comparing results from lab studies with results from a field trial this novel methodology is verified to deliver reliable and externally valid results. In terms of modelling the relationship between waiting times and QoE the WQL hypothesis is postulated and validated for simple web usage scenarios. The attempted extension of the WQL towards more complex web browsing identifies challenges and practical issues on a perceptual and technical level which lead to a rejection of the hypothesis for QoE prediction in case of complex web browsing. However, a thorough analysis of the resulting interactions on the network and application layer, and respective perceptual events, showed that numerous practical issues and challenges on a technical as well as on a perceptual level exist. As a result, the subjectively perceived page-load-time could be identified as an interactional metric, which could be used as potential input parameter for QoE modelling approaches.

In order to properly include interactivity and related measures into quality formation models, Chapter 5 identifies five interaction performance aspects that should be considered in a QoE perception model to capture interaction related impairments. Following, a perception model is proposed that allows to detect these interaction performance aspects for interactions between two or more entities. In addition, it is also able to describe (re-)actions to conversational input signals in the form of

output signals, which then serve as input signal(s) for the other interacting entity and vice versa. Thereby, interactional processes between two or more entities can be explained. This perceptual model is then integrated into an existing model of the quality formation process that was initially proposed for static input signals. By this modification, the updated model can be used to describe also interactive quality formation processes and it considers interaction performance aspects for the formation of its QoE output.

Finally Chapter 6 summarises the contributions of the work conducted in this thesis.

Chapter 2

QoE for Interactive Internet Applications: Background

Taking the subjective view of the end-user into consideration for dimensioning, maintaining and operating telecommunication networks requires the understanding of QoE influence factors and the relationships between technical QoS parameters and QoE. In order to study these relationships and influence factors, subjective QoE testing and according analyses are needed. Interactive applications are especially interesting in these respects as they pose special requirements towards QoE assessment methodologies: 1) they have to be designed in a way that the interactive nature of the targeted application is ensured throughout the QoE assessment. 2) interaction cues have to be monitored and / or recorded in addition to QoS parameters and subjective QoE ratings. Such data is essential for analysing interaction patterns and their deterioration due to impairments. Such an analysis is necessary for revealing additional influence factors and dimensions not captured in conventional test and analysis methodologies.

First, the present chapter provides a review of the historical development of the QoE concept and an analysis of the current status of the concept and the related framework. In the next step, commonalities of interactive applications are introduced, together with a discussion of different delay ranges and how sensitive different interactive application categories are to these ranges. In addition two interactive Internet applications are selected to serve as prototypical applications for the analysis of the transmission delay influence on QoE. Furthermore the third section analyses existing QoE methodologies from standardisation bodies, as well as from related work and respective analysis methods for their appropriateness to assess

QoE of interactive applications. Based on these analysis, challenges and requirements for subjective assessment and analysis methodologies, which have to be met for a holistic analysis of QoE for interactive Internet applications, are summarised.

2.1 The Concept of QoE¹

For a better understanding of the QoE concept it is helpful to make a brief review of the recent history of communications quality assessment. In the early 1990s, the notion of Quality of Service (QoS) attracted considerable attention in telecommunications, nurtured by articles such as [173], in which the authors described their conceptual model of service quality and in which the ultimate instance for the service quality judgement was the respective customer. This user or customer centricity is also reflected in the ITU-T definition of QoS, which underlines the subjective roots of the service quality concept despite being oriented rather towards the view of a telecommunications provider or manufacturer:

Quality of Service is the totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service. [121]

However, contrary to this original definition, most QoS-related work actually focused on the investigation of purely technical, objectively measurable network and service performance factors such as delay, jitter, bitrate, packet loss – effectively reducing quality to a purely technology-centric perspective (cf. [1], [179]). This focus shift towards network- and system-level performance parameters is also reflected in the QoS definitions that became dominant in the networking community as the following QoS definition by the IETF exemplifies:

"A set of service requirements to be met by the network while transporting a flow." [75]

Due to this deviance from its subjective focus the concept of QoS got less attractive to domains such as audio and video research, where historically subjective quality assessment played a major role in comparing, e.g. codec performance. A

¹This section is based on original work from the author with adaptations as published in [21], where he was responsible for Section 3.2 and actively contributed text and figures to Section 3.1 and Section 3.4 and original work from the author with adaptations as published in [9] where he contributed text and figures throughout the respective chapter.

countermovement gained momentum which took up the notion of *Quality of Experience*, which was introduced in the context of television systems by [147]². The notion of QoE was rapidly adopted not only in the context of mobile communications (cf. [194]) but also in the domains of audio and video quality assessment (cf. [165, 176, 186, 214]). However, each service type (voice, video, data services, etc.) tended to develop its own QoE community with its own research tradition. In addition it has to be noted that some domains do not even use the notion of QoE but rather use the terms "subjective quality" or "user-perceived quality" although using the conceptual model that goes back to QoE (cf. [60, 62, 103]).

This has resulted in a number of parallel attempts to define QoE (as outlined in [179, 181]), accompanied by an equally large number of QoE frameworks and taxonomies (see [149] for a comprehensive overview). However, today the definition by ITU-T Rec. P.10 (Amendment 2, 2008) is still the most widely used formulation of QoE, defining the concept as:

QoE is the overall acceptability of an application or service, as perceived subjectively by the end user. [119]

Note 1: includes the complete end-to-end system effects.

Note 2: may be influenced by user expectations and context.

During discussions at the Dagstuhl Seminar 09192 in May 2009 (cf. [163]) it was pointed out that among others the notion of "acceptability" in the above definition, is somehow problematic as the concept of acceptability demands a certain (usage) context of the service (cf. [6]) to yield reproducible results across different assessments of QoE or acceptability respectively. In addition, a new definition of acceptability was proposed as follows:

Acceptability is the outcome of a decision [yes/no] which is partially based on the Quality of Experience. [163]

In this respect, ITU-T Rec. P.10 captures the essence of QoE by highlighting some of its main characteristics: subjectivity, user-centricity, and multi-dimensionality. Particularly concerning the latter aspect, most frameworks and definitions found in the literature highlight the fact that QoE is determined by a number of hard and soft *influence factors*, attributable either to the technical system, the usage context,

²It can not be figured out with 100% certainty who introduced the notion of QoE into the domain of multimedia quality assessment, however the work by [147] is one of the earliest ones that used the notion in the same understanding as it is still used nowadays.

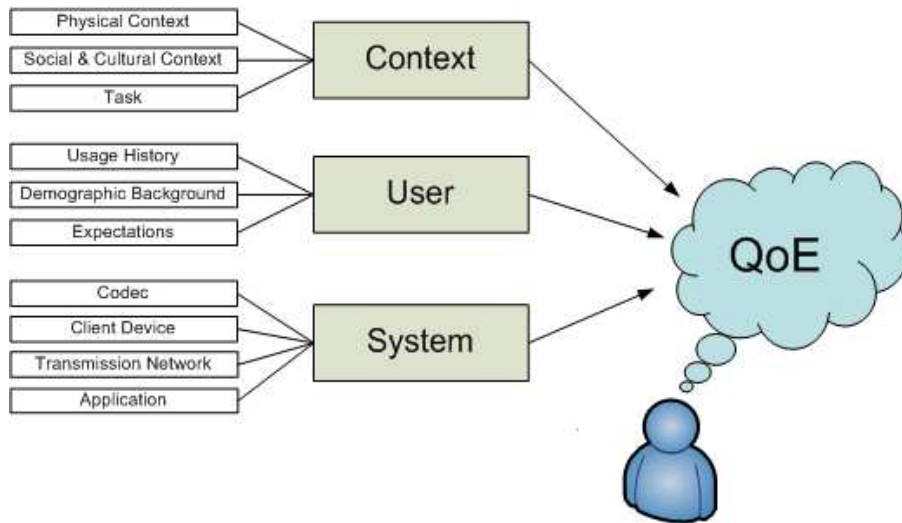


Figure 2.1: QoE influence factors belonging to context, human user, and the technical system itself.

or the user him/herself (see Figure 2.1). This means that whether a user judges the quality of, e.g. a mobile video service as good (or even excellent) not only depends on the performance of the technical system (including traditional network QoS as well as client and server performance),³ but to a large extent also on the context (task, location, urgency, etc.), the user himself (expectations, personal background, etc.), as well as the interaction process with the system (or another user). The resulting level of complexity and broadness turns reliable and exact QoE assessment into a hard problem. Indeed, this is also one of the main reasons why, as of today, the scientific QoE community remains fragmented and has not agreed on a common QoE definition as well as a unified QoE framework yet.

As one of the most recent initiatives, the COST Action IC 1003 has published a QoE definition whitepaper to further advance the required convergence process regarding this subject [1]. Version 1.2 of this whitepaper defines:

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

³Note that the technical system generally comprises of a chain of components (sender, transmission network elements, receiver) that connect the service provider with the end-user. All these elements can influence technical QoS (and thus QoE) on different layers, predominantly in terms of network- and application-level QoS.

Thus, it advances the ITU-T definition by going beyond merely binary acceptability and by emphasising the importance of both pragmatic (utility) and hedonic (enjoyment) aspects of quality judgment formation⁴. In addition to QoE influence factors, the whitepaper also highlights the importance of *QoE features*, i.e. recognised characteristics of the individual's experience that contribute to service quality perception. These features can be classified on four levels: direct perception (e.g. colour, sharpness, noisiness), usage situation (e.g. accessibility, stability), service (e.g. usability, usefulness, joy), and interaction (e.g. responsiveness, conversation effectiveness). The latter one is of particular interest in the context of this thesis and the following section will further discuss interaction and its relation to QoE.

2.2 Definition of Interactive Internet Applications and Selection of Analysed Services

In order to clarify how Internet applications can be categorised as interactive, how QoE of such applications is affected by different transmission delays, and which interactive Internet applications are of particular interest for the aim of this thesis, Section 2.2.1 discusses which (inter-)actions typically take place via the communication channel and derives common characteristics which are shared across interactive applications. Then, Section 2.2.2 discusses delay ranges that typically appear in communications systems and resulting deteriorations of QoE for interactive applications. Furthermore, it shows that different human timing systems are involved (for different delay ranges) in the regulation of interaction behaviour as well as for QoE formation based on waiting times. Finally, it concludes which interactive Internet applications are most suitable to identify the QoE impact of delays or waiting times, respectively.

2.2.1 Characteristics of Interactive Applications⁵

The discussion in the preceding section revealed that QoE is a pluridimensional concept, which is also reflected in the discussion about influencing factors that is taken

⁴The definitions of the terms used as well as further details can be found in the QoE definition whitepaper [1] itself.

⁵Parts of this section are based on original work of the author with adaptations as published in [49] where he was acting as the lead author for the respective chapter and contributed text and figures.

up in the current Qualinet whitepaper [1]. Despite these insights the framework presented in the whitepaper still focuses mainly on a single person that experiences QoE of a certain single stimulus. Thereby it overlooks the important perspective that the interaction process between different entities (and related problems due to impairments) is a major factor of QoE for interactive services. Traditionally, this new and essential perspective has been addressed mainly in the context of human-to-human (H2H) – and, to a lesser extent, human-to-machine (H2M) – communication, while more recently also machine-to-machine (M2M) aspects have gained rapidly increasing relevance. Therefore, the discussion in this section will follow a rather broad approach, and discuss the corresponding fundamental concepts and notions related to the interaction process in an abstract way, including all different basic scenarios, and thereby address research question RQ1 (introduced in Section 1.2).

As far as underlying technology is concerned, it is mainly the intermediate communication channel – and more specifically its *two-way delay* characteristic – which is responsible for the need to distinguish interactive from non-interactive QoE⁶. Of course, this delay has a direct impact on the quality perception itself (as shown by [141] and further discussed in Section 2.2.2), but beyond that it may also massively influence the information sending or receiving behaviour of the individual communication partners involved.

The topic of interactivity is a widely used concept which is rooted in several different research traditions. As the aim of this chapter is the assessment of interactivity and the identification of the effects interactivity exerts on QoE, it is essential to differentiate between the different phenomena all identically labelled interactivity.

It is common to all understandings of interactivity that interaction can only take place if certain interactive acts are performed by at least two actors communicating with each other. Nevertheless, the nature of the interactants (humans, machine, media) as well as the way how they interact with each other are a crucial point of differentiation between the existing concepts of interactivity. A classification along the most prominent categories of interactivity has been proposed by [198], distinguishing between:

Interactivity as Process is interaction taking place between human subjects where subsequent messages consist of responses to prior messages or requests in a coherent fashion. Note that, in principle, the roles of the interactants are

⁶Also other distortions in the communication channel such as e.g. echo or noise do impact the interaction behaviour of interactants. However, throughout this thesis the focus will be on transmission delays.

reciprocal and can be exchanged freely.

Interactivity as Product occurs when a set of technological features allows users to interact with the system.

This classification already points towards the different scholar traditions of human interaction and human-to-system (computer) interaction. Human interaction researchers are rather strict in defining interactivity such as Rafaeli [177]. In their understanding, *true* interaction can only take place between human interactants when their roles (within the interaction) are 100% reciprocal. In contrast, scholars in human-to-machine interaction are less stringent and talk about interactivity as soon as interactive actions are exchanged between entities, even if the roles of the entities are not reciprocally interchangeable. In this chapter, however, the aim is to analyse the influence of interactivity on QoE for several types of different services as targeted in this thesis (including both H2H and H2M interaction)⁷. Hence, the following definition of interactivity from [49] is chosen as common ground:

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

Without loss of generality, the further discussion throughout this section and the thesis in general will be restricted to request-response patterns which are considered to be the common ground for both H2H and H2M interactions⁸.

Figure 2.2 depicts the fundamental structure of interactive communication, the request-response patterns. While the x-axis refers to time, one can see requests (REQ) and related responses (RES) exchanged between a user A and a receiver B via the intermediate transmission channel with constant one-way delay. Messages are assumed to exhibit an underlying fine granular structure (for more details on the dashed circle see Figure 2.3). Requests can be initiated by both sides, and responses typically follow them in time, however, in certain cases (cf. dashed circle) responses

⁷However, it is pointed out that the definitions of interactivity are also valid for other interactive services such as sensory experiences and interactive gaming.

⁸Human interaction scholars might argue that restricting interaction to request-response patterns is no longer analysis of *true* interaction but rather analysis of *quasi* interaction (cf. [177], [156] and [198]). However, as other services in addition to H2H interaction are targeted in this thesis it can be assumed that this restriction is adequate for identifying the influence of interactivity on QoE, for all of these services.

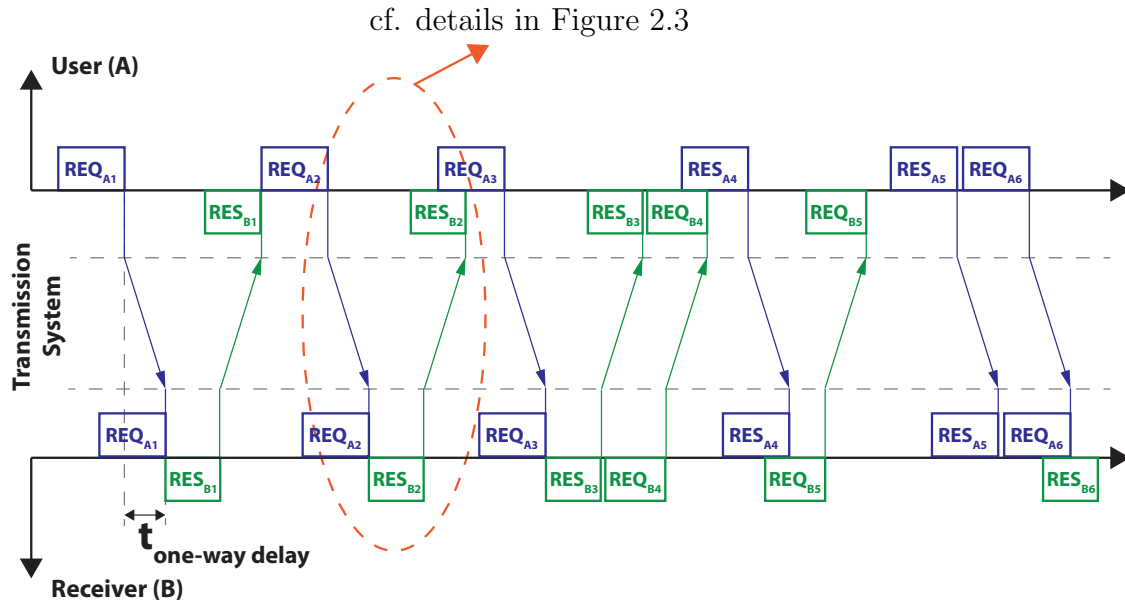


Figure 2.2: Illustration of a series of requests and responses throughout an interaction that constitute the request-response pattern

are started already before the end of the request transmission or are interrupted by additional arriving messages. Eventually, this can even lead to largely different perceptions with respect to the actual interaction pattern as pointed out in [100], leading for instance to the distinction between active and passive interruptions. A more detailed discussion of differing perceptions of interactional realities between interacting entities will follow in Section 3.1.2.2 and Section 3.1.4.

For a better understanding of the relation between request and related response and the influence of transmission delays Figure 2.3 depicts in detail what events can be observed and what timings are related to these events. User (A, top of Figure 2.3) issues a request which is transmitted to the receiving side (B, bottom of Figure 2.3). Now, the receiver (B) processes the request and starts responding by sending data to user (A) again⁹. In both directions, messages may exhibit a fine granular structure, which is shown in Figure 2.3 as a sequence of arrows where different thicknesses are used to indicate the “semantic intensity”¹⁰ of the corresponding content. Following the model outlined in [178], one can for instance assume that the most important pieces of information (e.g. key answer facts in human conversation or HTML format

⁹This is the description of an “ideal” conversation. Self-evidently, it can also happen that the receiver (B) responds with a request (counter question) as indicated on the bottom of Figure 2.2 right of the red dashed circle with two green boxes where RES_{B3} is followed by REQ_{B4} .

¹⁰In that context “semantic intensity” denotes the amount of semantic information contained in the response per bit.

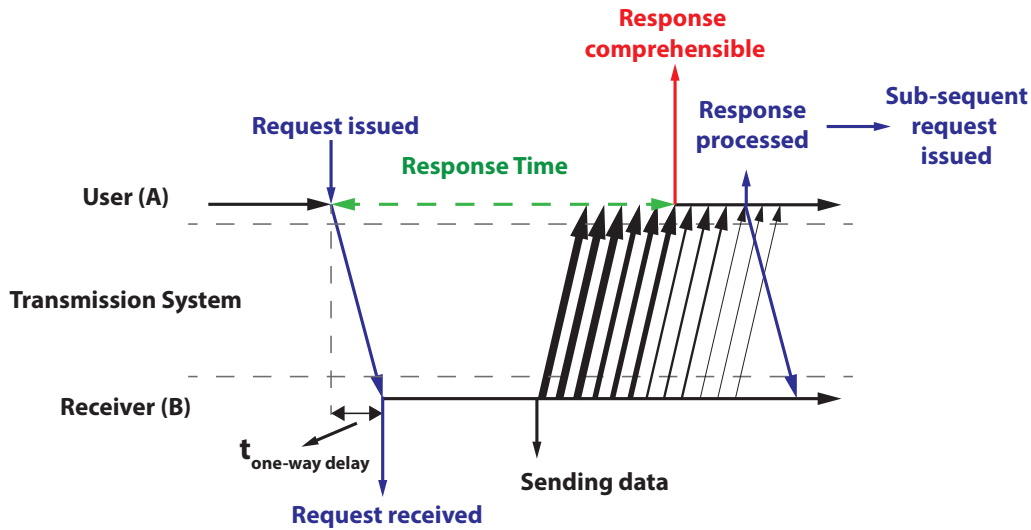


Figure 2.3: Illustration of an interactivity constituting request-response pattern based on [85], [29] and adapted with notions from [178]

instructions in Web traffic) are contained in the earlier parts of a response, while with ongoing message length, the corresponding information becomes less dense and/or less important ¹¹. As a consequence, the receiver might be tempted to start her next action already before the entire message has arrived. While such a behaviour is typically observed in everyday communication, from an overall system behaviour it can also very naturally be interpreted as a Nash Equilibrium that maximises the overall information exchange between both participants [178].

Hence, after a certain time, from the viewpoint of user (A) the transmitted response leads to an intermediate rendering result which is considered already sufficient by the user (this rendering state is defined as "response comprehensible"). Now user (A) starts processing the response. However, receiver (B) might keep on responding (e.g. in case of a long utterance or a heavy web page). After processing the response, user (A) issues a subsequent request, thereby starting a new *request-response cycle*. Here, it is important to understand that the issuing of the follow-up request by user (A) does not necessarily take place after the complete response has been received at user (A). He issues the request when he has acquired sufficient information from

¹¹Here it should be noted that this is a description for a prototypical "normal" conversation. Natural conversations might deviate to a certain extent, e.g. rhetoric tricks like keeping the payoff of a joke for the end of the utterance would not adhere to this decline in semantic intensity.

the (possibly technically incomplete) response and has accordingly processed it¹². The essential characteristic here is a certain relation between the response and a preceding event (in the simple case only the relation to a single request).

Together with the above definition of interactivity, it gets clear that *the request-response characteristic* is a *distinct feature of interactivity*, which thereby answers research question RQ1. Considering the differentiation between H2H interaction and H2M interaction, it can also be said that in terms of the receiving side (B), the nature (human, machine, etc.) of the entity answering the request is not essential for establishing an interactive request-response pattern, which makes it applicable to both interaction types and respective applications.

At present, the Internet applications enlisted in Table 2.1 are considered to share this interactivity related characteristic¹³. In the following section typical delay ranges in communication networks and their impact on interactive Internet applications are discussed.

2.2.2 Interactivity and Delay Impairments¹⁶

Delays introduced along the communication chain are a major issue for other above mentioned interactive Internet applications, as these delays increase the response time (cf. Figure 2.3) and thereby strongly deteriorate the interactive process of the entities which further leads to worse QoE as a result of these disturbances. The degree of deterioration introduced (by the delay) into the interaction process also depends on the interactivity inherent to the application considered. For instance, browsing through a simple online photo album has a very low degree of interactivity, and delay impacts the response time just once per picture view (which is not a

¹²This model is based on observations of interactions in H2H communication as reported in [2, 3, 52], where users were interrupting the other person frequently, and observations of H2M interaction where similarly, users while web browsing [8, 10] were navigating further on a web page through clicking on a respective link before the web page was fully loaded. This lower bound of sufficient information (for issuing a subsequent request) might be defined in two ways: 1) with a relative or absolute amount of information (e.g. 70% of rendered screen area, or fully rendered screen) 2) based on the considerations from [178] where the bound is reached after the entropy of user (A) gets smaller than the entropy of the response of user (B) in order to maximise the amount of information exchanged

¹³The enlisted applications are only a snapshot valid while writing this thesis and might be subject to future changes and other upcoming applications.

¹⁴This type of streaming is also referred to as progressive download.

¹⁵This type of streaming is also referred to as progressive download.

¹⁶Parts of this section are based on original work of the author with adaptations as published in [29] where he was acting as the lead author for the publication and contributed text and figures.

Category	Applications
Audio	Streaming (UDP, RTP) vs. Streaming (HTTP) ¹⁴ Telephone services (VoIP), multiparty telephone conferencing
Video and Audio-Visual applications	Streaming (UDP, RTP) vs. Streaming (HTTP) ¹⁵ Videotelephony (dyadic / multiparty) Video-conferencing (dyadic / multiparty)
Browser based applications	Web-Browsing File downloads
Cloud applications	Cloud Gaming Remote Folder Access Remote Desktop Access Online Office Applications (MS Office 365, Google Docs)
Online games	Several types of games that connect through the internet such as first person shooter, massively multiplayer online games etc.

Table 2.1: Overview of current interactive Internet applications

big impairment in this case), compared to cloud gaming where the response time adds up for every of the numerous request-response patterns and timing of game actions is critical. Therefore, despite the communality of these applications as being interactive, the sensitivity of the applications towards delays can strongly differ. A categorisation of different delay ranges and respective sensitive applications for these ranges is given by [116] and shown in Figure 2.4¹⁷.

From this categorisation one can see that time sensitive applications such as real time human-to-human communication over IP (VoIP, video telephony, online document sharing combined with audio- or videoconferencing) are affected by delays below one second. Therefore, they demand communication networks that ensure transmission delays in a range below one second in order to guarantee a disturbance free interaction process. In contrast, applications such as web browsing or file downloads are impaired by delays larger than two seconds (cf. [62, 167, 182]).

These different time ranges of delay impairments are not only related to different applications but they are also related to different human timing systems used for the processing of the request-response pattern within the (human) user as described in [69]. These are, millisecond timing for delays below one second and interval timing

¹⁷The differentiation between error tolerant and error intolerant applications is not further discussed in this thesis as TCP/IP and its property of reliable packet delivery does not introduce packet losses but these losses are translated into delays.

Error tolerant	Conversational voice and video	Voice/video messaging	Streaming audio and video	Fax
Error intolerant	Command/control (e.g. Telnet, interactive games)	Transactions (e.g. E-commerce, WWW browsing, Email access)	Messaging, Downloads (e.g. FTP, still image)	Background (e.g. Usenet)
	Interactive (delay $\ll 1$ s)	Responsive (delay ~ 2 s)	Timely (delay ~ 10 s)	Non-critical (delay $\gg 10$ s)

Figure 2.4: Recommended delay categories for different applications from [116]

for delays above one second up to hours as shown in Figure 2.5. The circadian timing system relates to long-term QoE or service QoE as discussed in [99, 159] and is therefore not considered in this thesis. The existence and applicability of these different timing systems is of particular interest as the processing of delays and their relation to human (interaction) behaviour as well as the sensation of delays (or waiting times) differ for the different timing systems as numerous psychological studies have shown [55, 56, 83].

The main difference is that the millisecond timing system processes timings on a sensory and automated (= unconscious) basis, whereas the interval timing system is based on conscious processing of time [69]. Therefore, in case of the millisecond timing, transmission delay is not directly sensed by human interactants (as it has been shown in qualitative results for voice communication reported in [2, 3], where interactants were complaining about the inattentiveness of the other interlocutor but they were not able to name a communication system deficiency). Despite the lack of conscious awareness of the delay the human sensory system reacts on it through automated processing (i.e. unconsciously). In case of human interactants involved in human-to-human communication, the interaction behaviour can be unconsciously altered which creates certain communicative problems as unintended interruptions or double talk. These unconscious behavioural changes and delay induced problems influence the subjective quality impression of the interactants in turn.

On the other hand, for delays above one second the human body utilises the

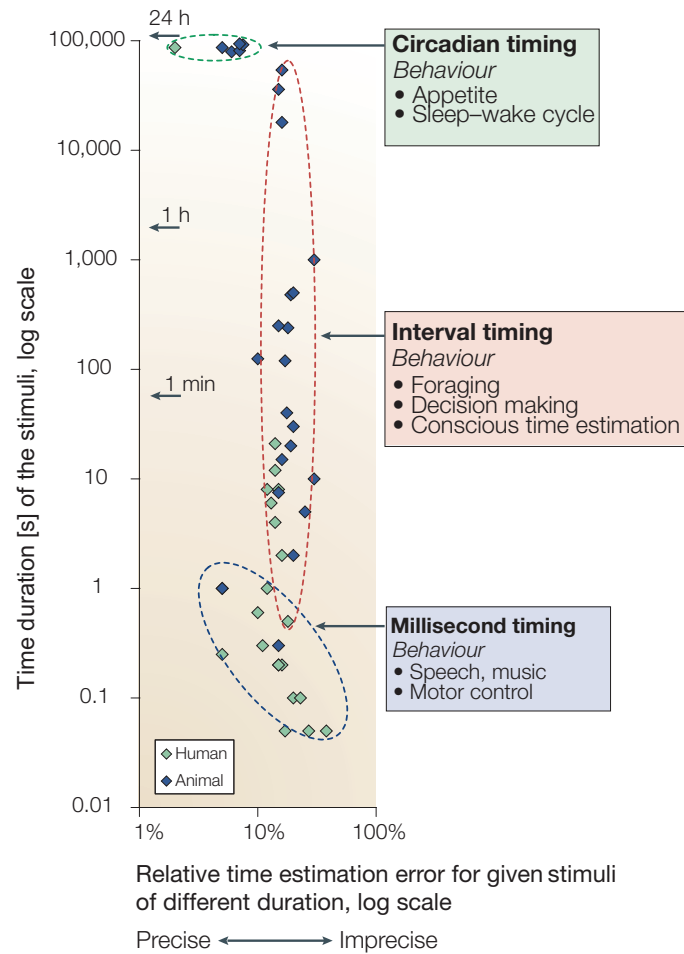


Figure 2.5: Overview of the human internal timing systems (that are involved in the processing of transmission delays) and time ranges in which the timing system applies as well as error ranges that apply for each of these timing systems, taken from [69] and adapted towards behavioural description for each timing system.

interval timing system to process the wait or slowdown of interactions. In case of human-to-human mediated interaction, delays above one second lead to severe changes in the interaction behaviour (as will be shown in Chapter 3) which plays a dominant role in the user sensation of the problem with the communication system, therefore the (conscious) processing of delays or waiting times above one second can be left aside for such services. In contrast, for applications that naturally involve longer delays in their interaction process such as file downloads or web browsing, the time elapsed after a request is sent can easily exceed one second, without compromising QoE and the interactive process respectively (cf. [62, 182]). In such cases the delay impairment will exert influence on the interaction process, which can still be

processed unconsciously, but the human interactant will also consciously process the delay or waiting time respectively, and will incorporate this conscious experience in his QoE judgement. For the conscious processing of waiting times psychological results on the relation between waiting time and user satisfaction, as reported in [200] or [93] will be of interest and will be further reviewed in Chapter 4.

In terms of perceived interactivity as a result of delays or waiting times the classification by [157] is interesting in the light of the different timing systems and their related time ranges as discussed above. The classification shown below (cited by [167] from [157]) enlists three time ranges of response times (i.e. delays) and how interactivity is perceived for these three time ranges in the context of human computer interaction:

0.1 s is about the limit for having the user feel that the system is reacting instantaneously, meaning that no special feedback is necessary except displaying the result.

1.0 s is about the limit for the user's flow of thought to stay uninterrupted, even though the user will notice the delay. Normally, no special feedback is necessary during delays of more than 0.1 s but less than 1.0 s, but the user does lose the feeling of operating directly on the data.

10 s is about the limit for keeping the user's attention focused on the dialogue. For longer delays, users will want to perform other tasks while waiting for the computer to finish, so they should be given feedback indicating when the computer expects to be done. Feedback during the delay is especially important if the response time is likely to be highly variable, since users will then not know what to expect.

These ranges and the related perceptions of interactivity are inline with the insights regarding conscious and unconscious processing of delays and waiting times respectively in the context of human timing systems, and confirm the relevance of separately analysing delays below one second and above one second.

Nevertheless, QoE is not automatically linked with these times as there are also other influencing factors to be considered, such as service or application, expectations etc., as discussed in Section 2.1. As a further consequence, QoE evaluation methodologies have to be properly designed to capture the delay influence for the application under study and have to consider the timing system involved (and thereby also consider conscious or unconscious processing of delay impairments). In order

to keep the problem space manageable in this thesis, a selection of services to be targeted has to be done. Ideally, these applications have to deal with delays related to the above mentioned timing systems of millisecond timing and interval timing. Therefore, the following two applications will serve as prototypical applications throughout this thesis:

Internet Telephony as example application for unconscious alteration of the interaction process involved in human-to-human mediated voice communication over IP which is strongly affected by below one second delays, thereby falling under the processing of the millisecond timing system

Web Browsing serving as a representative of applications adhering to the WIMP¹⁸ interaction paradigm that typically deals with delays or waiting times above one second, and therefore will be subject to processing by the interval timing system.

For both of these applications waiting times and delay impairments respectively, and their influence on QoE perception will be analysed. Furthermore, inclusion of interaction process related measures into current QoE frameworks and QoE prediction models will be reviewed. In this respect, the following section will review existing QoE assessment methodologies regarding their ability to establish interactive processes for the selected applications, and their capabilities to acquire measures of the interaction process which can be used for identifying interactional problems.

2.3 QoE Assessment Methodologies¹⁹

The central question for QoE research and engineering is how to operationalise the concept in terms of performing reliable, valid, and objective measurements. This challenge is framed by the overarching questions '*How can we quantify QoE and how can we measure it?*'. Since inclusion of the end-user's perspective is the defining aspect of QoE, conducting measurements merely on a technical level (e.g. by just assessing conventional end-to-end QoS integrity parameters) is not sufficient. Thus, QoE assessment methodologies are needed that act as translator between a

¹⁸The acronym WIMP denotes "windows, icons, menus, pointer", a style of interaction using these elements of the user interface.

¹⁹Parts of this section are based on original work from the author with adaptations as published in [21], where he was responsible for Section 3.2 and actively contributed text and figures to Section 3.1 and Section 3.4.

set of technical (QoS) and non-technical (subjective and contextual) key influence factors, interaction performance indicators, user perception, and ultimately, user experience. QoE assessment methodologies can be categorised into subjective and objective quality assessment methods.

As the focus of this thesis is on subjective QoE evaluation methodologies, identification of additional influence factors, and their relation to QoE, objective assessment methodologies will not be discussed. Regarding further information on objective quality assessment methods, the interested reader is pointed to a comprehensive overview of such methods in [21], [160, 163, 165, 174].

Subjective quality assessment methodologies are based on gathering information from human assessors (frequently referred to as 'test participants' or 'test subjects') who are exposed to different test conditions or stimuli during the process. In general, a panel of assessors is presented with various system parameterisations or media qualities (e.g. different downlink bandwidths, different transmission delays, or audio clips encoded using different settings) which lead to some form of explicit, or implicit response. In most cases, quantitative methods derived from neighbouring disciplines, such as psychophysics and psychometrics are used to obtain information regarding assessors' judgment in the form of ratings that describe their perception of the respective quality experienced (i.e. QoE, cf. [163]). In addition, qualitative methods such as focus groups, interviews, or open profiling [197] are used, particularly in order to find out which influence factors or features contribute to QoE and how [145].

Subjective tests are typically conducted in a controlled laboratory²⁰ setting and require careful planning in terms of which variables and influence factors need to be controlled, measured, and monitored. To this end, recommendations like ITU-T Rec. P.800 [114] and ITU-T Rec. P.805 [120] provide detailed guidelines regarding choice of test conditions, rating scales, room setup, as well as sequencing and timing of the presentation. The typical result of a subjective test campaign are the individual assessors' ratings which are aggregated into so-called mean opinion scores (MOS). The MOS expresses the average quality judgment of a panel regarding a certain test conditions, the related overall quality experienced or the specific quality along a certain quality dimension (e.g. picture quality) [119]. It is based on an ordinal five-point scale: (1) bad; (2) poor; (3) fair; (4) good; (5) excellent. Note that most test designs rely on absolute scales (like the aforementioned five-point scale), but

²⁰In addition to the lab, field trial methods for conducting studies under real-world conditions [166], [6] as well as cost-effective crowdsourcing methods [72, 108] have become popular in subjective QoE assessment.

also relative Differential MOS (DMOS) or continuous methods are being used (see ITU-R BT.500 for details). An overview of existing standardised methodologies within ITU is given in Figure 2.6.

Application	Media	Conversational (CONV)/Non-conversational (NONCONV)	Subjective test methodology
Telephony	Speech	NONCONV	[ITU-T P.800] [ITU-T P.830] [ITU-T P.835]
		CONV	[ITU-T P.800] [ITU-T P.805]
Video telephony	Multimedia (Note)	CONV	[ITU-T P.920]
Video streaming (Mobile TV/IPTV)	Video	NONCONV	[ITU-T P.910] [ITU-T J.140] [ITU-R BT.500-12]
	Audio	NONCONV	[ITU-T P.830] [ITU-R BS.1116-1] [ITU-R BS.1285] [ITU-R BS.1534-1]
	Multimedia	NONCONV	[ITU-T P.911]
Web browsing	Data		none

Figure 2.6: Existing subjective QoE assessment methodologies standardised by ITU-T, taken from [124].

For the majority of assessment methodologies the MOS has become the de facto standard metric for QoE, a development that led to considerable debate (cf. [61, 110, 134, 144]) on how QoE should be measured over the last decade. Beyond the controversial use of ordinal grades for computing averaged scores, this debate is also nurtured by the fact that assessors' judgments are influenced by various user- and context-related parameters due to the pluridimensional nature of QoE. Therefore, a number of authors have proposed to complement subjective QoE ratings with alternative measures free from distortion by user opinion [68, 81, 101, 134, 138, 144,

154, 197, 213]. Such *objective* QoE measures [68, 144] can be task performance²¹, physiological indicators²², or user behaviour in general²³. In the light of interactive Internet applications *user behaviour measures* are of particular interest as they also take the interaction process itself into account which is deteriorated by transmission delay as already discussed in Section 2.2.1. Hence, certain *user behaviour measures* can be used to derive or compute interaction performance measures and thereby identify disturbed interaction behaviour due to temporal impairments.

Summarising, subjective QoE assessment methodologies for interactive Internet applications face certain challenges in terms of QoS requirements, interaction behaviour and related content. Within those the most crucial ones are

- fertilising interaction over certain time spans, and
- tracking of interaction cues.

Considering these requirements, standardised assessment methodologies as well as related work is discussed in the following two sections for the two prototypical services VoIP and browser based applications.

2.3.1 Subjective Speech QoE Assessment Methodologies

For speech QoE assessment two different approaches for QoE assessment can be differentiated: listening quality tests and conversational quality tests, whereas the latter ones can be further sub-divided into methodologies where the (quality) evaluating subject is not involved in the conversation and just follows a recorded conversation passively by mainly listening and answering some contentual questions regarding the recorded conversation, and methodologies where the quality evaluating subject is actively involved (speaking with another interlocutor) in the conversation.

The most widely used assessment approaches are listening (or: listen-only) tests as described in [114]. Typically, these tests utilise a number of short speech samples (approx. 5 s duration), that have been previously recorded and contain a certain set of audible impairments. The test subjects are then asked to listen to the samples and issue a QoE rating for each sample. Such tests allow for the evaluation of a large number of different test conditions within short durations. However, this gain in execution speed results in a considerable loss of external validity and QoE

²¹E.g. quality and speed of goal completion [144].

²²E.g. heart rate, skin conductance [154, 213]. These are often used to assess emotional states.

²³E.g. speaker alternation rates, cancellation rates, viewing times [81, 101, 138].

dimensions evaluated. Therefore, the acquired subjective scores mainly represent the signal (or system) fidelity only. This is based on the fact that listen-only tests by their nature do inherently neglect the context of interactive conversations and its dynamics. Thereby, they fail to include the interaction performance aspects which play an essential role in human-to-human interaction as shown in [101] and reviewed in Section 2.2.1.

In order to overcome these shortcomings, conversational test methodologies have been developed. The aim of conversational speech QoE assessment methodologies is to include the interactional context and interactional impairments caused by transmission delay into the gathered QoE ratings. As already mentioned above, they can be sub-divided in active and passive conversation tests. For passive (listening) conversation tests two different approaches do exist: The approach described in [115] uses recorded samples of, e.g. echo impaired conversations, which include signals from both end points. A slightly different approach is proposed by [211] where simulated conversations are created from short term samples (5 s to 6 s) in a way that a meaningful dialogue of 1 min to 2 min is rendered. For both approaches the resulting samples then contain a conversational structure and can include double talk as well as speaker interruptions and following speaker changes. The subjects then listen to the recorded conversations (comparable to non-interacting bystanders in normal conversations) and issue their respective QoE ratings. The methodology proposed in [211] also includes questions to the listening subjects. The questions are related to the content of the preceding sample(s) and have to be answered verbally by the listeners. This enhances the attention of the subjects as well as their sense of involvement in the recorded conversations. Passive conversation tests address certain limitations of listening tests and partially cover interactional deficiencies in the conversational structure as well, such as double talk, interruptions and speaker changes. However, they fall short in incorporating the effects of transmission delay on the conversational structure and are not able to consider human adaptation strategies used to compensate delay induced conversational problems.

Active conversation tests as described in [113, 120, 162, 176, 183] (cf. Figure 2.6) overcome these limitations. The main aim of these kind of tests is to establish a (real) conversation between two or more subjects²⁴. For initiating such conversations several different scenarios have been defined in [120]. From these, the most prominent ones are: 1) short conversation tests (SCT's), which aim to mimic everyday life

²⁴As the focus of this thesis is on dyadic interaction over VoIP the remainder of this section concentrates on scenarios for two interlocutors

situations such as travel arrangements, or ordering food from a delivery service over the phone. In their practical implementation each of the interlocutors is assigned a certain role in the dialogue (e.g. (A): hotel reception desk, (B): tourist searching for a free room) and is then asked to interact with the other person upon this role, and 2) random number verification tests (RNV's) where both subjects receive first a table with numbers, and are then asked to verify which of the numbers in the table on their side (A) do match with the numbers in the table of the interlocutor on the other side (B). For both scenarios the signals of the interlocutors are then transmitted over the system under test and impairments to be tested have to be inserted in real time. After finishing a scenario both subjects are then asked for their QoE ratings on certain scales. One criticism for these kind of conversation tests is that the subjects direct too much attention to the conversations itself [211], hence less attention is left for properly assessing quality features of the conversation. However, they feature a high degree of realism by considering transmission delays, accounting for different usage situations through different tasks involved, and capturing interlocutors' conversational behaviour adaptations. Thereby they capture communication dynamics and are hence the only methodologies that allow to properly address the impact of transmission delays in the interactional context.

Despite all these theoretical advantages, current methodologies for active conversational tests treat some of these variables as pure experimental variables and only measure user perceived quality on a certain (MOS) scale. Thereby they fall short in considering interaction performance aspects (cf. Section 2.2.1) which could be obtained by additionally measuring conversational surface parameters. Such *conversational surface parameters* can be used to *quantify communication problems* induced by transmission delay and can give valuable information about the interactional state of the conversation, usage situations (e.g. degree of interactivity), and user characteristics as shown in [101] [2,3]. Combining such measures with perceptual quality scores would constitute a big step towards *covering the dimension of interaction performance* addressed in Section 2.2 to a larger extent. In that respect, Chapter 3 will give a comprehensive overview on related work regarding conversational surface parameters and new approaches taken by the author to overcome the neglect of the interactivity dimension.

2.3.2 Subjective Web QoE Assessment Methodologies

In general, the term Web QoE stands for the Quality of Experience of interactive services that are accessed via the browser and based on the HTTP protocol [35]. In contrast to audio and video quality assessment methodologies, where several accepted and even standardised testing methodologies exist (cf. Figure 2.6), there is far less guidance in terms of proper testing methodologies for Web QoE. Concerning browser based applications, it has been widely recognised that in contrast to the domains of audio and video quality, where psycho-acoustic and psycho-visual phenomena are dominant, *end-user waiting time is the key determinant of QoE*²⁵ [60,167,202]. The longer users have to wait for the web page to arrive (or transactions to complete), the more dissatisfied they tend to become with the service.

Another main difference towards existing audio and video QoE assessment methodologies focusing on static media experiences is the interactive nature of the task and related user behaviour. Typically, the user does not issue a single request which is then answered by a short single media experience while web browsing, but rather goes through a series of such request and responses. Figure 2.7a depicts two request-response patterns involved in web browsing where $T_1 + T_2$ or $T_3 + T_4$ respectively, characterise the waiting time for one page view. A web session, however, consists of several of such waiting times which are typically of different length (cf. Figure 2.7b). In that respect several web studies show that these waiting times are embedded in an interactive *flow* of page views (cf. [193,210]). Even new pages with plentiful information and many links tend to be regularly viewed only for a brief period. Thus, users do not perceive web browsing as a sequence of single isolated page retrieval events but rather as a *flow experience* (cf. [193]). Understanding these differences in usage behaviour is essential for deriving realistic assessment methodologies for these applications. The notion of flow implies that the quality of the web browsing experience is determined by the timings of multiple page-view events that occur in a certain time frame during which the user interacts with a website.

Therefore, a testing methodology for web browsing QoE must ensure that such request-response patterns are issued throughout an evaluation. In order to achieve this goal, two different approaches can be distinguished: 1) a defined number of requests or 2) a defined duration of one web session. Approach 1) as used in [86,126], [35] demands two requests and subsequent responses and page views as depicted in Figure 2.7b (therefore addressing just a subset of page views of a

²⁵In Chapter 4 this hypothesis will be tested.

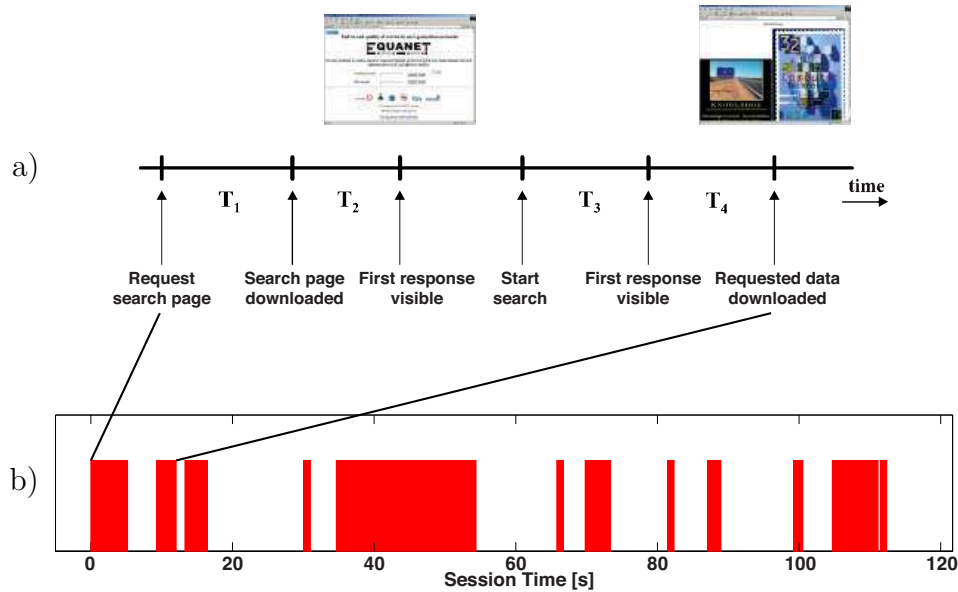


Figure 2.7: a) Waiting times related to request-response patterns in web browsing [126] and b) a web session as series of page views with different waiting times.

whole web session as indicated by the zoom beam in Figure 2.7). After completion, the user is then prompted for his quality rating on an absolute category rating (ACR) scale. The independent variables here are the $T_1 + T_2$ and $T_3 + T_4$ waiting times. Contrary, approach 2), which was utilised by the author in [5–8] uses pre-defined session times. For each session the user is asked to execute a certain task on the given webpage while network parameters (e.g. downlink bandwidth, round trip time) are varied as independent variable. After the session time is elapsed, the quality rating is gathered. Whereas approach 1) considers the overall session time as independent variable against which the MOS are plotted, approach 2) uses network level parameters as independent variable which then influences the waiting times for each request-response pair, respectively.

While the former approach allows exactly controlling waiting times, the latter approach guarantees a more realistic *flow based* web browsing experience for the user, resulting in a series of waiting times (cf. Figure 2.7b) the user is exposed to. In Chapter 4 it will be shown how these different approaches can be jointly used to explore the influence of waiting times on Web QoE, and to understand what additional perceptual phenomena can thereby be identified.

2.4 Challenges and Requirements for QoE Assessment Methodologies for Interactive Internet Applications

The discussion in the preceding sections have shown that it is important to include interaction performance aspects into QoE modelling and analysis. Accordingly, QoE assessment methodologies have to consider these aspects. Therefore, they must 1) use interactive scenarios for QoE assessment and 2) they have to provide additional measures of interaction performance that can be used for analysing the influence of impaired interaction on QoE in addition to currently considered influence factors.

Reviewing the concept of QoE in Section 2.1 has clearly underlined the pluridimensional nature of QoE which is widely acknowledged. However, it also became clear that the interaction process of interactive applications is not yet adequately recognised nor instrumentalised in related QoE frameworks. Section 2.2 identified the request-response pattern as a commonality of the interaction process across interactive applications that is deteriorated by delay impairments. Furthermore, it was shown that two different delay ranges can be distinguished in terms of human perception of time related stimuli or impairments respectively, hence these different delay ranges exert different influences on the interaction process. Two example applications have been selected to represent these different delay ranges, namely *Internet Telephony* and *Browser based Applications*. Finally, the review of existing QoE assessment methodologies in related work and standardisation comes to the conclusion that existing methodologies do not sufficiently consider interaction related aspects. Whereas in the case of Internet telephony, testing methodology facilitating interactivity exists but not satisfactory considers characteristics of the conversation process itself, in the case of browser based application current methodologies fall short in establishing an interactive flow experience that facilitates the interactive process for, e.g. web browsing.

To sum up, these results lead to the following requirements regarding subjective QoE assessment methodologies for the identification of the delay impact on QoE for interactive Internet applications:

Interactivity is crucial for the targeted example applications. Therefore, assessment approaches have to ensure that the resulting task execution throughout an evaluation session adheres to the request-response pattern and establishes a certain level of interactivity.

Interaction Performance Metrics have to be captured and properly analysed to understand changes and interaction defects caused by delay in case of Internet Telephony.

Flow Experience has to be established for a high degree of realism in Web QoE assessment.

Waiting Time can either be a singular (response time) experience of an application or appear in a series of page views as multiple response times. How waiting times in either of these two approaches relate further to the overall QoE has to be analysed for identifying the relation between waiting time and QoE.

Within the context of this thesis, these requirements will be met as follows: in Chapter 3 subjective tests will be conducted based on scenarios that facilitate certain conversation interactivity levels, and the resulting interaction process will be thoroughly analysed by means of conversational surface structure analysis. Then the gathered measures will be used to derive QoE prediction models that incorporate interaction related metrics. For web browsing as second target application covered in Chapter 4, first a subjective testing methodology that ensures flow experience and related interactivity levels will be derived and verified. In the second step, this methodology will be used to identify the relation between waiting time and QoE. Finally, the gathered data is utilised to derive a model of QoE for web browsing as a function of the waiting time.

Chapter 3

Internet Telephony

In the telecommunication industry the impact of transmission delays on customer satisfaction has been a constantly recurring topic of interest. However, the interest towards the impact of delay from the research community was modest compared to research on the impact of degradations such as packet-loss, jitter, noise, and codec distortions on user-perceived voice quality. In Figure 3.1 delay related scientific publications as reactions to technological changes in the 1960's, early 1990's and around 2000 are depicted over time. It can be seen that the introduction of different transmission technologies as well as the extension of distances telephone calls were transmitted over, which led to severe changes in the transmission delays, has led to scientific work addressing related conversational quality problems¹. However, a large share of these studies was devoted towards the impact of transmission delay on user satisfaction or conversational speech quality. Although approaches have been made to statistically describe changes in communication behaviour and related parameters due to transmission delays, they fall short in understanding the root causes of these changes or they did not consider important parameters such as interruptions in their analysis. As a result there are only a few QoE prediction models published that do consider delay and conversational parameters.

The aim of this chapter is to address the lack of analysis for conversational quality and related models under delay influence by providing, 1) a reliable evaluation on how transmission delay affects Quality of Experience (QoE) of voice communications in different controlled conversational contexts, 2) analyse the impact of transmission delay on conversational surface structure, 3) introduce two new conversational

¹The initial introduction of telephone systems and a first increase in transmission delays in the 1920's and 1930's is not included in this figure as respective publications are not known to the author

metrics that capture the delay influence for feedback cues e.g. interruptions in order to address research question RQ2 (introduced in Section 1.2), and 4) to present an updated version of the E-model that considers interactivity metrics for QoE prediction (thereby addressing research question RQ3, introduced in Section 1.2), and 5) to define new thresholds for acceptable transmission delays based on aforementioned conversation analytical results for scenarios of different interactivity.

This chapter is organised as follows. Section 3.1 reviews related work in the field of speech quality and conversational parameter extraction and introduces two new metrics that capture the delay influence in human mediated conversations. Section 3.2 describes the experimental setup used to quantify the delay impact on QoE for certain conversational scenarios. Conversational quality results of our study as well as the analysis of the conversational surface structure and the newly introduced metrics are then presented in Section 3.2.3. Finally, Section 3.3 proposes an approach to use two of the newly introduced interactivity metrics for QoE prediction in a modified E-model and Section 3.4 concludes the achievements of the chapter and presents delay guidelines for conversational voice services.

3.1 Background: Perceived Quality, Conversational Analytics and Communication Theory

The influence of delays on user behaviour and perceived quality in telecommunication systems has been mainly studied from two different viewpoints: 1) The conversational quality viewpoint which attempts to analyse the relation between conversational quality and underlying transmission delays and 2) A conversation analytic approach that tried to understand how human interaction behaviour changes due to transmission delay. Although some of this work includes both viewpoints to a certain extent, none really deepens the interrelation between both viewpoints. A rough classification of the related work is shown in Figure 3.1. The allocation of related work towards the categories was done according to the category of results the specific related work contributed. Therefore, related work that is categorised as quality centric might also discuss conversation analytic aspects and is therefore listed in both categories. The publication dates show that three waves of related work are present. The first wave around the 1960's was triggered by the availability of geostationary satellite communication links and the concomitant long transmission delays. The second wave in the early 1990's was a reaction to the introduction

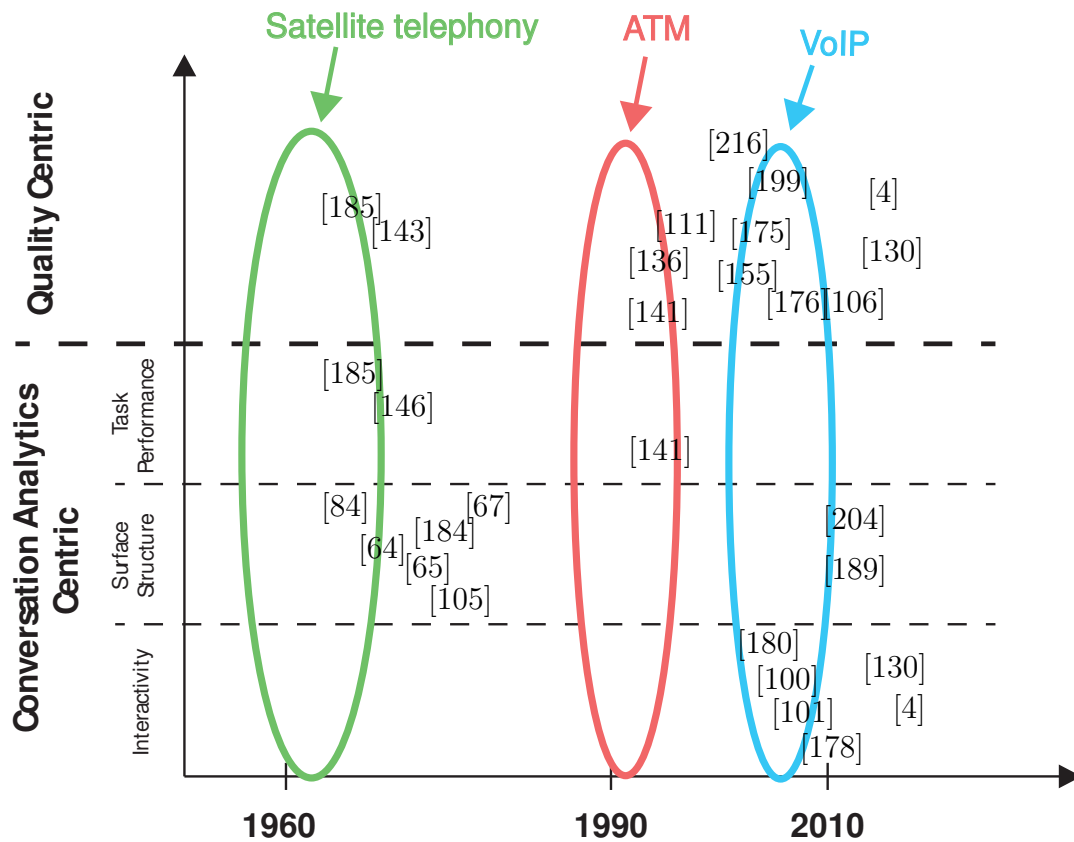


Figure 3.1: Categorisation of related work on the influence of delay in telecommunication systems along analysis methods and year of publication

of digital signal processing equipment, ATM technology and the uprise of mobile communication networks, all of them being suspected to introduce additional delays in the transmission chain. The last wave of related work started around the mid 2000's has been induced by transmission delay related problems arising from VoIP systems caused by imminent characteristics of packed-switched networks such as varying packet delivery times dependent on the packet routing.

The overarching categorisation is performed along the two dimensions of *quality centrality*, where the main focus is on the relation between transmission delay and perceived conversational quality further discussed in Section 3.1.1, and *conversation analytics centrality*, where the focus of the mentioned publications is the change of interaction behaviour due to the transmission delay further elaborated in Section 3.1.2. The latter dimension is further subdivided into different approaches of

conversation analytics². Contributions from related work towards understanding the influence of transmission delay on conversational quality and conversation behaviour are discussed in the following two sections.

3.1.1 Impact of Delay on Perceived Quality

Initial work on the relation between transmission delay in presence of echo and user acceptance has been conducted by [143,185]. The authors assessed the acceptance of long transmission delays in test circuits over a number of weeks through counting the number of rejected calls for each delay condition. In their first study [185] they analysed the relation between delay and call rejections (the rejection rate is related to the more widely used acceptance rate as: $R_{Acceptance} = 1 - R_{Rejection}$) in case of *echo afflicted* transmission channels. In this study the rejection rate of delay impaired calls was considerably high. Even for 100 ms one-way delay approximately 12% of the calls were rejected ranging up to a rejection rate of about 34% for the 600 ms condition. The latter study [143] analysed the same relation between delay and acceptance with the difference that the test circuits did not induce talker echo³. The results showed a considerably decreased rejection rate. In numbers, the rejection rate for the 300 ms one-way delay condition dropped down to 0% and the rejection rate for 600 ms one-way delay were well below 5%. Inline with these findings are the results from [105] which didn't show a severe change of quality ratings (on a 5-point scale) for one-way delays up to 600 ms in echo-free conditions.

Although these studies represent first steps towards assessing the influence of transmission delays and give insights in the relation to conversational quality or closely related measures, they can not directly be compared to the related work discussed in the remainder of this section for the following reasons: the results reported in [185] and [143] do not measure conversational quality nor was a task assigned to the participants, hence the established interactivity between different participant pairs might have severely fluctuated and thereby influenced the delay sensitivity of different conversations. In contrast the study in [105] uses already certain conversational scenarios ranging from free conversation to jointly solving a puzzle to facilitate constant interactivity levels, however it falls short in analysing

²The notion *conversation analytics* is used throughout this thesis as I want to clearly distinguish the quantitative approaches towards the analysis of conversations (as used in this thesis) from qualitative *conversation analysis* as introduced by [187] and used in social sciences.

³This was achieved by using 4-wire circuits in contrast to normal 2-wire circuits as used in the first study. (cf. [143])

the results with the scenarios (and hence differing conversational interactivity) as influencing variable and summarises ratings of all different scenarios in one overall conversational quality measure. Due to this methodological differences these results are not included in the comparison of related work presented in the following and Figure 3.2 and Figure 3.3.

Another point raised by the comparison of results between [185] and [143] is the negative influence of echo. In case of echo in conjunction with transmission delays the call rejection rate was substantially higher compared to calls without echo in the transmission system. Similar results have been shown by [98] where the authors compared conversational quality of delay impaired conversations with and without echo and showed that conversations with echo were rated significantly worse than conversation where no echo was present. These results show that echo influences conversational quality strongly as described in [118], [123] and overlays other effects of transmission delay on the conversational process. Due to this strong negative influence of echo in the presence of transmission delay, *only echo-free communications circuits and related work will be considered* in the remainder of this thesis.

In terms of recent speech quality measures such as mean opinion scores (MOS) as defined in [120], seminal work on the effect of delay on perceived conversational quality of telephone systems was performed by [141]. The authors assessed the perceived quality of different conversational scenarios for different delays and delay detectability thresholds, respectively. Their results show that delay detectability thresholds of untrained participants were up to 1120 ms (one-way delays) depending on the conversational scenario. In contrast, their obtained perceived conversational quality results were surprisingly much more delay sensitive across highly interactive scenarios and dropping by 0.4 MOS points up to 1.0 MOS for moderate transmission delays of 125 ms and 250 ms already (cf. Kitawaki (random number verification, RNV) and Kitawaki (read numbers) in [141] and Figure 3.3). This means that the participants were not able to detect certain delays but did rate these conversations critical in their perceived quality scores. Such a result might be explained by the fact that the participants were not able to name the degradation (delay) but nevertheless were (unconsciously) aware of conversational problems when issuing their ratings. Contrary, a similar study by [136] presented results where transmission delays of up to 600 ms were rated only slightly worse in terms of conversational quality by 0.1 MOS. One could argue that these differences can be based on differences in the conversational interactivity inherent to the scenarios used for the tests. However, this is not the case for these two studies as both utilised comparison and verifica-

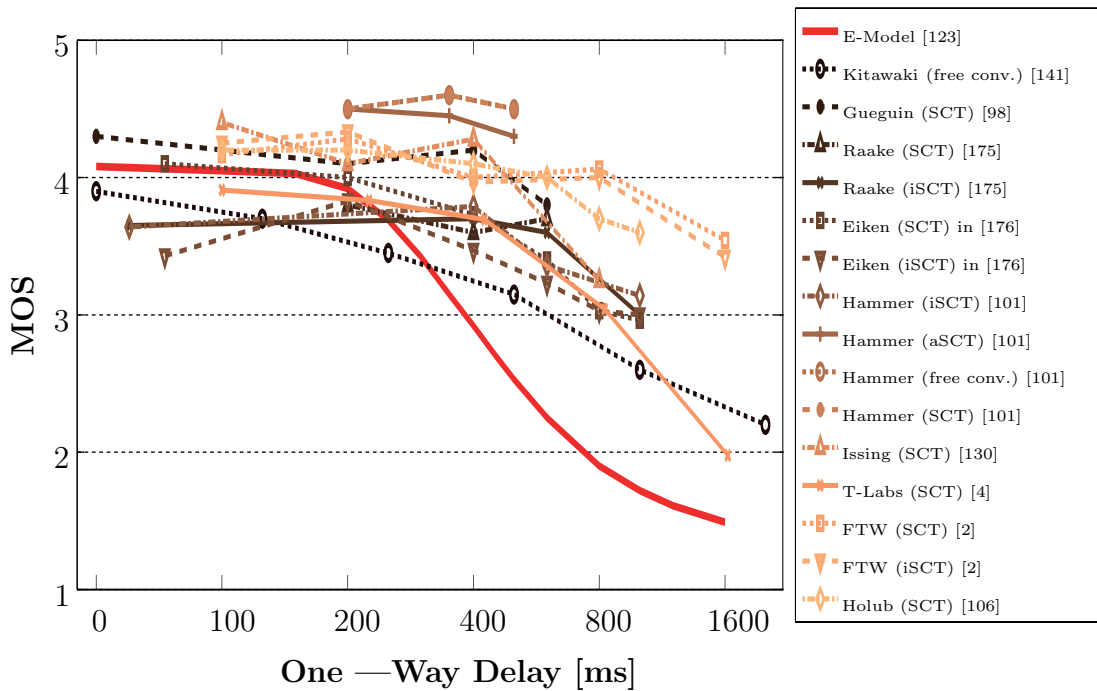


Figure 3.2: QoE (MOS) vs. transmission delay from related work for SCT and free conversation scenarios of comparable conversational interactivity

tion tasks respectively, which are assumed to establish rather high conversational interactivity levels in the resulting test conversations.

In order to consider scenario related differences in results from related work, Figure 3.2 and Figure 3.3 depict only results of scenarios with comparable conversational interactivity levels. In addition, both figures show conversational quality estimates from the E-model [123] which is the most widespread model that takes transmission delay into account for its predictions. Before comparing the results within scenarios it is important to mention that the technical setup of all these studies differed to a certain extent. Therefore, the absolute values of the results can differ due to different configurations such as used codec, loudness ratings etc. which induce an offset in terms of MOS. Nevertheless, the gradient of the resulting rating curves and the related drop in conversational quality between low and high delay values should be comparable.

The results in Figure 3.2 show conversational quality ratings obtained with scenarios of mid to low conversational interactivity such as short conversation tests (SCT) or free conversations (FC). The results showing the strongest negative influence of transmission delay are the ones reported in [141] and [4]. Both of them show a difference between the conversational quality for the lowest and highest de-

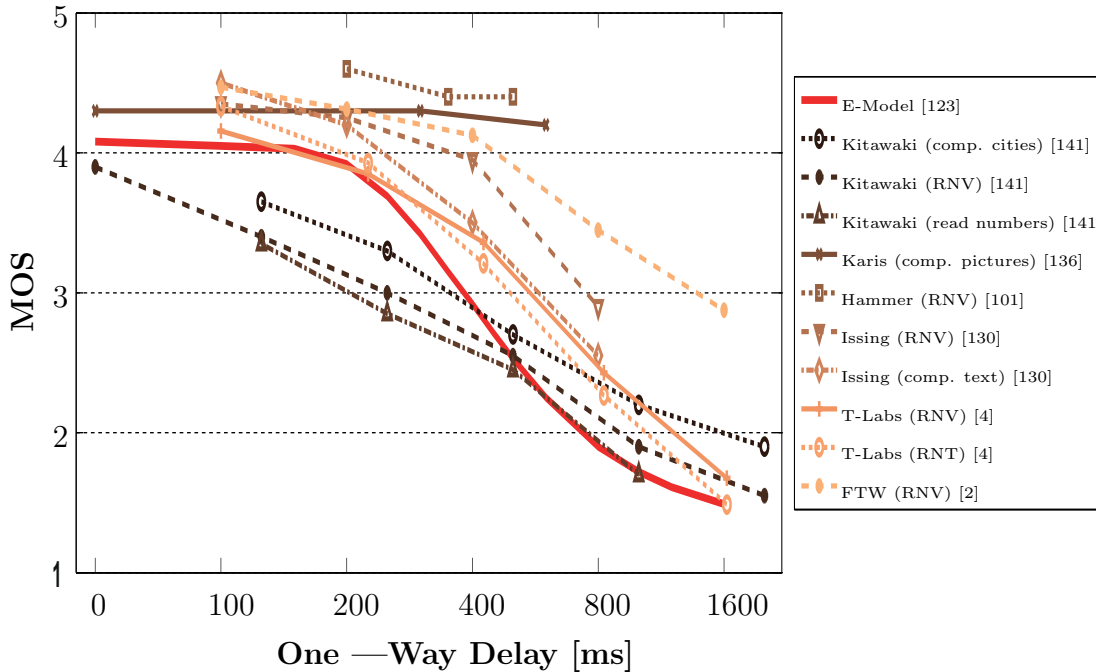


Figure 3.3: QoE (MOS) vs. transmission delay from related work for RNV and comparison task scenarios of comparable conversational interactivity

lay level in the range of 1.7 to 1.9 MOS points with the results in [4] being the only results reporting such a negative influence since the introduction of VoIP services. For the other results one can roughly distinguish two more groups, with the results from [175] and [176] being generally lower rated across the delay range compared to the results from [101], [98], [106], [2] and [130]. As mentioned above already, these offset differences might have been caused by different technical settings of the studies. But the more important comparison between these two groups is that the conversational quality drop between the lowest and highest delay setting is considerably lower as aforementioned in the range of 0.1 to 0.9 MOS. In terms of variance between slightly different conversation scenarios of similar interactivity it is interesting to see that the results by [101], which were acquired with four different scenarios (SCT, interactive short conversation test (iSCT), asynchronous short conversation test (aSCT) and free conversation), do yield comparable differences in conversational quality between the lowest and highest delay setting. Therefore, it can be concluded that scenario related differences within these results do not account for the strongly diverging conversational quality ratings in Figure 3.2.

In a similar fashion as the related work for the low to mid interactivity in Figure 3.2, Figure 3.3 compares conversational quality results from related work which

were obtained with conversation scenarios that induce high conversational interactivity in the resulting conversations. Within these results one can distinguish two groups of results based on their differences in conversational quality ratings between low and high delays. The first group is constituted by the results reported in [141], [4] and the text comparison results in [130] and shows a difference of about 1.5 to 2 MOS between lowest and highest delay setting. Group 2 contains the results from [136], [101], [2] and the RNV scenario from [130] with differences only in the range of 0.1 up to 1.3 MOS. In comparison to Figure 3.2 it is obvious that these higher interactive scenarios show a stronger influence of transmission delay on conversational quality. The results from the first group do even get close to the conversational quality estimates from the E-model [123] which are indicated with the solid red line. However, the differences of the results within the highly interactive scenarios themselves are still considerable.

Attempts to consider the impact of different interactivity levels and the relation to quality on MOS scales has been taken by [4], [100, 130]. Whereas the first study was mainly focusing on the impact of delay on interactivity rather than quality (and is therefore discussed in detail in Section 3.1.2.3), the latter two publications introduce models of the form $MOS = f(delay, interactivity)$ for conversational quality prediction. The linear model used in [130] was only used to prove the influence of delay and conversational parameters on the quality ratings and was not intended for high prediction performance of conversational quality. In contrast, the model presented in [4] proposes a modification of the E-model that takes into account conversational parameters (and thereby reaching beyond conversational interactivity alone) and yields good prediction performance results. It will be further discussed in Section 3.3.

From the above discussion it can be concluded that considering conversational interactivity levels through differentiation of conversational scenarios (and hence conversational interactivity levels established) can not explain the strongly diverging impact of transmission delay in the related work on conversational quality. This discrepancies can be explained with the following five influence factors:

- I:** the training of the test subjects might exert a strong influence on the acquired results of the respective study, e.g. the participants in [141] experienced delay effects on communications quality for about thirty minutes in the training phase and were therefore proficient in detecting delay impairments.
- II:** the experience of the test participants with transmission delays and their abil-

ity to detect related problems can play an important role, e.g. a part of the participants in [141] were trained experts and employees of the laboratory the tests were conducted in.

- III:** the perceived nature of the impairment might not be directly ascribed to the quality questions asked in the study. This can be of particular importance for the delay impairment as test subjects often ascribe conversational problems caused by delay to the personality or the affective condition of the other interlocutor.
- IV:** the time span between the experiments and the change in telephone systems in the meanwhile (mobile phones, VoIP) may have altered the expectations and experiences of subjects participating towards more delay tolerance in the more recent studies.
- V:** it is well known that the results of subjective tests depend on certain influence factors such as task execution, briefings, demographic variables etc.

Whereas the latter two factors can not be quantified a posteriori very well nor can they be out-ruled for upcoming studies, the factors I to III can be tackled by considering the course of the conversations itself rather than solely relying on methodologies that use ACR scales only. By tacking the resulting conversation structure into account further information about the conversational behaviour of interactants can be acquired which will help in understanding the causes of strongly diverging conversational quality results.

Therefore, conversation analytic approaches which set out to deeper analyse the impact of delay on the surface structure of conversations will be discussed in the following section.

3.1.2 Conversation Tests and Delay Impaired Conversations

The impact of transmission delay in telecommunication systems has not only triggered conversational quality centred research but has also resulted in the interest of the changes in interactant behaviour due to the delay impairment which lead to a number of approaches how to study this behavioural change. While the related work in the previous Section 3.1.1 is rather homogeneous, the different approaches in conversation analytics are more diverse and are therefore sub-categorised into *task performance*, *conversational surface structure* and *interactivity* related approaches

(cf. Figure 3.1) accordingly. As some of the related work discussed within this section addresses more than one category its contribution is discussed in each sub-category separately.

3.1.2.1 Task Performance

Task performance as a measure of the impact of certain influence factors has been widely used over decades in psychology (cf. [152]), sociology (cf. [142]) or usability research [168]. Typical task performance measures are task completion time, completion rate, error rate, error frequency, error probability or recognition rates. Such measures can be utilised in management studies where task performance is studied under the influence of e.g. goal setting, motivation or feedback as described in [82] or in group sociology where the task performance of the group can be influenced by group cohesion or leadership organisation in the group [142]. In speech communication research, intelligibility is a widely used task performance measure for assessing the quality of spoken dialogue systems [158].

In the context of mediated interaction the underlying assumption is that task performance is related to the ability of the communication system to ensure appropriateness of communicative interaction between the interactants. If different impairments hamper the proper interaction between the interlocutors the efficient completion of the given or implicit communicative task can not be sustained. Examples of such impairments are e.g. noise, which impedes intelligibility of the speech signal on the receiver side and thereby increases error rate for word recognition tasks or packet losses which leads to dropouts and hence lost information that might be necessary for the successful task completion. Theoretical considerations suggest that long transmission delays in the communication path can e.g. cause delayed arrival of interruption cues which then leads to longer completion time as longer utterances are exchanged. Another effect of transmission delays can be that floor control through speech pauses is elongated as each speaker tends to wait longer until taking the floor when collisions due to transmission delay have happened before.

In terms of studies on task performance and their relation to transmission delays the authors in [185] have conducted a study on call rejection rates due to transmission delays where they analysed the relation between call duration and rejection rate. They showed that increasing call duration (caused by increased transmission delays) led to an increasing number of rejected calls. Duration is a prototypical task performance measure, however as they didn't set a certain task for the analysed calls

this duration measure was affected by the different call purposes and the involved test subjects in addition to delay induced lengthening of the calls. Hence, insights from their results regarding the relationship between call duration and transmission delay cannot be gained.

Another approach towards task performance as measure for the impact of delay is reported in [146]. There, the authors defined the efficiency of a conversation with respect to the number of words needed for the completion of a visual comparison task: the less words were used the more *efficient* they labelled the conversation. Transmission delay decreased efficiency (hence task performance) of the conversations significantly for one-way delay of 900 ms compared to the 0 ms and 300 ms conditions. Between 0 ms and the 300 ms condition there was no statistically different number of words used for completing the task. Similarly to [185], the authors in [141] utilised a duration related measure: The ratio of completion times with delay to completion time without delay of certain tasks, termed conversational efficiency. Their finding was that conversational efficiency dropped with rising transmission delays and that no saturation of this effect was visible up to 1000 ms one-way delay.

Although these examples prove that transmission delay has a negative impact on task performance they fall short in identifying which conversational problems and conversation behaviour changes were caused by the delay and have led to the drop in task performance. Therefore, the following section focuses on the actual conversation structure and its alteration by transmission delay to get a deeper insight into the conversational problems caused by this impairment.

3.1.2.2 Conversational Surface Structure

Initially the analysis of human conversations has been used in ethnology and sociology for analysing social interaction of human beings in everyday encounters (not limited to mediated communication) and has been developed further into the research approach of conversation analysis as introduced by [187]. In this context conversational surface structure can be defined as follows:

Conversational surface structure: is the sum of observable or measurable interaction cues and interaction behaviour between two or more human interactants during the course of interaction.

By definition this concept includes all types of interaction cues that can take place in embodied communication such as verbal, non-verbal, visual, tactile, olfactory etc. cues. In the context of this section which is targeted towards mediated voice

communication the considered cues will be limited to verbal and non-verbal cues conveyed in the acoustic channel (= vocalic interaction cues⁴).

In the context of telecommunications initial work on the analysis of conversational surface structure has been conducted by [64, 65]. In these studies the author statistically analysed sojourn times of certain conversational states such as: talk duration, double talk duration, mutual silence time etc.. In addition he also analysed the probabilities of state transition in order to use these results for a statistical model of human telephone conversations as reported in [66]. By verifying predicted conversations of his model with human conversations he identified differences for conversational state and transition probabilities when transmission delay was present in the telephone system. Therefore, in [67] he applied his framework to the analysis of delay impaired speech in echo-free telephone circuits. The obtained results show statistical significant changes in sojourn times for certain states such as double talk and mutual silence for both delay conditions of 300 ms and 600 ms one-way delay. Also transition probabilities between states differed between the delay impaired conditions. Another finding reported by him was the fact that the remote speaker B was experiencing the local speaker A different than speaker A actually behaved as a consequence of the transmission delay. This is an important insight in terms of analysis methodology as it emphasises the importance of analysing each interlocutors *subjective conversational reality* as a result of the delayed transmission channel. In a similar fashion [105] analysed sojourn times of conversational states in delay impaired conversations up to 600 ms one-way delay. Contrary to the results from [67] he didn't find any statistical differences between conversations without delay and delay impaired conversations.

Another conversational state related measure was reported in [189, 204], defined as the ratio of the duration the participants actively speak or listen to the total call duration termed conversational efficiency⁵. Naturally this measure decreases over time with increasing mutual silence due to delay, hence giving not much more insight into changing user behaviour than mutual silence itself. In addition the authors have also defined conversational symmetry (cf. [189, 204]) as the ratio between maximum and minimum silence as perceived by an interlocutor throughout a (delay impaired) conversation, with conversational symmetry = 1 for ideal communications. The underlying rationale of this measure is that speaker alternations (equal to turn taking)

⁴A more detailed discussion on vocalic interaction cues can be found in [53] and [133].

⁵Despite its naming this ratio is not related with task performance and the conversational efficiency measure from [141] discussed of in Section 3.1.2.1

are affected by delay. Hence, the more constant this takes place the less distorted is the conversation. A major drawback of this ratio is the fact that mutual silence and conversational pauses respectively naturally diverge throughout a conversation and are also dependent on interlocutor behaviour and preferences, e.g. any kind of misunderstanding in the conversation can cause long mutual silence periods which will then lower the conversational symmetry measure without transmission delay being the cause. An additional problem is the fact that conversational symmetry is only computed for the conversational reality (=mutual silence period for speaker A, cf. Figure 3.4) of one interlocutor, which is not affected by delay, and it does therefore not consider the conversational reality of the counterpart (= arriving mutual silence period at speaker B, cf. Figure 3.4) which can strongly deviate due to the transmission delay. Furthermore, it does not consider interruptions in the human interaction process, which are an essential feedback cue for floor control and turn taking. Therefore, also this measure is not very useful to determine the influence of delay on conversational and turn taking behaviour in particular. An in depth discussion regarding interruption based measures will follow in Section 3.1.3.

In their work on interactivity of delay impaired conversations [100,101] (discussed in detail in Section 3.1.2.3) also analysed conversational states and state transition probabilities for delays up to 1000 ms. Additionally they conducted the analysis for several different conversation scenarios, namely random number verification (RNV)⁶, short conversation tests (SCT)⁷, asymmetric short conversation tests (aSCT)⁸ and free conversations (FC) (cf. [100,101]). Significant changes were only reported for the RNV scenario with decreasing talk duration⁹ and increasing mutual silence for delay increases from 200 ms to 350 ms and longer. In order to better capture the subjective perception of conversational events by the interlocutors the authors introduced *active interruptions* and *passive interruptions* which they defined as follows: "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 talking myself." (cf. p.57 [101]). These two measures account for the *subjective conversational reality* of the related speaker

⁶People have to match two lists of numbers through conversation.

⁷Interlocutors go through information desk scenarios with both sides having certain information which has to be exchanged.

⁸In contrast to the SCT scenario one person holds all information and the other person has to request this information (therefore asymmetric) in this scenario.

⁹termed *speech activity* in [100,101] but changed to talk duration for comparability issues to the results of [67].

(and therefore have to be computed for each speaker separately). Nevertheless, both of these measures do not analyse the result (in case of an active interruption) of the interruption on the other interlocutors side nor if the initiation of the interruption was deliberately issued by one of the speakers or was caused by transmission delays (in case of a passive interruption). How these measures can be further used to accomplish such an analysis is discussed in Section 3.1.4.

3.1.2.3 Interactivity

The inter-relation between a conversation's degree of interactivity and its vulnerability to transmission delays has been widely acknowledged and (implicitly) addressed in tests conducted in [141, 216], where the different scenarios used (implicitly) established different degrees of interactivity in the resulting conversations. The obtained results revealed the vulnerability of higher interactive conversations (in [141] the RNV, the 'comparison of cities'¹⁰ and the 'reading numbers in turns'¹¹ tasks were highly interactive). However, interactivity as influence factor has for a long time not been explicitly addressed and analysed in conversational quality related research.

Fundamental work on the measurement of interactivity and its relation to conversational quality and delay impairments has been conducted by [100–102, 180]. In a first step they developed metrics for the measurement of interactivity and in a second step they analysed the influence of delay on conversations of different interactivity on the dimensions of conversational quality and conversation surface structure. For the measurement of conversational interactivity they invented three different approaches, namely: *conversational temperature* (which relates thermodynamical principles to human communications), *entropy* (which is based on a speaker turn model, cf. [101]) and the *speaker alternation rate* (SAR, represents the number of speaker alternations per minute, cf. [101]). By comparing the results of these three different approaches they came to the conclusion that SAR can be used as a *simple and efficient metric providing a meaningful representation of interactivity* [101]¹². Recent work in [130] has taken up SAR as an interactivity measure used for modelling conversational quality when transmission delay is present. Further [4] uses a slightly altered version of SAR [4] to extend the E-Model [123] and improve

¹⁰In this task lists of cities had to be matched instead of numbers as in the RNV.

¹¹Here no matching of numbers had to be achieved but the numbers had to be read in an alternating fashion.

¹²The surface structure and quality related results from the studies conducted within the framework of this thesis are discussed in Section 3.1.2.1 and Section 3.1.2.2

its prediction performance.

The related work discussed in this section has shown, first that higher interactive conversations are more vulnerable to transmission delays than conversations with low interactivity levels, and second that interactional levels of conversations can be measured. Furthermore, the interactivity metric SAR proposed by [100, 101] can be efficiently used for comparing interactivity levels of conversations and is used for analysing conversational data in this thesis in Section 3.2.3.3.

3.1.3 Communication Theory based Considerations¹³

The above mentioned conversation analytical approaches (cf. Section 3.1.2) do make use of several communication theoretical assumptions without explicitly mentioning them nor providing a theoretical framework justifying the analysis along certain communication theoretical dimensions. E.g. the statistical description of conversations as used by [65] or [100] might be subscribed to the syntactic dimension of human communication processes. Therefore, a brief review of communication theoretic considerations will reveal how the analysis of delay impaired conversations and related conversation analytic approaches are framed in communication theoretic dimensions.

Based on the division of the field of semiotics (as being the study of signs and sign processes) into the three dimensions of *semantics*, *syntactics*, and *pragmatics*, the authors in [206] have proposed to transfer these divisions to the field of human communication (which can be considered as a special case of sign exchange). By doing so they came up with the following dimensions and related examples:

Syntactics deal with problems of message transmission, like, e.g. codes, channels, capacity, noise, redundancy and other technical and statistical properties of the transmission system

Semantics refer to the meaning of messages exchanged, and require mutual agreement about the meaning between sender and recipient.

Pragmatics describe the influence of the communication process and the transmission system on the behaviour of all participants.

¹³Parts of this section are based on original work from the author with adaptations as published in [19, 53]

From these examples it seems obvious that the targeted analysis of interlocutor behaviour and its alteration due to (disturbed) communication processes is strongly related to the *pragmatic aspects* of human communication and that this dimension is most relevant for the analysis of delay impaired conversations. However, a clear division between these three dimensions is not straight forward as there exist several interrelations between them. It is not a goal of this thesis to strictly separate approaches according to these dimensions but the presented work will focus mainly on pragmatic aspects of the communication process. Therefore, the question arises if the conversation analytic approaches discussed up to now address the *pragmatic dimension*, or if they rather lean towards the other two dimensions of syntactics and semantics. Applying the dimensions strictly according to the examples in Section 3.1.2.2, one could for instance assign the results of [65,66] also to the *syntactic dimension* as it analyses statistical properties of the communication. On the other hand, the application of this statistical description to delay influenced conversations as in [67,100] and the analysis of changes in the statistics caused by transmission delay can be ascribed to the *pragmatic dimension*. This example shows that conversation analytics per se are not automatically tackling the *pragmatic dimension* and one has to properly select interaction cues that are relevant for this dimension and for the identification of delay impairments respectively.

How to select the most important interaction cues? From a communication theoretical point of view a communication system can be treated as a cybernetic system [205]. Its desirable state is a stable equilibrium also termed *homeostasis*¹⁴. As a system (e.g. two interlocutors communicating) will never be in a total homeostasis the system control unit (the human interlocutors) is always concerned with trying to reach such an equilibrium by using feedback signals. In terms of human communication, feedback is twofold as it serves for turn taking as well as for interrupting the other interlocutor. Hence, properly distinguishing between turn taking feedback and system stability feedback is difficult. However, if the system gets disturbed e.g. by transmission delay feedback cues consequently increase to establish homeostasis again. Therefore, *feedback cues such as interruptions* and their influence on the communication process can serve as a measure of the delay induced disturbance.

From this discussion and the results from related work discussed in Section 3.1.2 one can conclude that a majority of the conversation analytic approaches from the related work does not directly tackle the relevant *pragmatic dimension* of human

¹⁴Homeostasis: self regulation of the communication system

communication, nor feedback related interaction cues. Most of the discussed approaches focuses on a description of certain conversational state probabilities and an analysis of probability changes due to the delay impairment. In this respect, the work of [100, 101] is a positive exception as the authors also utilised feedback cues such as *active* and *passive interruptions* for identifying the delay impact on conversations. The advantage of these measures is their ability to successfully express how interruptions are perceived in the conversational reality of each interlocutor. However, they do not reveal how delay impacts the arrival of these *active interruptions* on the receiver side or which *passive interruptions* are caused by the the delayed transmission channel. Therefore, two novel metrics are presented that consider this delay related interruptions in the forthcoming section.

3.1.4 New Conversational Metrics (UIR, I^3R)¹⁵

Previous sections have already made evident that it is important to consider the *subjective conversational reality* of each interlocutor (cf. Section 3.1.2.2, Section 3.1.3) and that this can be achieved by the *active* and *passive interruption* metrics introduced by [100, 101]. *Active interruptions* occur when an interlocutor (B) starts to speak, while he can still hear his counterpart (A) talking. In contrast, a *passive interruption* occurs when an interlocutor (A) becomes interrupted by the (delayed) arrival of a counterpart's (B) utterance which was issued by B. Additionally the discourse about the role of interruptions as feedback cues and how they are used to control distorted conversations from a conversation theoretical point of view (cf. Section 3.1.3) has underlined that an analysis of interruption related metrics is promising to understand the destructive nature of the delay impairment.

However, it has also been shown that the *active* and *passive interruption* metrics do not differentiate between interruptions caused by the technical system's delay and natural interruptions issued by the interlocutors. In order to overcome this differentiation problem the *unintended interruption rate (UIR)* has been introduced

¹⁵The new metrics introduced in this section are based on original work by the author with adaptations as published in [2], [3]

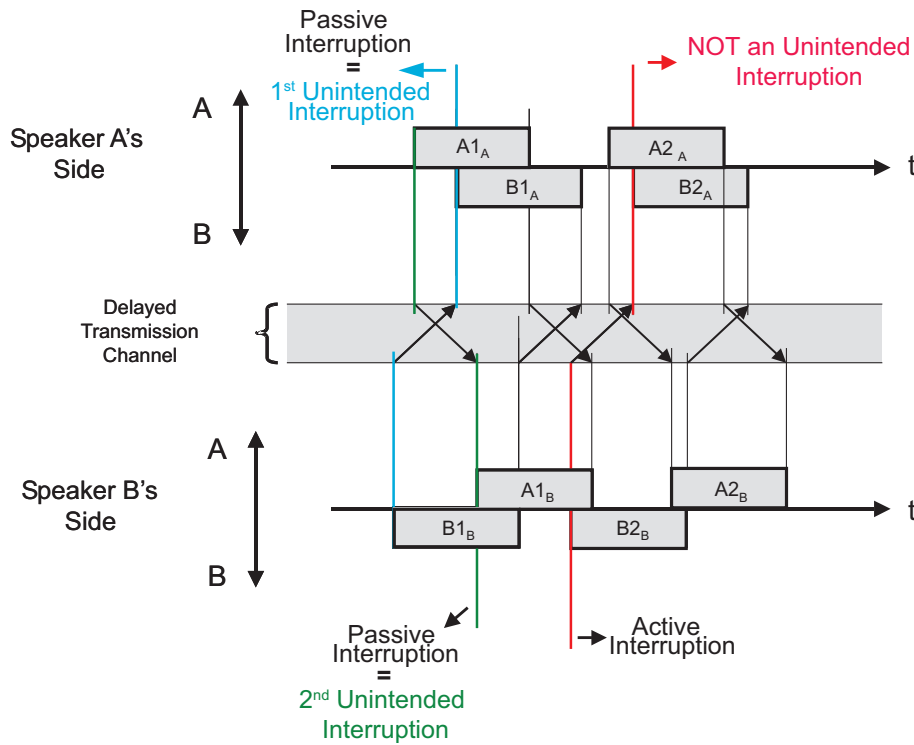


Figure 3.4: Unintended Interruptions Rate UIR

in [2] which is defined as follows and depicted in Figure 3.4:

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

While interruptions can be introduced by the delay, as in the UIR case, the opposite case can also come to pass: Deliberate interruptions of an interlocutor are not able to interrupt the other interlocutor due to the time shift caused by the transmission delay. Thus, the ability to interrupt the other interlocutor is hampered. To express this in a quantitative way, the *interruptive (and) intended interruptions rate* (I^3R) has been introduced by the author in [3] and [15].

The I^3R captures interruptions which are intended by one speaker and, despite the interfering delay, manage to interrupt the other speaker.

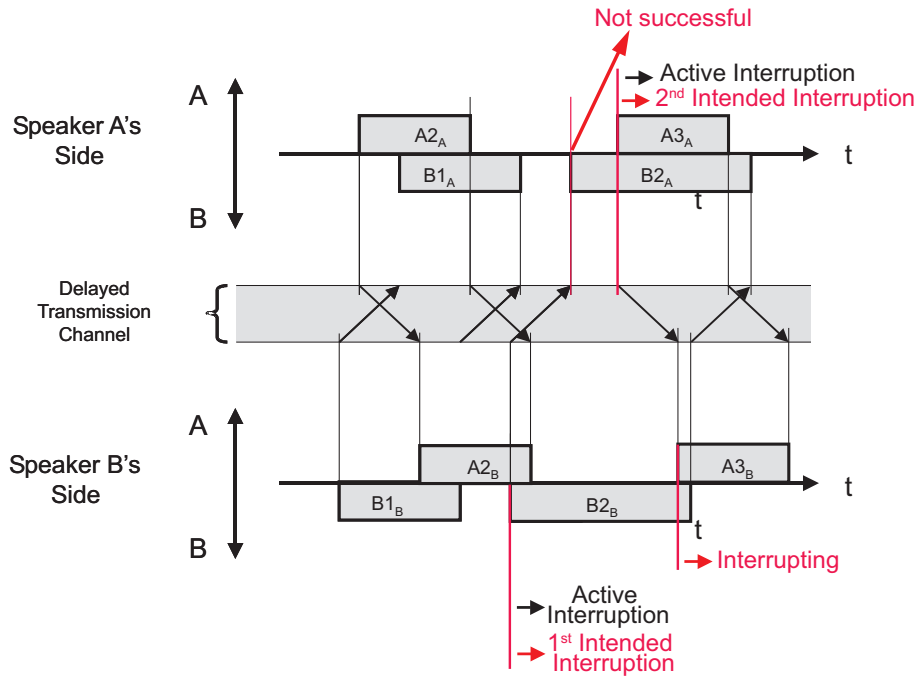


Figure 3.5: Interruptive (and) Intended Interruptions Rate I^3R

A graphical representation of I^3R is shown in Figure 3.5. As a result of this definition, the I^3R feature behaves contrary to the UIR with increasing delays as less intended interruptions arrive properly timed at the opposite interlocutor.

The advantage of these conversational metrics is their ability to differentiate between *disturbances introduced by the system's transmission delay* from interruptions deliberately issued by one speaker in order to interrupt the other. Thereby, being to the best of the author's knowledge, *the first work analysing these delay induced conversational defects* and providing an answer to research question RQ2. The relevance of these metrics, for truly understanding conversational behaviour changes and conversational QoE for delay impaired conversations, will be discussed in Section 3.2.3.4.

In terms of the practical implementation of these definitions, synchronised recordings of both speaker signals and the actual one-way delay for each transmission path are needed. The recordings are first converted to talk spurts by using voice activity detection, and in a second step *each speaker's interactional reality is reproduced* by shifting the talk spurts according to the transmission delay. For each speaker it can then be determined if he was (passively) interrupted (e.g. passive interruption of $B1_B$ in Figure 3.4), and if this interruption was caused by an active interruption (=in-

tended interruption). If no active interruption (as the 1st unintended interruption by speaker A in Figure 3.4) caused this interruption in the other speaker’s conversational reality, it was an *unintended interruption*. Furthermore, the two interactional realities also allow to identify if an active interruption (2nd intended interruption that interrupts utterance B2_A in Figure 3.5) in reality A interrupts the same utterance (B2_B in Figure 3.5) in reality B, and hence constitutes an *interruptive (and) intended interruption*.

3.2 Subjective Experiments¹⁶

The aim of the conversational studies described in this section is twofold: First, they should provide subjective conversational quality ratings and second, they provide synchronised audio recordings of the delay impaired conversations. Such synchronised recordings allow the computation of the aforementioned conversational parameters as well as the computation of the introduced metrics and a detailed analysis of their behaviour under delay influence. The acquired dataset consists of two studies: Study 1 was conducted in the i:Lab premises of FTW in Vienna whereas Study 2 was conducted in the lab facilities of Telekom Innovation Laboratories in Berlin.

3.2.1 Technical Setup

In both studies, subjects were seated in two separate, acoustically treated rooms and connected through a VoIP system as depicted in Figure 3.6. Both facilities were set up according to [114,120]. For Study 1 participants utilised VoIP clients on standard consumer grade laptops in conjunction with monaural headsets, whereas in Study 2 the participants were using Snom 870 VoIP telephones for conversation. In both setups it was ensured that no echo was perceivable in the transmission path even for long delays. In order to gather recordings of both interlocutors, their speech signals were captured synchronously by microphones placed in front of the participants on their tables and stored on a centralised server. These synchronised signals were then used for the extraction of the respective talk spurts. For the transmission delays first the minimum achievable delay was measured and then complemented with

¹⁶This section is based on original work from the author with adaptations as published in [2], [3], [4]. In terms of test execution, the author has setup and supervised the tests in Study 1 and contributed actively to the preparations of Study 2 which was executed at Telekom Innovation Laboratories in Berlin. All results and their analysis as shown and discussed within this section have been computed by the author.

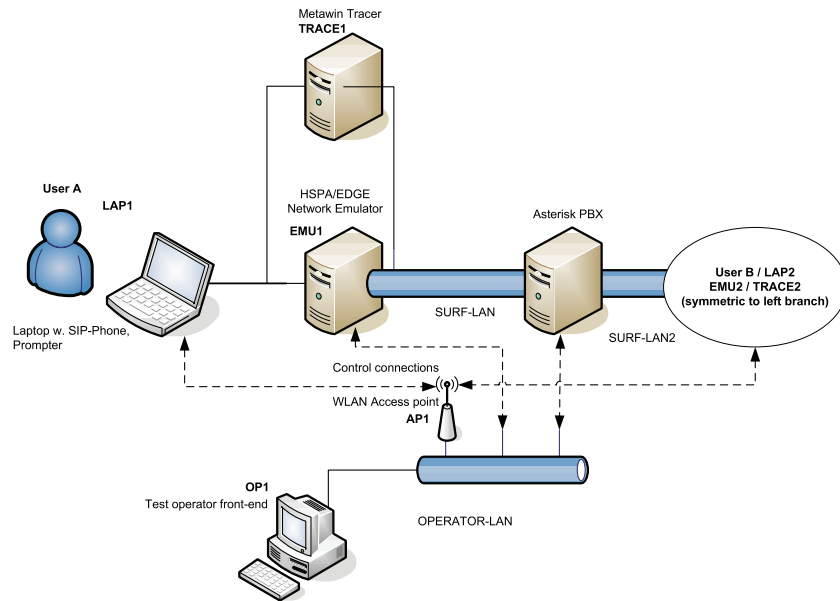


Figure 3.6: Testbed at FTW's i:Lab

adjustable delays from the network emulator in order to achieve the delay settings as depicted in Table 3.1. In order to examine delay ranges where severe degradations occur for sure we decided to expose users to one-way delays up to 1600 ms.

3.2.2 Tasks and Test Procedure

In order to assess different degrees of conversational interactivity, test subjects were asked to accomplish different scenarios. We chose the scenarios based on the recommendations in [120, 162] and according to the degree of interactivity they introduce in the conversation (cf. [101]). Based on these results the short conversation tests (SCT) was chosen as a low conversational interactivity task whereas the random number verification (RNV) task represented high conversational interactivity.

While the SCT and the RNV task were performed in both studies, the random number verification timed (RNT) task was only performed in Study 2. Scenarios of the latter type were considered for a prize written out for the fastest and most correct pair of conversation partners. Hence, the participants were encouraged to finish the scenario as fast as possible. The rationale behind this decision was that the introduced time pressure would foster even higher interactivity and trigger higher sensitivity for transmission delays within the test participants. The task sheets provided to the participants were identical for both studies. Within one test session

participants experienced five different delay conditions, ranging from 100ms up to 1600ms in order to span a decent range of transmission delays, for each scenario as described in Tab. 3.1.

After arrival the subjects were informed regarding the procedure and the nature of the scenarios. In order to acquaint them with the scenarios, they were asked to practice each scenario type with the lowest delay setting (100ms) in a warm-up condition. After completing the warm-up conditions we followed up with the different delay conditions. Subsequent to each test condition (approx. 1 to 3 minutes, depending on the time needed to complete the task), participants were asked to rate the integral perceived quality of the system on a 5-point absolute category rating (ACR) scale ranging from 1.0 (bad) to 5.0 (excellent) according to [114]. In Study 2 the RNTs were clearly indicated so that participants were always aware of whether the current scenario counted for the competition or not.

In Table 3.1 an overview of the used scenarios the user demographics and the technical settings for each study are given.

	Study 1 (FTW)	Study 2 (T-Labs)
Number of subjects	34	48
Mean Age of subjects	23.15 (SD=3.36)	30.44 (SD=8.36)
Female/ Male	F: 11 / M: 23	F: 24 / M: 24
Network	VoIP + NetEm	VoIP + NetEm
Codec	G.711	G.711
Subject Nationality	Austrian / Spanish	German
Conversational Tasks	SCT ₁ , RNV ₁	SCT ₂ , RNV ₂ , RNT
Delays[ms]	100,200,400,800,1600	100,225,425,825,1625

Table 3.1: Experimental conditions and locations of both studies

3.2.3 Result Analysis

The analysis of the gathered dataset is structured as follows. Starting with a discussion of the subjective quality ratings and their differences in Section 3.2.3.1, a description of qualitative observations regarding the task execution and resulting conversations will be given in Section 3.2.3.2. Furthermore, certain conversational states and their changes due to the introduced delay will be analysed in Section 3.2.3.3. The impact of transmission delay on the feedback abilities of the interlocutors via interruptions and the changes of interlocutors' interruption behaviour will be finally discussed in Section 3.2.3.4.

3.2.3.1 Quality Ratings

Figure 3.7 shows the obtained MOS values for both studies and all used conversation scenarios versus the transmission delay introduced. Surprisingly, the subjective ratings for identical scenarios differ strongly. Although external factors were kept as close as possible as described in Section 3.2, the resulting MOS ratings differ up to 1.6 MOS (in case of the SCT scenarios at 1600ms delay) for the same delay settings and same scenarios as shown in Figure 3.7(a). This was unexpected as we put considerable effort into making the setups as close as possible via using the same technical settings as well as identical instruction materials for the subjects. Of course there are further factors that can influence subjective results (cf. [162]) in general such as e.g. scale usage due to preceding expectations, rating behaviour based on demographic properties etc. Taking such an offset due to scale usage into account one can subtract the offset values from the best conditions and thereby "normalise" the ratings (for the normalised plot in Figure 3.7(b) the SCT scenario from Study 1 was used as baseline for normalisation). Doing so reduces the maximum MOS difference down to 1.3 MOS in case of SCT and 0.9 MOS in case of the RNV scenario as depicted in Figure 3.7(b). Both these differences are still considerably different on a statistically significant level, therefore differences in the conversational quality ratings due to scale usage can not be the cause in these cases.

The conclusion from the quantitative ratings is that a significant difference in quality ratings for same scenarios at identical delay values exists even in the case of normalised ratings, hence the subjective testing methodology is not producing sufficiently valid results (which is not assumed as the used methodology has been proven over decades to produce comparable results across different labs) or the conversational behaviour and the conversational interactivity respectively must have been different between the studies. In order to better understand the differing conversational behaviour that arose in the two studies (which has been the most probable root cause of these differences), it is essential to quantify the conversational behaviour of interlocutors and thereby identifying conversational problems that have been caused by the delay impairment, which are obviously not captured in conventional a posteriori quality ratings on ACR scales as recommended in [114]. A first glance on potential conversational problems that can emerge is given in the next section and will be followed by a detailed conversation analytic discussion of differences in interaction performance aspects and their potential effect on the subjective quality rating behaviour of the subjects.

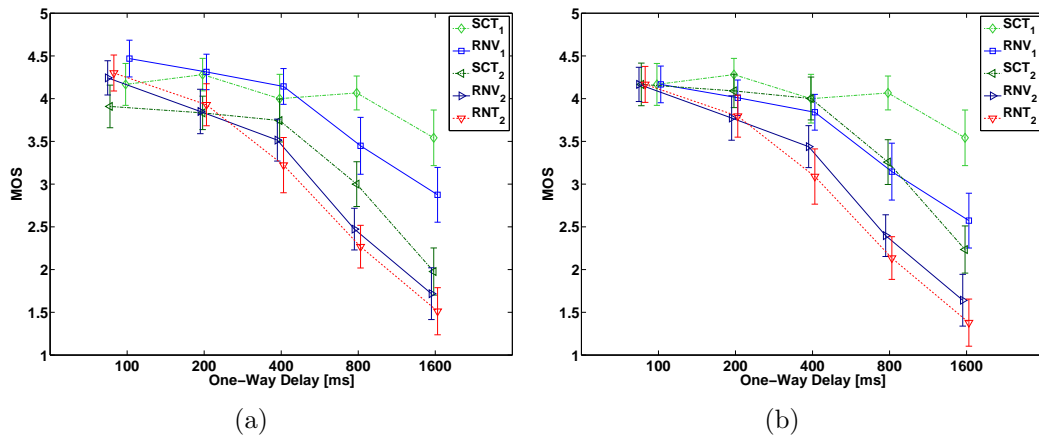


Figure 3.7: Subjective quality ratings (MOS) for different transmission delays as acquired in Study 1 and Study 2 for (a) and normalized to the ratings of SCT₁ for (b).

3.2.3.2 Qualitative Analysis

Due to the strong differences in subjective quality ratings as reported in the preceding Section 3.2.3.1, a qualitative analysis of the recorded conversations has been done to understand if differences in the conversational behaviour were present despite the identical instruction material.

The qualitative analysis of the RNV scenarios reveals that different task execution for RNV scenarios has resulted in different conversational behaviours. In Study 1 the resulting conversations of the RNV₁ scenario were well structured (as intended) but still contained elements of natural speech behaviour. These elements were introduced by the frequent use of conversational shortcuts which were not prevented by the test supervisor. Such conversational shortcuts aggregated several numbers from the task sheet in one utterance, contrary to the intention to use one utterance per number, forcing the other participant to interrupt in case of deviating numbers. Contrary, in the RNV₂ scenario the supervisors did not allow such conversational shortcuts which resulted in more structured conversations. This qualitative finding will also be visible and discussed in the quantitative conversational surface structure metrics shown and discussed in the following subsections. For the SCT scenarios such differences are less pronounced. The only perceivable difference is a slightly faster pace of the conversation among Study 1 speakers. A natural explanation of such a difference could be the different demographics in Study 1 compared to the demographics of Study 2. However, this behaviour could not attributed to Aus-

trian or Spanish group affiliation but rather to the individual interaction behaviour of certain interlocutor pairs. The subsequent sections will mainly focus on the delay influence on the conversational parameters and metrics but will also refer to the qualitative findings reported in this section.

3.2.3.3 Conversational States and Speaker Alternation Rate

For the computation of the different conversational surface parameters discussed in this section the speech activity over time of both interlocutors is needed. Generally, this is referred to as talk spurts. Talk spurt information has been extracted from the synchronised recordings of participants' speech by using a long-term-spectral-envelope voice activity detector (LTSE-VAD) as described in [153]¹⁷. The most relevant parameters for the extraction of conversational talk spurts are described in Table 3.2 below, all other parameters were kept at their default values. For the computation of the conversational parameters and metrics described in Section 3.1.2 and used in the remainder of this chapter, we shifted the synchronised talk spurts according to the amount of delay set for the respective condition (cf. Figure 3.4 and Figure 3.5).

Parameter	Setting
minimum vocal duration	200 ms
minimum silence duration	100 ms
frame length	25 ms

Table 3.2: Parameterisation of the LTSE-VAD algorithm used for the computation of talk spurts.

The initial conversational parameter analyzed is the *speaker alternation rate (SAR)* introduced by [100] as a measure for the degree of conversational interactivity. Figure 3.8 shows that for RNV and RNT the speaker alternation rate is significantly higher than in the SCT scenarios, thereby confirming the higher interactivity introduced by these conversational scenarios. Furthermore, the SAR is higher for the RNT task compared to RNV₂, especially for lower delay values confirming the effectiveness of the competition appeal. In terms of differences between the labs it is visible that the random number verification scenarios conducted in the T-Labs vicinities (RNT and RNV₂) resulted in a considerably higher SAR. This

¹⁷An implementation of the algorithm was kindly provided by one of the authors.

is in line with the qualitative observations reported in Section 3.2.3.2 and provides evidence of different task execution.

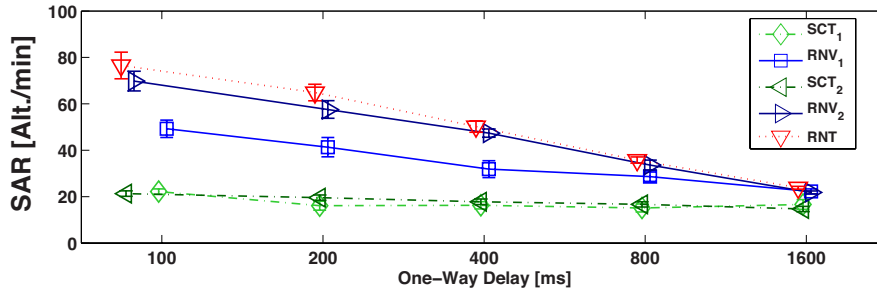


Figure 3.8: Speaker alternation rate (SAR) vs. one-way delay

In terms of delay influence the scenarios are differently impacted. While the random number verification scenarios are strongly impacted by increasing transmission delays, as the interlocutors have to wait longer for their opponent to respond, this is not the case for the short conversation tests. This is explained by the natural ability of human subjects to cope well with delay in case of close-to-natural conversation scenarios such as SCT. However, changes to other conversational parameters such as *mutual silence* as can be seen from Figure 3.9 show that human adaptation induces certain changes in conversation behaviour also in case of the SCT scenarios. The probability of *mutual silence* (*MS*) increases for all scenarios with rising delays. Interestingly, the state probability of *double talk* (*DT*) is not really affected by the delays but rather determined by the scenario and the lab the test took place in. This indicates once again the influence of scenario and different task execution (cf. Section 3.2.3.2) on the resulting surface structure of the conversation.

Although the discussed conversational parameters do render the influence of delay in a quantitative manner (in case of *SAR* and *MS*), they do not give deep insight in the changes of the conversational interaction process and how feedback cues are impacted. Therefore, the next section discusses the influence of transmission delay on interruption related conversational metrics.

3.2.3.4 Interruptions

Within the discussion in Section 3.1.3 it has already been shown that feedback cues such as interruptions are especially interesting for the analysis of the delay impact.

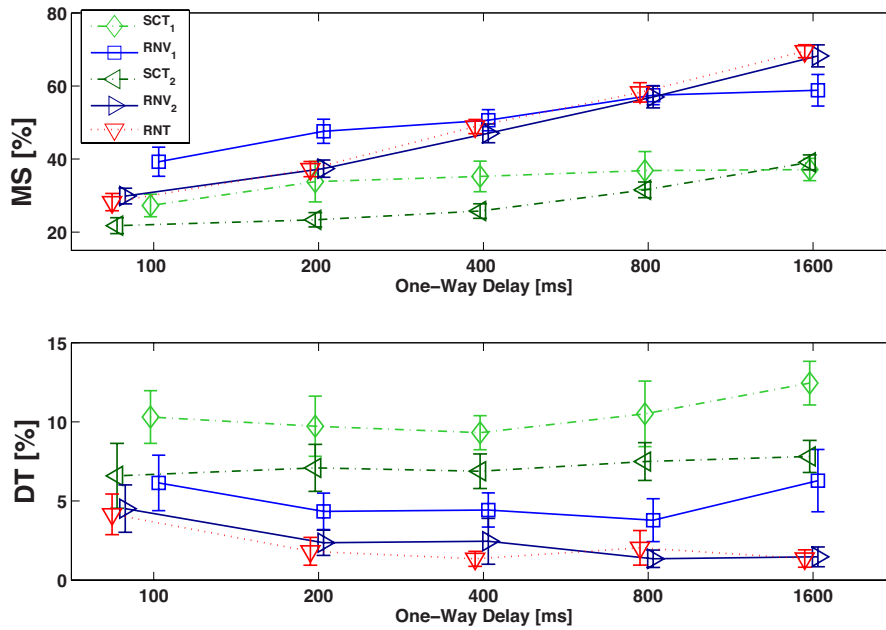


Figure 3.9: Mutual silence (MS) and double talk (DT) vs. one-way delay

In this section such changes in conversational interruption behaviour due to delay are discussed.

The *Active and passive interruption rates* (*AIR*, *PIR*) introduced by [100] and discussed in Section 3.1.2.2 are depicted in Fig. 3.10. It is evident that for the *AIR* the speaker behaviour changes for the *RNV* scenarios such that the interlocutors attempts to interrupt the opponent are decreasing with increasing delays. This means that the subjects do adapt their conversational behaviour such that they issue less interruptions and rather wait for conversational pauses to take the floor. For the random number verification scenarios this behaviour is reasonable as it is a very structured scenario which allows completion even without the use of interruptions. On the other hand, the *SCT* scenarios which are closer to conversational speech as the *RNV₂* and *RNT* scenarios show only moderate changes in the *AIR* behaviour (except for *SCT₁* at 1600ms), thereby suggesting that the interlocutors do not change their (active) interruption behaviour strongly in case of delay. Also for *RNV₁* this holds partially true, hence proving that the conversational shortcuts used there (cf. Section 3.2.3.2) resulted in a conversation behaviour closer to normal conversations compared to *RNV₂*. The *PIR* as shown in the lower plot of Figure 3.10

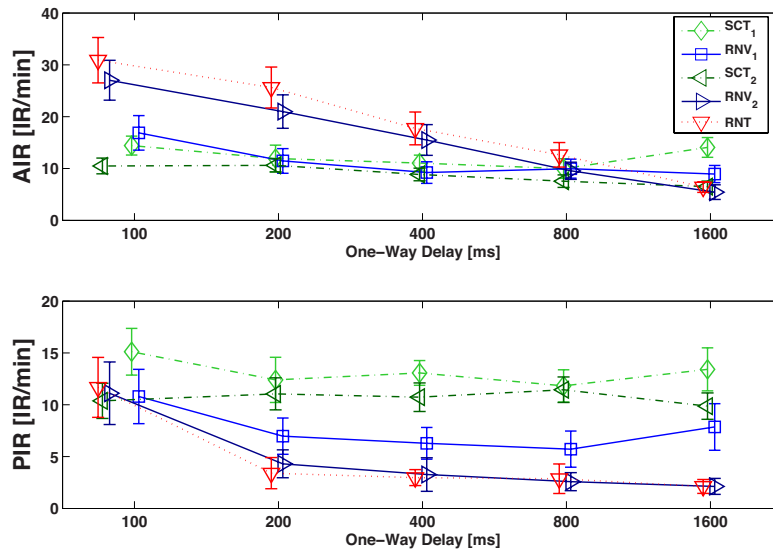


Figure 3.10: Active (AIR) and passive interruption rate (PIR) vs. one-way delay

expresses the interruptions an interlocutor experiences from the opponent being either intentionally issued or being caused by the delay. As discussed above for the AIR, the random number verification scenarios are more affected by the delay than the SCT scenarios.

Both of these metrics do not capture which AIR's at the sender side arrived (despite the delay) properly at the other receiver side nor which PIR's received were deliberately issued (hence being AIR's). To overcome these limitations the new metrics introduced in Section 3.1.4 are discussed in the next paragraph.

An incorporation of above mentioned critic towards the PIR metric is achieved in the *unintended interruption rate (UIR)* metric. This metric counts only unintended interruptions caused by the transmission delay, which can be identified by detecting passive interruptions on the receiver side B which have been issued on the sender side A while the sender was not hearing speaker B speaking and were therefore not actively intended by speaker A but caused by the transmission delay induced shift of utterances.

The lower plot in Figure 3.11 shows the relationship between transmission delay and UIR. For the SCT scenarios as well as the RNV₁ the UIR rises with increasing delays, meaning that the interlocutors get (subjectively) interrupted more often although the opponent is not interrupting more often as is shown in Figure 3.10

with the AIR rate. Also for this metric it is apparent that the random number verification scenarios are less affected by the delay impairment due to given speaker change structure of the scenario.

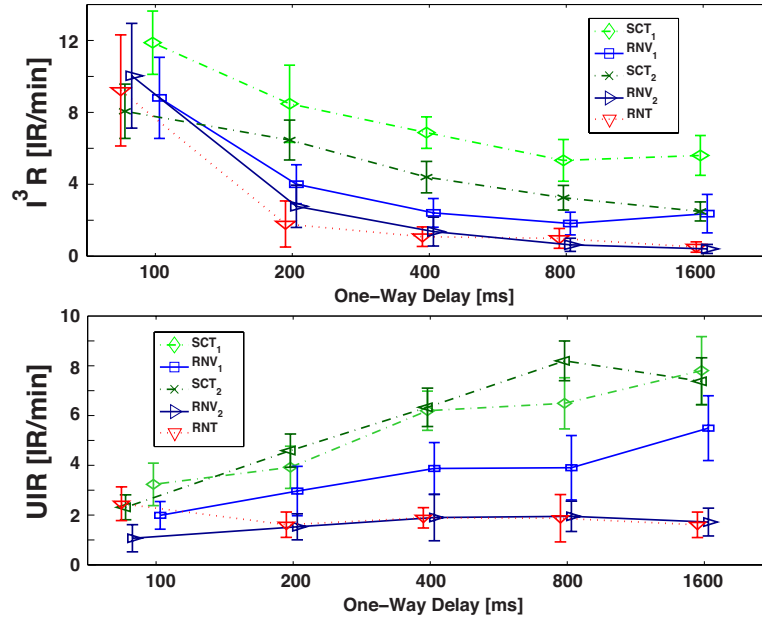


Figure 3.11: Interruptive (and) intended interruption rate (I^3R) and unintended interruption rate (UIR) vs. One-Way Delay

In a similar fashion the I^3R metric identifies which actively issued interruptions (AIR's) reach their goal, hence manage to interrupt the opposite interlocutor. Its behaviour for increasing delays as depicted in Figure 3.11 reveals the fact that the delay decreases the ability to successfully interrupt the other interlocutor for all scenarios, thereby proving that also this metric is able to capture the destructive nature of delay on the interlocutor's ability to successfully issue feedback backchannels. Its effect is differently pronounced for the two scenarios with being stronger for the random number verification scenarios.

However, one might argue that the interlocutors also adapt the amount of deliberate interruptions they issue as shown with the decrease of the AIR in Figure 3.10. In order to consider this argument the I^3R s are related to the respective AIRs in a next step. This ratio $\frac{I^3R}{AIR}$ provides information of the interlocutors ability to interrupt the opponent intentionally and how this ability is influenced by the transmission delay. A ratio of 100 % means that all active (and intended) interruptions fulfill their purpose and interrupt the other interlocutor. Obviously this ability is

highest for the lowest delay setting and an increase in delay reduces the amount of AIR's that reach the opponent successfully. While this is steadily decreasing for the close-to-normal conversations it drops harshly for the RNV₂ and RNT conversations. Both metrics, the I³R and the $\frac{I^3R}{AIR}$ show that they are a powerful tool for analysing the success of the possibility to interrupt the opposing interlocutor. Additionally, both metrics are able to capture differences in interactivity between the scenarios, as well as the influence of the transmission delay. Therefore, they qualify to serve as an additional input for QoE prediction models. In the following section an update for the E-model considering these two parameters as additional input will be proposed and analysed in terms of prediction performance.

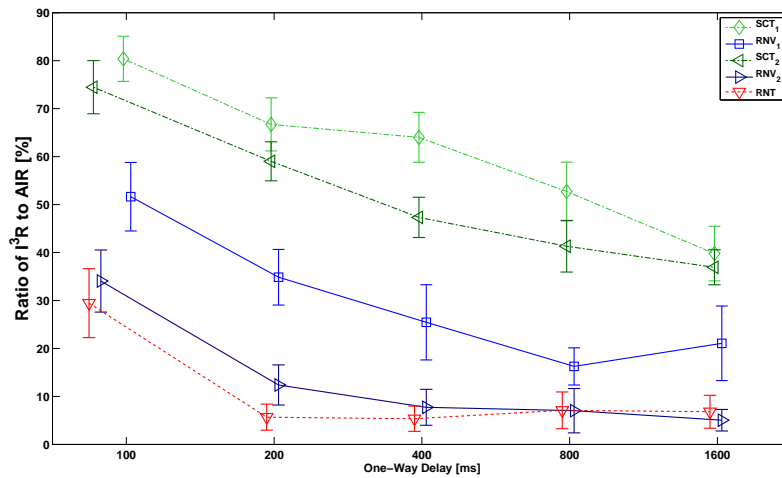


Figure 3.12: The ratio between interruptive (and) intended interruption rate (I³R) and unintended interruption rate (UIR) in [%].

3.3 Interruption Metrics, Delay and Their Impact on Conversational QoE¹⁸

The by far most popular model for the estimation of the influence of transmission delay on conversational quality is the E-model [123] standardised by the ITU-T as G.107 which is commonly used for network planning. It incorporates a large number of influence factors ranging from codecs over terminal parameters, expectation factors to transmission parameters like echo path loss, packet loss and (last but not least) transmission delay and predicts conversational speech quality. For the prediction of conversational quality it utilises the transmission rating scale also called R-scale (cf. [123]). The R-scales summarises all type of degradations in order to obtain the resulting conversational quality value R as follows:

$$R = R_0 - I_s - I_d - I_{e,eff} + A \quad (3.1)$$

With R_0 representing the basic signal-to-noise ratio, I_s as simultaneous impairment factor summarising all impairment factors occurring simultaneously with the voice transmission (e.g. loudness, quantisation distortions etc.), I_d represents impairments caused by delay and echo, the effective equipment impairment factor $I_{e,eff}$ representing impairments caused by low-bitrate codecs and impairments due to randomly distributed packet loss, and the advantage factor A that enables consideration of usage contexts and related expectations towards conversational quality, respectively (cf. [123]).

In the context of this thesis the factor I_d is of particular interest as it accounts for delay related impairments in the E-model [123]. It can be further decomposed into:

$$I_d = I_{dte} + I_{dle} + I_{dd} \quad (3.2)$$

where I_{dte} denotes the impairments caused by talker echo, I_{dle} denotes impairments caused by listener echo and I_{dd} expresses the impairments caused by echo-free delay.

¹⁸The E-model modification and its description as presented in this section is based on joint work on conversational quality from Alexander Raake, Janto Skowronek, Katrin Schoenberg and the author with adaptations as published in [4]. The E-model modification was proposed by Alexander Raake. Modelling of the relationship between the interruption metrics and the extended E-model parameters sT and mT has been done by the author. In terms of data acquisition for modelling through subjective tests, the author has set up and supervised the tests in Study 1 and contributed actively to the preparations of Study 2 which was executed at Telekom Innovation Laboratories in Berlin.

As the former two impairments have not been introduced in the experiments they are considered to be zero in the following modelling process. The factor Idd contained in Equation (3.2) (cf. [123]) is calculated as follows¹⁹:

$$Idd = \begin{cases} 0 & \text{for } T_a \leq 100\text{ms} \\ 25 \left\{ (1 + X^6)^{\frac{1}{6}} - 3 \left(1 + \left(\frac{X}{3} \right)^6 \right)^{\frac{1}{6}} + 2 \right\} & \text{for } T_a > 100\text{ms} \end{cases} \quad (3.3)$$

with

$$X = \frac{\log_{10} \left(\frac{T_a}{100} \right)}{\log_{10}(2)} \quad (3.4)$$

where T_a is the mean one-way delay on the transmission path. However, (3.3) does not take into account any conversation related parameters in its estimation of the delay impairment. As a result the predictions of the current E-model are pretty conservative and therefore do not predict conversational quality very accurately as shown in [16], [96, 98, 106, 162, 175]. For the conversational quality ratings acquired within the framework of this thesis the respective MOS scores have been transformed to the E-model R-scale according to the formula given in [123] and further recalculated to Idd' values according to the following formula from [4]:

$$I\vec{d}'_{xx,scen} = \vec{R}_{xx,G.107} + I\vec{d}_{xx,G.107} - \vec{R}_{xx,scen} \quad (3.5)$$

With xx indicating the test lab ($xx \in [\text{FTW}, \text{T-Labs}]$) and $scen$ indicating the test scenario ($scen \in [\text{SCT}, \text{RNV}, \text{RNT}]$). Here, $I\vec{d}'_{xx,scen}$ is the vector for the impairment values for the given lab and scenario to be used for model development, with each entry representing one delay condition. $\vec{R}_{xx,G.107}$ is the vector with the predictions provided by the current version of the E-model for the different delay conditions, and $I\vec{d}_{xx,G.107}$ the respective delay-related impairment factor vector and $\vec{R}_{xx,scen}$ being the to the R-scale transformed quality ratings from the respective tests conducted in the context of this thesis (cf. [4]). An example of the difference between the predictions and the real values is shown in Figure 3.13

In order to enhance the prediction performance of the E-model for Idd (cf. Equation (3.1)) it has been updated with two additional parameters mT and sT in [4]²⁰

¹⁹In order to stay conform with the E-model we directly utilise the formulas as given in [123]

²⁰In the ITU-T Study Group 12 general meeting on December 12th 2013 this update of the E-model has been approved by the general assembly of Study Group 12 and will be in force from 2014 onwards.

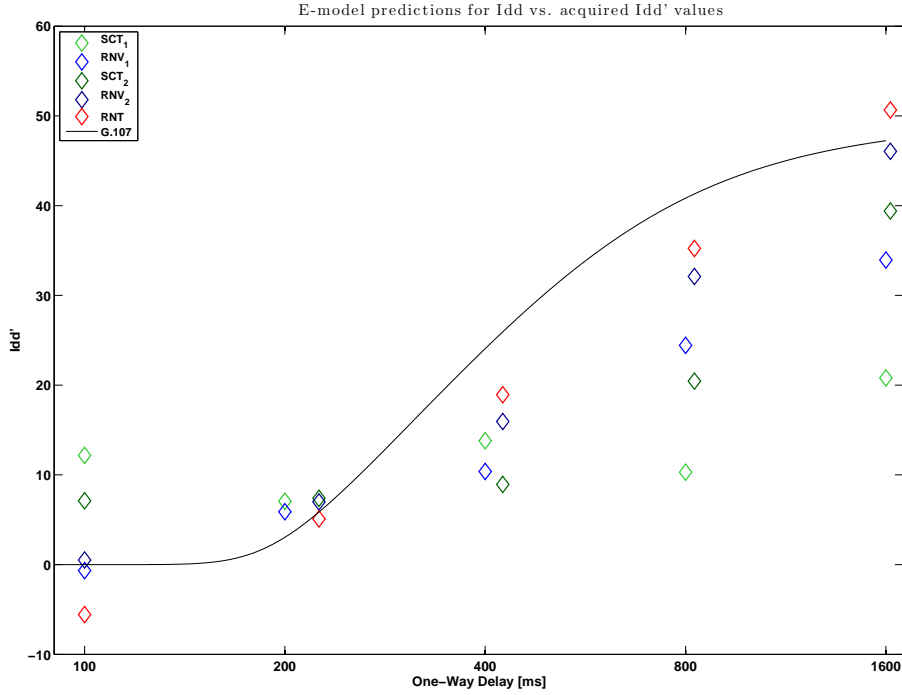


Figure 3.13: Idd values predicted by ITU-T Rec. G.107 (shown as solid black line) versus Idd' values for different conversational scenarios acquired in Study 1 and Study 2 (shown as colored diamonds).

shown in Equation (3.6), where mT denotes the minimal perceivable delay and sT is the delay sensitivity²¹ of the conversation (cf. [4]). The modification is chosen in a way that for the values of $mT = 100\text{ms}$ and $sT = 1$ the E-model modification is compatible to the actual version of the E-model in terms of Idd prediction, thereby ensuring conservative estimation of conversational quality for its use as transmission planning tool (cf. [4]).

$$Idd' = \begin{cases} 0 & \text{for } T_a \leq mT \\ 25 \left\{ (1 + X^{6sT})^{\frac{1}{6sT}} - 3 \left(1 + \left(\frac{X}{3} \right)^{6sT} \right)^{\frac{1}{6sT}} + 2 \right\} & \text{for } T_a > mT \end{cases} \quad (3.6)$$

²¹in this context *delay sensitivity* indicates the vulnerability of the considered conversation to transmission delays, with $sT=1$ denoting very high vulnerability and $sT = 0.1$ denoting that the conversation is barely delay sensitive.

with

$$X = \frac{\log_{10}\left(\frac{T_a}{mT}\right)}{\log_{10}(2)} \quad (3.7)$$

Based on this E-model modification respective sT and mT values for all scenarios of Study 1 and Study 2 have been derived through least squares curve fitting for the given function in Equations (3.6) and (3.7) (shown in Table 3.3). The results and the respective R^2 and RMSE values are shown in Figure 3.14. It can be seen that the inclusion of these two additional parameters in Equation (3.6) can considerably enhance prediction performance to $R^2 = 0.92659$ and RMSE = 3.9625 compared to the prediction performance for the current E-model of $R^2 = 0.3048$ and RMSE = 12.1629.

Scenario	mT	sT
SCT _{1,G.107 extension}	16.48	0.21
RNV _{1,G.107 extension}	135.29	0.44
SCT _{2,G.107 extension}	144.42	0.41
RNV _{2,G.107 extension}	120.07	0.72
RNT _{G.107 extension}	116.90	1.01

Table 3.3: mT and sT Parameters for IDD' prediction of the different conversational scenarios. Calculated with least squares fitting of the function given in Equations (3.6) and (3.7).

Although this E-model modification enhances prediction performance as shown above it still does not take into account conversational interactivity and related surface structure parameters for the selection of respective sT and mT parameters.

In order to close this gap the I^3R and $\frac{I^3R}{ATR}$ are used and related with sT and mT following the approach in [4]²².

The modelling procedure includes the following steps: First the mean values for the I^3R and $\frac{I^3R}{ATR}$ across the delay values have been calculated as $mean(I^3R)$ and $mean\left(\frac{I^3R}{ATR}\right)$. In a second step the sT and mT values have been plotted against these mean values followed by the third step of again least squares curve fitting with respectively chosen logarithmic and exponential functions. The functions for the curve fitting have been chosen based on the supposed relations outlined in the relation between the mean values of the interruption metrics and the model parameters.

²²In [4] the corrected speaker alternation rate (SARc), a surface structure measure introduced by Katrin Schoenberg in [3], has been used for this relation. The achieved prediction performance will be compared to the prediction performance with I^3R and $\frac{I^3R}{ATR}$ at the end of this section.

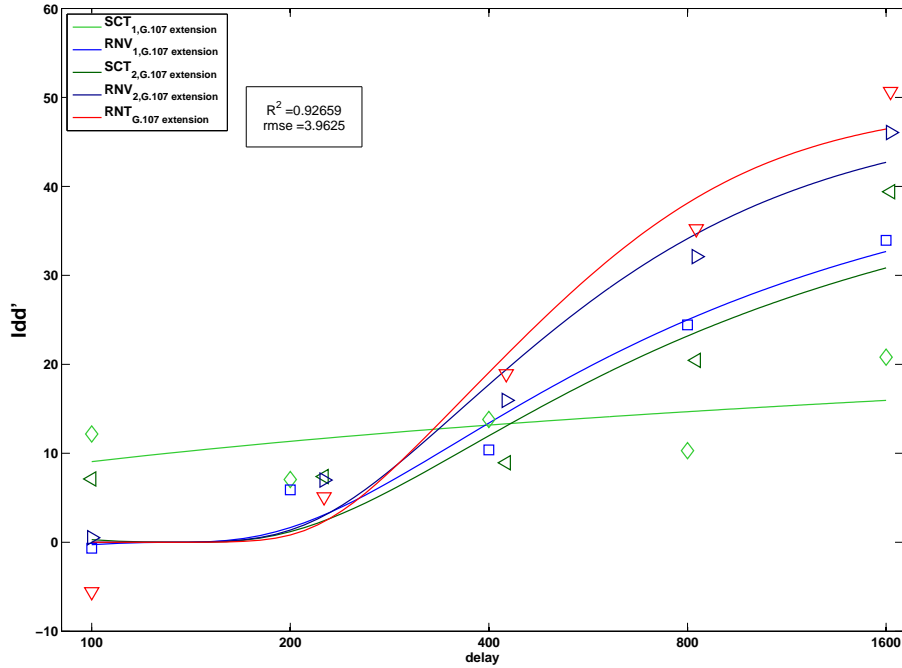


Figure 3.14: Idd' predictions from the extended E-model based on least square curve fitting results for mT and sT in Equation (3.6) and (3.7) for different conversational scenarios)

The candidate functions for the curve fitting of sT were:

$$sT = \alpha \cdot e^{\beta \cdot interruptionmetric} + \gamma \quad (3.8)$$

$$sT = \alpha \cdot \log(\beta \cdot interruptionmetric) + \gamma \quad (3.9)$$

with the same candidate functions also used for mT. After determining the coefficients for the chosen functions, as shown in Equations (3.10) - (3.13) for the two interruption measures, the respective sT and mT values for each scenario have been calculated and used to compute the respective Idd' values of the E-model modification which are shown in Figure 3.15 together with their prediction performance measures. The respective base functions were chosen according to the resulting stability and reasonability of the resulting predictions for sT and mT. E.g., the choice of the exponential function (cf. Equation (3.8)) in Equation (3.10) is based on the resulting asymptotic behaviour of sT for high I^3R values.

It can be seen that the computation of mT and sT based on the $mean(I^3R)$

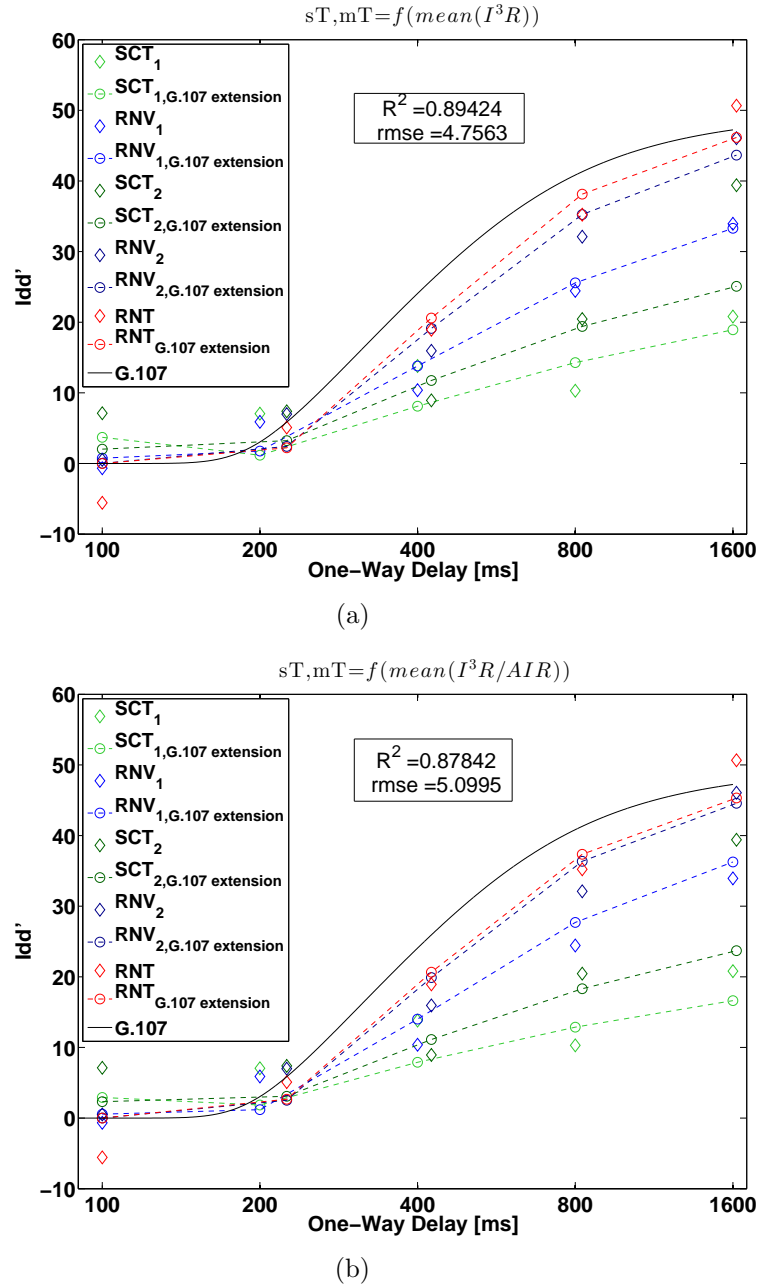


Figure 3.15: I_{dd}' values calculated with the E-model modification from Equation (3.6) and (3.7) and their prediction performance measures. The sT and mT parameters used for these plots have been calculated with Equations (3.10), (3.11) based on $\text{mean}(I^3R)$ for (a) and (3.12), (3.13) based on $\text{mean}\left(\frac{I^3R}{AIR}\right)$ for (b) for all scenarios.

values of the different scenarios yields the best prediction performance with $R^2 = 0.89424$ and $RMSE = 4.7563$. This holds also true when compared to the results from [4] where the speaker alternation rate corrected (another surface structure measure, introduced by Katrin Schoenenberg in [3], not discussed in details in this thesis) was used to derive sT and mT for the E-Model modification given in equation (3.6) and (3.7) and achieved $R^2 = 0.8623$ and $RMSE = 5.4263$. All of these results show that incorporating conversational surface structure measures for predicting Idd' enhances the prediction performance considerably compared to the performance of the current E-model with $R^2 = 0.3048$ and $RMSE = 12.1629$.

$$sT_{(I^3R)} = 16.4935 \cdot \exp(-1.1452 \cdot \text{mean}(I^3R)) + 0.2676 \quad (3.10)$$

$$mT_{(I^3R)} = 93.6729 \cdot \log(5.4785 \cdot \text{mean}(I^3R)) + 7.8719 \quad (3.11)$$

$$sT_{\left(\frac{I^3R}{AIR}\right)} = -0.8505 \cdot \log(-10.9674 \cdot \text{mean}\left(\frac{I^3R}{AIR}\right)) + 2.8946 \quad (3.12)$$

$$mT_{\left(\frac{I^3R}{AIR}\right)} = 40.8847 \cdot \log(16.8528 \cdot \text{mean}\left(\frac{I^3R}{AIR}\right)) + 24.4774 \quad (3.13)$$

3.4 Conclusion and Lessons Learned

This section summarises the insights given and results provided by the newly introduced conversational metrics, arguing for updated delay thresholds in conversational speech services. Finally, it discusses the results obtained with an E-model modification that utilises the $\text{mean}(I^3R)$ measure to incorporate the conversational reality of the analysed conversation into conversational quality prediction.

3.4.1 Appropriateness of New Interruption Metrics to Capture the Delay Impact on Conversations

The communication theoretic discussion at the beginning of this section has motivated the significance of the pragmatic dimension of human communication and underlined the importance of interruptions as feedback cues in disturbed human conversations. In a second step, two new metrics have been introduced that are able to differentiate interruptions that were deliberately issued by the interlocutors, or have been introduced by the delay induced shift of utterances. These two metrics represent an answer to research question RQ2.

The UIR expresses the number of passive interruptions at the receiver side that were induced by the transmission delay and were not issued as intended interruptions by the sender. This is supplemented by the I³R, which captures interruptions deliberately issued by one interlocutor (in order to interrupt the opponent) and manage to successfully interrupt him despite the transmission delay. As a third step these measures have been applied on a dataset of two conversational lab studies. The results reveal that both measures are able to give additional insight into the conversational interaction process, the impact of delay on the ability to issue feedback cues and how conversational disturbances are introduced by delay.

In addition it has been shown how the ratio of I³R to AIR provides valuable information about the interlocutors ability to interrupt the opponent intentionally and how this ability is influenced by the transmission delay.

3.4.2 Updated Delay Thresholds

The work presented in this section has successfully shown that a quantitative analysis of the conversational surface structure can be used to identify the impact of transmission delays in mediated communication systems on the human communication behaviour itself. This constitutes a new aspect that helps to better understand the conversation process which gives a deeper insight into conversation dynamics than the current quality assessment approaches that are only based on a posteriori assessment of conversational quality through ACR scales. Due to the gained insights into the delay impaired conversation process the following conclusions can be drawn:

1. There are clear differences between the SCT and RNV scenarios in terms of delay impact on their conversational structure.
2. Different task execution of identical scenarios can be identified by the conversational metrics. In RNV₁ conversational shortcuts and natural conversation elements have resulted in a lower grade of delay sensitivity for this scenario.
3. In case of the random number verification scenarios certain parameters and metrics (PIR, I³R, ratio of I³R to AIR) do decrease strongly from the lowest delay condition of 100 ms to the 200 ms condition already. This means that the conversational behaviour and the interruption abilities of the interlocutors are severely degraded already for 200 ms one-way delay.
4. For the SCT scenario, the analysis of the conversational parameters and metrics for the close-to-natural conversations, as in the SCT scenario, reveals the

ability of human interlocutors to cope with high transmission delays by adapting their conversation behaviour. Therefore, no clear saturation point of the curves can be identified.

Based on the above mentioned conclusions in terms of conversational behaviour under delay influence the following thresholds for telephone systems are proposed:

Highly interactive conversations: The obtained results suggest to keep the restrictive **150 ms** delay requirement from the original E-model

Normal / Free conversations: There is no evidence that transmission delays up to **800 ms** have a strong negative impact on the resulting conversational structure.

These proposed delay thresholds are in line with numerous results reported in literature such as [98] [106] [101], [136], [175] and [130] where it has been shown that the currently recommended delay threshold of max. 150ms from [117] is far too conservative for a broad range of conversation scenarios ranging from low interactivity to high interactivity. It has also to be noted that the strong influence of transmission delay as indicated by [117] has been only seen in the results from [130, 141] and in case of the RNT results presented in this thesis which confirms the validity of the above claim to relax the strict delay requirement regime for casual and free conversations.

3.4.3 New Interruption Metrics as Model Factors

In terms of conversational quality prediction under transmission delay, it has been shown that an E-model modification can yield considerable better prediction performance, by incorporating two additional parameters sT and mT in the calculation of the delay related degradation I_{dd} of the E-model.

In a second step it has been shown that these two parameters can successfully be computed by using conversational surface parameters such as $mean(I^3R)$ and $mean\left(\frac{I^3R}{AIR}\right)$, with $mean(I^3R)$ yielding the best prediction performance. This successfully answers research question RQ3, and demonstrates that conversation analytic metrics can be used to capture the delay sensitivity of interactive conversations and that the related information can be used for improving prediction performance of conversational quality models.

Chapter 4

Browser Based Applications over HTTP

In the past decade the Internet changed information retrieval processes of most people. The most widespread mean of informations access have since been web browsers (referred to as browsers in the remainder of this chapter). Technically speaking, information retrieval in browsers is characterised by a request – response scheme (as introduced in Section 2.2) where the user issues a request for a search result, a web page, a file download and so forth. Typically, the response to these requests is not instant but rather delayed to a certain extent (influenced by the type of request and the type of desired response). As a result, *user-perceived quality is largely dominated* by these *response times or waiting times*, respectively (cf. [62], [182]). An illustration of response times and waiting times respectively has been provided in Figure 2.3 in Section 2.2. Such waiting (time) for information is a recurring issue in browser based applications since their adoption. A citation within [137] puts an experience of early World Wide Web (WWW) adopters in the middle of the 1990s in a nutshell by asking “Tired of having to make coffee while you wait for a home page to download?”. This reminds of the growth of waiting times in Europe during the afternoon when the American users became active; users associated WWW with “World Wide Wait” [190]. In the meantime, the early-afternoon problem lost importance due to massive worldwide installations of server and network capacities. Today, we are facing other slowpokes such as overloaded terminals and access networks, or ineffective service chains.

In order to identify and model the influence of waiting times on QoE for browser based applications, this chapter discusses QoE assessment approaches, results from

related work, and the related shortcomings in Section 4.1. To overcome the identified deficiencies, Section 4.2 proposes a novel QoE assessment methodology that considers the interactive nature of the targeted applications. In a further step this methodology is used in a series of three user studies to collect a set of QoE training data for modelling purposes. Section 4.3 discusses fundamental psycho-physical relationships between QoS and QoE and how such relationships can be utilised for Web-QoE modelling. In Section 4.4 a logarithmic relationship is applied for mapping waiting times to QoE. The results prove that such a mapping is feasible for simple scenarios. A detailed discussion on shortcomings and challenges related with more complex scenarios concludes this chapter and thereby lays out a further roadmap for modelling requirements and improvements.

4.1 Related Work on QoE for Browser Based Applications¹

The importance of limited waiting times for successful e-commerce was investigated already in the early days of the Web. Results from [217] postulate an 8 s limit of page download time to be kept in order to avoid user churn. In the study [62], users were given tasks in a web-shop with deliberate additional delays. Most interesting are some citations of user reactions, such as

If it's slow, I won't give my credit card number.

This is the way the consumer sees the company...it should look good, it should be fast.

As long as you see things coming up it's not nearly as bad as just sitting there waiting and again you don't know whether you're stuck.

You get a bit spoiled. I guess once you're used to the quickness, then you want it all the time.

Obviously, waiting times affect user trust into the system and the company behind it, and can easily become showstoppers once money gets involved. Furthermore, decreases in waiting times increase user expectations on performance. Based on

¹This section is based on original work from the author with adaptations as published in [3,5–7,50] where the author actively contributed text to the related work sections of the publications.

the knowledge and experience of how quick responses could be given, subsequently growing waiting times are perceived as particularly disturbing.

In the context of browser based applications and waiting times, a number of guidelines exist. E.g. [116] defines maximum waiting times for such applications, unfortunately without empirical evidence on how violations of these guidelines do impact user perception. [167] gives similar recommendations about which waiting times are acceptable to facilitate perceived interactivity of interactive web applications. A major problem with such guidelines is the fact that neither the data used deriving them nor the methodologies used to derive this data are known. Another issue with guidelines of that nature is their frequent change which raises questions regarding their validity for long term planning of application needs. An example are the guidelines reported by [87, 97, 182] that postulated waiting times of 8 s as acceptable in 1999 and adapted this threshold to 4 s in 2006 and furthermore down to 2 s in 2009. Certainly, human expectations change over time depending on the experiences they encounter in everyday usage of the Web. Nevertheless, the aforementioned changes seem pretty drastically considering relatively stable evaluation of waiting times in related domains like time perception psychology (cf. [94]). Therefore, such guidelines are of limited use for deriving QoE models.

One step further towards real data describing the influence of waiting times on QoE are studies as reported in [62, 89]. They expose users to certain waiting times in the context of web browsing and assess their attitudes and quality perception respectively. In a similar way, studies like [80, 92] tried to quantify the influence of time fillers or design characteristics on the evaluation of waiting times.

The major problem with all these results is the fact that they are 1) derived with strongly diverging assessment methodologies and are therefore hardly comparable across each other, 2) the measures and scales used can often not be translated or compared to MOS, which is commonly used in the QoE community, and 3) the procedures they exposed their subjects to do often not consider the interactive nature of the application they try to evaluate. However, reliable and realistic assessment methodologies are a key element for gathering valid data for QoE modelling purposes. In the following section, existing assessment methodologies will be reviewed and their shortcomings identified.

QoE Testing Methodologies for Browser Based Applications In order to reliably assess QoE for browser based applications, appropriate testing methodologies are required. Surprisingly, scarce guidance regarding assessment methodologies

for Web QoE exists in related work as well as in standardisation. Concerning satisfaction scales used, early studies such as [61,62] have introduced measures of user satisfaction (3-point scale) and acceptability for web services. Later studies such as [172,188] have mainly utilised this satisfaction scale and extended it to a 5-point scale. Recent work on Web QoE [126], [112] and [192] has converged so far that utilisation of the MOS methodology and ACR scales from video and speech quality assessment (cf. [114]) has emerged as a de-facto standard for Web QoE evaluation. That such an adoption is reasonable has been proved by [5] where we² showed that such a transfer of scaling methods to new service categories is valid, even if the nature of the experience is different.

Although such an adoption holds true for utilised scales, this is not the case for test procedures itself. In contrast to the audio and video quality domains where psycho-acoustic and psycho-visual phenomena are dominant, temporal phenomena such as waiting times and latency characterise QoE of data services. As a result, this difference demands alternative approaches for measurement procedures and fore-closes transfer of test procedures from audio and video quality test methodologies. The above mentioned studies [61,62,112,172,188,192] as well as the study conducted for gathering the modelling data for [126] utilised a very simple page view procedure where users were requesting a single web page and waited for the page to load or two successive page loads thereby establishing a single or dual request-response action (cf. Figure 2.7 in Section 4.2). Although such a basic test procedure fulfils the reproducibility requirement, it falls short in establishing a realistic browsing experience for the test user. Data regarding web browsing behaviour as reported in [148,208] shows, that in terms of page views median values range between 3 to 17 page views per session, and session time values range between 150s and 405s, for different web pages. These values prove that a simple search task with single or dual page views, as used by above mentioned studies, does not reproduce real web browsing behaviour well. In addition, such a procedure does not reflect the flow based interactive nature inherent to web browsing as shown by [193]. In their work they identified five factors that contribute to flow experience in web browsing. Amongst speed, attractiveness, ease of use, challenge (content of the website), interactivity has been shown to have a causal relationship with flow experience.

In this section we have discussed that a major drawback of existing assessment approaches for Web-QoE are the test procedures the subject has to go through

²The author was co-author for this publication and actively involved in the test setup and data analysis of the test data.

before issuing the quality rating. In the following section a novel test procedure and related tasks that address these issues are introduced.

4.2 Subjective Experiments³

The purpose of the two lab studies (A and B) and the field study (C) described in this section was: 1) the verification of an appropriate subjective testing methodology for interactive browser based applications in order to provide an answer to research question RQ4 (introduced in Section 1.2), and 2) the acquisition of data for answering research question RQ5 (introduced in Section 1.2) regarding the relationship between waiting times and QoE for browser based applications. Therefore, this section discusses first a novel test procedure developed by the author for conducting QoE tests for browser based applications. In a second step, test setups used for data acquisition are described as well as experimental conditions and user demographics of all three studies are reported. Finally, example results are discussed regarding their reliability and external validity.

4.2.1 Novel Subjective Testing Methodology and Related Tasks

Due to the high complexity of real world web browsing in terms of numerous load times and page views per web browsing session we decided to include two different test procedures with related differences in browsing complexity in the studies. This is helpful for the analysis and modelling stage, as one can first work on the lower complex scenario, and with the learnings from this scenario one can then further analyse the more complex ones. For file downloads and simple browsing tasks the procedure described in [126] and for the more complex web browsing tasks a novel test procedure described in the following section were used. As discussed in Section 2.4 and the preceding section a test procedure for assessing QoE in browser based applications should meet the following requirements:

- Being close to real web browsing where people are browsing and interacting with web pages in order to acquire certain information. The procedure they go

³This section is based on original work from the author with adaptations as published in [3,5–7,50] as well as subjective experiments conducted within the ACE and ACE 2.0 projects at FTW. The author was responsible for the test design, test supervision, data analysis and parts of the technical setup.

through within this methodology should ensure that people get into a browsing or flow experience mode rather than a pure page loading mode.

- Subjects should be exposed to a certain QoS level over a time period rather than for one event, in order to experience several request-response cycles for the subjective evaluation.
- Certain tasks for each technical condition should stimulate the interaction between the web site and the subject for each test condition.
- The content, e.g. the web site should be interactive and has to provide a sufficient number of sub-pages, such that the subject can browse through the web site over several conditions without getting bored.

To meet these requirements, the tasks given to the subjects are important as they structure the subjects behaviour and the given length of the technical conditions after which the subjects issue their retrospective quality rating. Thereby, tasks strongly contribute to the first three requirements mentioned above. Hence, the following four principal tasks associated frequently with nowadays web usage are reasonable for subjective test of Web QoE:

- Browsing through an online photo album and selecting the favourite five pictures of personal taste.
- Browsing through a certain section (politics, sports etc.) of an online news page.
- Searching for vegetarian / Italian / Mexican etc. recipes on an online recipe database.
- Searching for a selection of five hotels one would choose in a given city.

Each of these tasks allows the subjects to interact with the given website for a certain time and throughout certain different technical conditions by varying the tasks slightly (e.g. changing the photo album for choosing the favourites, changing the news section etc.). In terms of session length, these tasks lead to web session durations of approximately 180 s, which is inline with results from [148,208] where it was shown that web session durations range between 150 s and 405 s. Furthermore, they ensure that a sufficient number of user clicks (= interactive acts) is performed and that the user experiences several different load processes throughout a web

session. After a session, the subjects are asked to rate their overall quality experience on a 5-point ACR scale as described in [114] and depicted in Table 4.1.

The *major advantage of these tasks* and the test procedure compared to the one described in [126] is that they establish a certain level of interactivity, which is one of the factors contained in the flow model of web browsing in [193]. Furthermore, the proposed tasks represent typical search and browsing tasks the subjects experience in their daily web routines. Thereby, they guarantee ease of use and give the subjects a sense of control about their actions, which are important factors of flow [170,193]. The use of real web pages and the amount of information available, makes sure that users can select content that is interesting to them and thereby prevents boredom. All together these tasks induce four factors from the flow model of web browsing as described in [193]: interactivity, ease of use, attractiveness (through the selected content, described in the next section) and challenge (also through realistic content with sufficient amount of information). The fifth factor of the model is speed, which is the independent variable for these experiments.

Grading Value	Estimated Quality	Perceived Impairment
5	excellent	imperceptible
4	good	perceptible but not annoying
3	fair	slightly annoying
2	poor	annoying
1	bad	very annoying

Table 4.1: ITU-T 5-point scale for Absolute Category Rating [114]

4.2.2 Test Content

Another *source of influence* in media related user tests is always the *content used*. This holds true for web content as well and ranges from low complex websites with only one or two visual elements to highly complex ones where a multitude of visual elements in different modalities is present (textual, visual, audio-visual). A *categorisation of content* has been achieved in [70] and is performed best *according to technical page complexity*. The complexity of a webpage can be captured best by the number and size of objects fetched to load the web page and also the different Internet media types (also referred to as multipurpose internet mail extension (MIME) in [88] e.g. image, javascript, CSS, text) across which these objects are spread. For user tests it is important to use content which is properly programmed for out-

ruling bad (web) development as source of influence (except it is the varied factor). In addition, attention has to be pointed towards the stability and performance of the server infrastructure of the used websites such that malfunctions through a test session can be minimised or detected.

For the content within the lab studies A and B we chose a representative set of web sites (photos, news, web 2.0, e-commerce) according to the above reasoning, namely: a customised photo album site using page sizes of approx. 250 kB (single photo view) and 500 kB (album view), *spiegel.de*, *expedia.de* as well as *chefkoch.de*. For the download scenarios, larger files (mp3, zip) ranging from 2.5 MB to 10 MB were used, and users asked to download the respective files and listen to them, or to unpack the zip file and view the contained pictures within.

The related tasks for above mentioned content are shown in Table 4.2. Similarly, users in study C were asked to provide ratings while web browsing and while downloading mp3 files (with a mean size of 5 MB) which were made available to them for the duration of the trial. For all the tests described in this section, monitoring of server and Internet performance of the network was achieved by passive network monitoring and a posteriori throughput estimation for the respective network settings. Thereby, we were able to identify server as well as delivery network related malfunctions.

<i>Task description</i>	<i>Abbrv.</i>
“Please browse through an online photo album and select your favourite five pictures”	PIC
“Try to get an overview of the current news in the given section”	NEWS
“Search for three recipes you would like to cook in the given section.”	COOK
“Please browse through the hotels in and select five you would like to stay in”	TRA
“Please download the given file”	DL
“Please go to the next picture / start a Google search query on ...”	PLT

Table 4.2: Subjects were asked to execute different tasks for each technical condition. The waiting time or downlink bandwidth was manipulated between the technical conditions.

4.2.3 Test Facilities and Study Setup

Studies A and B used a traditional lab setup, where the participants were performing web activities on a laptop connected to test servers and the Internet via a network

emulator (EMU). Monitoring of important application level performance parameters as objective page-load-time (PLT) was achieved through a browser plugin on the client laptop. The testbed architecture is depicted in Figure 4.1.

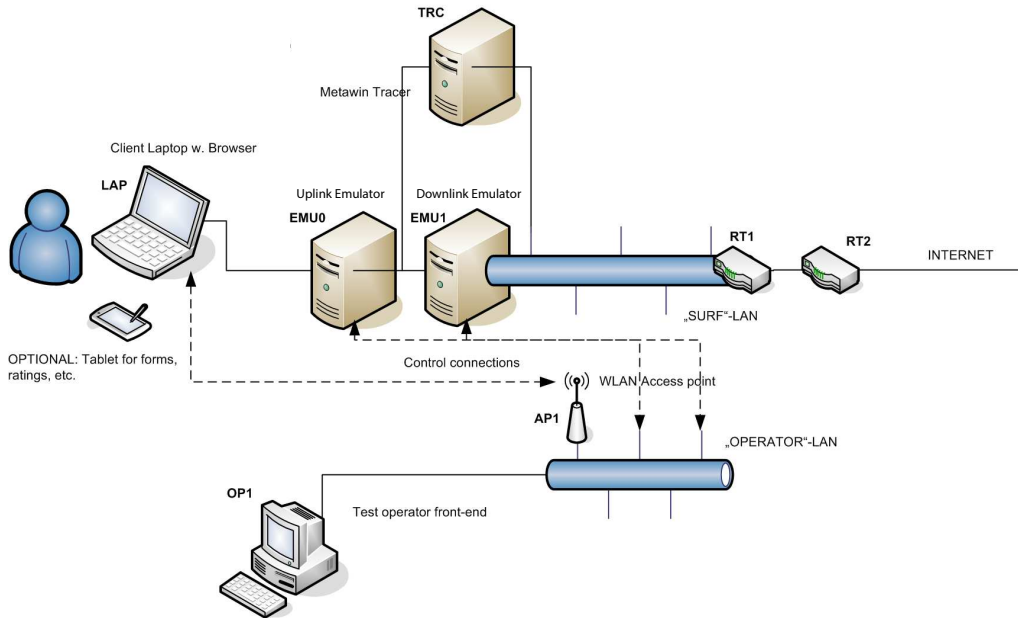


Figure 4.1: Technical setup of the two lab studies (study A and B).

We manipulated either the network throughput by limiting the downlink bandwidth (DL-BW) or directly manipulated the page-load-time (PLT) via Java scripts embedded in the respective web pages. In the case of web page load processes the page-load-time is defined as follows:

PLT is the time elapsed between the URL-request (e.g. caused by a click on a link) and the finished rendering of the Web page, as depicted in Figure 2.7 b).

Each lab study lasted for approx. two hours incl. briefing, training conditions, de-briefing interviews and a break of roughly 10 min in the middle of the test. For web browsing tasks and manipulated DL-BW, the test operator set different network conditions to be experienced, remotely started a browser session with the corresponding website, asked the user to perform a certain task and triggered electronic rating prompts after each condition. The condition durations from starting the browser session until the display of the electronic rating prompt varied from task to task as follows: In case of the free browsing tasks (PIC, NEWS, COOK, TRA from Table 4.2 the condition duration was approximately 180 s. For the PLT conditions, the users were asked to click through three pictures or to issue the requested

search queries and were prompted for their ratings immediately afterwards, hence condition durations were in the range of 5 s to 30 s. Finally, for the file download conditions (DL), the rating prompts were triggered after the file was successfully downloaded, therefore the condition durations ranged between 5 s to 300 s depending on the chosen downlink bandwidth.

	Study A	Study B	Study C
Number of subjects	26	32	17
Age of subjects	Mean= 35.42	Mean=28.39	Mean=30.51
Female/ Male	F: 10 / M: 16	F: 15 / M: 17	F: 8 / M: 9
Environment	Lab	Lab	Field
Indep. Variable	DL-BW	DL-BW, PLT	DL-BW
Service	DL	DL + Web	DL + Web
Year	2009	2011	2010

Table 4.3: Demographics and Experimental Conditions of the three Web QoE Studies.

In contrast to studies A and B, study C was carried out in the field. To this end, we used the 3G HSPA network of a leading Austrian mobile operator to enable subjects located in the city of Vienna to access the Web (see Figure 4.2). In order to emulate different network conditions, the IP traffic of trial participants was routed via our network emulator (EMU). The emulator shaped each user’s traffic according to different parameter sets which automatically changed every 30 minutes. Thus, participants experienced different quality levels and submitted their quality ratings in everyday contexts. In addition, we captured the network traffic (through the tracing Probe (TRC)) of each user in order to be able to measure e.g. the actual throughput the user was able to utilise, and thereby control the technical validity of the set parameters at the EMU. This was needed for study C as in the field setting the wireless link is subject to severe fluctuation in terms of throughput, RTT etc. as result of varying channel conditions. In cases where this a posteriori analysis yielded throughput estimates different (=lower) than the according parameter set for this time slot, we re-binned the related QoE rating to the corresponding true throughput condition.

Naturally, in the field setting task execution was slightly different compared to the lab. In terms of file downloads we asked the users to download three mp3 files of their choice per day from an online music portal they could access for free with an account provided by us. They were instructed to issue a rating at the

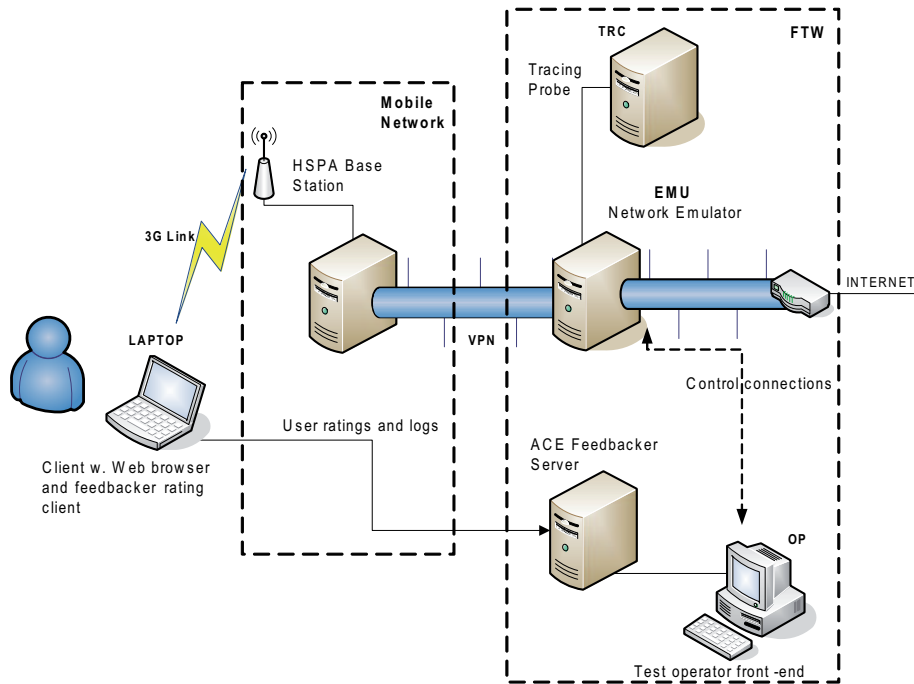


Figure 4.2: Technical setup of the field trial (study C).

end of the download via a custom application (ACE Feedback) executing on their laptops which forwarded the feedback to a central data collection server. The ACE Feedback application installed on the user laptops at the beginning of the test was running in the background, only showing up as icon in the system tray with two main tasks: First, it was monitoring the CPU load as well as the free memory available to the system in order to detect performance problems with the user laptop, and second, after clicking the icon it opened the questionnaire, recorded the user's quality ratings and forwarded them to the central data collection server. In case the user was not issuing enough ratings (the goal was three ratings a day per application which equals six ratings in total) the application reminded the user to issue the rating with a minimal invasive tooltip to the icon in the system tray. Download time as well as achieved throughput was extracted by the passive monitoring traces. Regarding web browsing we asked the users to access three times a day one website out of a portfolio of their favourite ten websites for approximately 3 min (=180s) and to issue a respective rating through the ACE Feedback. In total we aggregated six ratings on average per user and per day, totalling to approximately 130 ratings per user over a period of three weeks. The ratings were then correlated with QoS data characterising the network conditions the user experienced in the last three minutes

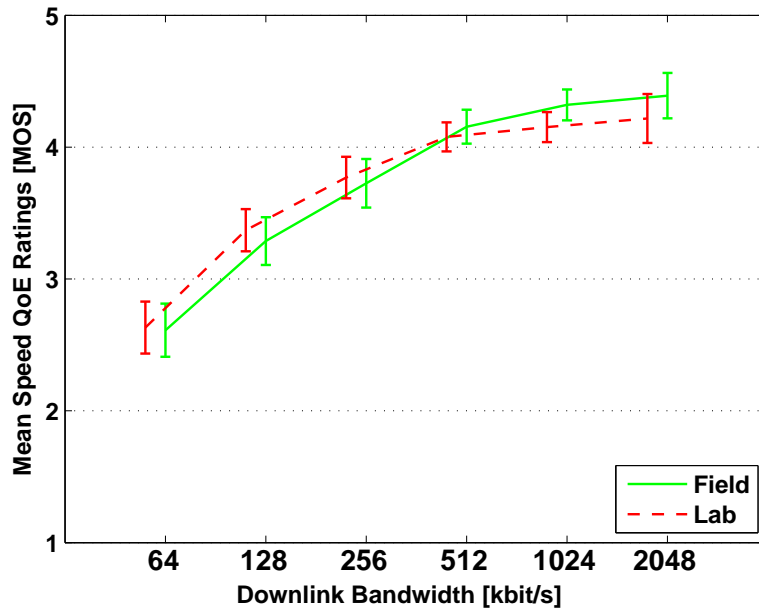
prior to issuing the rating. In cases where the network conditions didn't match with the parameters set through the EMU (due to e.g. bad wireless conditions) the ratings were assigned to the appropriate technical parameter settings for the data analysis. The demographic data of the subjects for each study is depicted in Table 4.3.

4.2.4 Verification of the Novel Subjective Testing Methodology

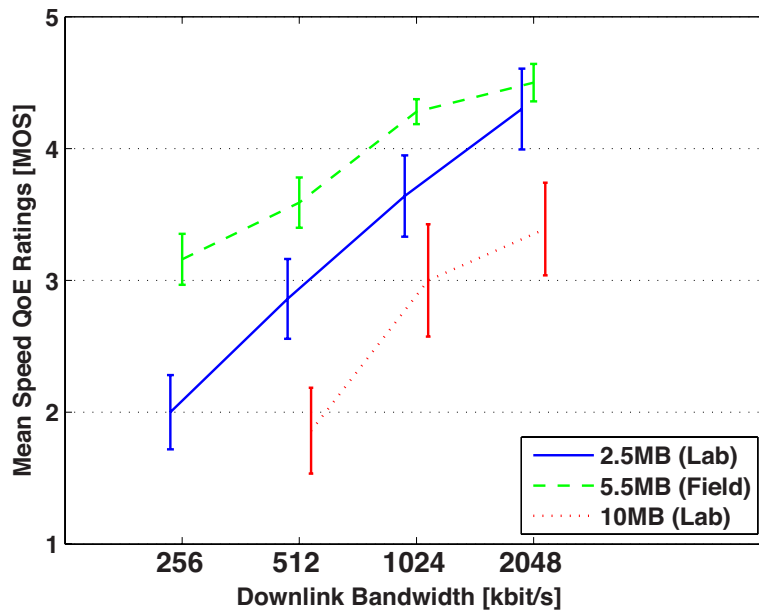
The results shown in this section are used to verify if the proposed test methodology produces reliable and valid results. An in depth discussion and analysis of the results in terms of QoE modelling follows in section Section 4.4.

Due to the different study environments 1) the lab (study A and B) and 2) the field (study C) we are in the fortunate position to compare the data gathered in the lab settings to the data gathered in the field. Such a comparison is of interest regarding the question of external validity of lab tests in the domain of browser based applications. In our case it was particularly interesting to compare the test methodology as used in the lab with results from users in their natural environment (which was considered to provide results of highest external validity).

The results of this comparison are shown in Figure 4.3. In case of web browsing the results are pretty coherent as can be seen from Figure 4.3(a) with more or less congruent rating curves across all downlink bandwidth settings. Regarding file downloads shown in Figure 4.3(b) the comparison is not that straightforward as we were using different file sizes in the lab and field environments. In case of file downloads, file size also has a certain influence on the resulting ratings which is explained by differing expectations regarding download times for different file sizes. However, the MOS ratings for the 5.5 MB file size (as used in the field study) does not lie in between the 2.5 and 10 MB ratings from the lab study as one would naturally expect. The most plausible explanation for this divergence is that file downloading in fact is a very simple, straightforward waiting task. In general, the perception of waiting times strongly depends on one's current attention span as influenced by task and situational context [191]. In the field study C, participants were in their natural environment exposed to several sources of distraction, whereas in the lab study A and B they had to wait and stare at screen until the file was downloaded and were not distracted at all. Therefore, users subjectively experienced waiting times to be longer which resulted in worse MOS ratings in the lab study. This is inline with findings from human time perception literature where subjects perceived temporal



(a)



(b)

Figure 4.3: Comparing rating data from lab and field environments for highly interactive web browsing ((a) across all websites) and file downloads (b).

stimuli to be shorter when other concurrent stimuli were present [56,191].

These results prove that lab studies, despite their lack of naturalness and the neglect of a multitude of influence factors, are able to provide valid results in case of an immersive task like web browsing. For file downloads, as a very strict waiting time task, results from lab tests can not be straightforwardly transferred to real life situations due to contextual influences. In Section 4.4.1 I will further analyse if the underlying (logarithmic) relationship between waiting time and QoE holds true, but with different parameters due to contextual influences.

Another finding within these results is the difference in rating behaviour between the two services. Web browsing ratings are rather moderate for very bad DL-BW conditions (2.82 MOS for DL-BW = 64 kbit/s) and go into saturation around 4.3 MOS from ca. 512 kbit/s onwards already. In contrast, file downloads are ranked far lower around 1.25 MOS for the low throughput conditions and continuously rise up to the highest DL-BW condition without showing a clear saturation point.

Summarising, these example results have verified that the subjective test methodology introduced delivers reliable and consistent results for web browsing across different contexts, and thereby provide an answer to research question RQ4. This also suggests that the achieved immersion and interactivity of the novel test methodology creates usage situations in the laboratory, that are comparable to real life usage situations, as experienced in the field trial, in terms of QoE perception. In contrast, the results for file downloads as plain waiting time task show that such a (static) assessment approach is prone to strong contextual influences. Furthermore, it has been showed that a difference in rating behaviour between web browsing and file downloads exists. Additionally, a result presentation mode including MOS as well as rating distribution information has been shown that gives comprehensive insights in user's QoE ratings beyond plain MOS plots.

4.3 Modelling QoE for Browser Based Applications by Identifying Fundamental Relationships between QoE and QoS⁴

Together with the increasing interest in QoE related research, also the modelling of fundamental relationships between QoS and QoE has received considerable attention (cf. [5, 8], [74, 86, 109, 192]). Fundamental relationships originate from the area of psychophysics where the main aim is to identify the relation between a physical stimuli (that can be sensed through human sensory organs) and the subjective perception of this stimuli with respect to its intensity. A major advantage of such fundamental relationships is that the interdependence between the stimulus and the subjective experience is constant and well described with a mathematical expression. Therefore, these relationships represent simple, unified and practicable formula expressing a mathematical dependency of QoE on network- or application-level QoS. They are thus applicable to online in-service QoE monitoring of QoS-related problems (e.g. as part of parametric planning or packet-layer models), enabling QoE management mechanisms that build on QoS monitoring [86]. Two prominent categories of such relationships which have been frequently observed in practice and have been discussed within QoE research are logarithmic and exponential relationships. Within the following two sections related work on these relationships used for QoE prediction models will be discussed.

4.3.1 Logarithmic Relationships – The Law of Weber-Fechner

A number of QoE experiments have identified relationships of the form

$$MOS = \alpha \cdot \log(\beta \cdot QoS) + \gamma \quad (4.1)$$

between QoE and QoS, be it in the context of web browsing (cf. [126] and [112]), file downloads [5] with waiting times as impairment or VoIP services [203] with packet losses as impairment.

Systematic studies of these observations [5], [181] revealed that these logarithmic relationships can be explained by the well-known Weber-Fechner Law (WFL) [207],

⁴This section is based on original work from the author with adaptations as published in [5, 8, 21] where the author actively contributed text to the respective related work sections, was involved in the test execution and performed the data analysis and modelling.

which in itself represents the birth of psychophysics as a scientific discipline of its own. In essence, the WFL traces the perceptive abilities of the human sensory system back to the perception of so-called "just noticeable differences" between two levels of a certain stimulus. For most human senses (vision, hearing, tasting, smelling, touching, and even numerical cognition) such a just noticeable difference can be shown to be a constant fraction of the original stimulus size. For instance weight estimation experiments have shown that humans are able to detect an increase in the weight of an object in their hands if this is increased by around 3%, independently of its absolute value. This is expressed by the differential equation

$$\frac{\partial Perception}{\partial Stimulus} \sim -\frac{1}{Stimulus} \quad (4.2)$$

As direct conclusion, the resulting mathematical interrelation is of a logarithmic form and can be used to describe the dependency between stimulus and perception over several orders of magnitude [207]. Where this dependency holds in the domain of QoE, typical stimuli have been shown to be waiting and response times as well as audio distortions, i.e. application-level QoS parameters directly perceivable by the end-user. Additionally, logarithmic relationships of the postulated form have not only been observed in the domains of psychophysics and perceived network performance, but also in the field of economics [181].

4.3.2 Exponential Relationships – The IQX Hypothesis.

The second example is the so called *IQX hypothesis* (exponential interdependency of Quality of Experience and Quality of Service) [86, 109] which describes QoE as an appropriately parametrised negative exponential function of a single QoS impairment factor. To demonstrate this mapping, iLBC-coded speech samples were sent over a network emulator that introduced packet losses. The resulting degraded samples were recorded and served, together with the original versions, as input to the PESQ algorithm (ITU-T Rec. P.862), which automatically calculates the corresponding QoE in terms of MOS values [107]. As a result the authors observed an exponential relationship of the form $MOS = \alpha \cdot e^{-\beta \cdot QoS} + \gamma$ between packet-loss and speech quality scores. The underlying assumption is that within a functional relationship between QoS and QoE, a change of QoE depends on the actual level of

QoE [86], implying the differential equation

$$\frac{\partial QoE}{\partial QoS} \sim -(QoE - \gamma). \quad (4.3)$$

which has an exponential solution.

Both types of relationships confirm the general observation that users are rather sensitive to impairments as long as the current quality level is already quite good, whereas changes in networking conditions have less impact when quality levels already are fairly low. However, they differ in terms of underlying assumptions: the WFL relates the magnitude of QoE change to the current QoS level, whereas the IQX hypothesis assumes that this magnitude of change depends on the actual QoE level. Furthermore, the WFL mostly applies when the QoS parameter equates to a signal- or application-level stimulus directly perceivable by the user (like latency or audio distortion), while the IQX applies in cases of QoS impairments on the network-level which are not directly perceivable (e.g. packet loss). However, this is only an observation from the related work and not based on empirical analysis. In order to prove this difference to be valid, one would have to translate QoS impairments not directly sensible to sensible impairments and then compare the two approaches in terms of prediction performance, which goes beyond this thesis. Therefore, this is only a side note to a further interesting question. Taken together, both relationships have been found helpful in explaining or obtaining new insights from passive measurements [192] and in the context of studying web applications and waiting times [74], [8]. Within the following section it is shown that QoS in IP networks can often be put on a level with waiting times and how this relates to QoE.

4.3.3 QoS equals Time for Browser Based Applications

The two foregoing sections have showed that fundamental relationships like the IQX hypothesis or the WFL are able to describe the relation between technical stimulus (QoS) and the resulting subjective experience (QoE). To achieve the aim of modelling QoE for browser based application based on fundamental relationships, it is important to identify the input stimulus. Foregoing discussions in previous chapters have already defined waiting times as dominant factor for the user experience. Nevertheless, other QoS parameters like packet losses or packet re-ordering are also well known to impact user perceived quality for IP-based services [63,103,176]. Therefore, this section explains why such other QoS parameters do not have to be considered

for the targeted application category.

Browser based applications have in common that their implementation utilises the HTTP protocol on the application layer, as well as the TCP protocol on transport and network layers, respectively. The HTTP protocol is based on a request-response pattern between server and client. Hereby, the term 'client' denotes an instantiation of the browser software at the end user device. For the end user, this request-response pattern results in a certain amount of waiting time. For example, when the end user requests a web page in her Internet browser, it takes some time until the web page is downloaded, rendered and displayed at the end user device. Below HTTP, the TCP protocol offers reliable transport of data between the client and server in general. In case of insufficient network resource issues, the protocol ensures that the end user receives all data requested by queuing strategies in case of insufficient throughput, packet re-transmission in case of packet losses or increased buffering in case of packet re-ordering [71]. The result is that those network impairments are perceived as waiting time by the end user (cf. [150]). Due to the increased use of HTTP in media delivery (e.g. audio and video streaming) also these kinds of services are nowadays affected by such temporal impairments in the form of rebuffering or stalling events (which are in turn waiting times again). In contrast, the occurrence of signal fidelity distortions as introduced by e.g. packet losses is declining. Therefore, waiting times are also a predominant impairment on the presentation layer for these kind of services, beside other temporal degradations such as varying video or audio quality as a result of adaptive streaming services [196]. However, as this is out of scope for this chapter the focus will be on waiting times for browser based applications. As the close relation between QoS and waiting time has been discussed above, the following section will analyse related work from psychology in terms of time perception and waiting times and their relation to user satisfaction.

4.3.4 Time Perception in Psychology

Work on human time perception covers a wide range of temporal perspectives on human behaviour (see [95] for a comprehensive review). This includes time estimation, perception of durations, the underlying timing systems in the human brain etc. The aim of this section is to outline certain characteristics of human time perception and carve out its parallels to quality perception in QoE. Furthermore, fundamental psychophysical relationships of human time perception are discussed, and it is shown

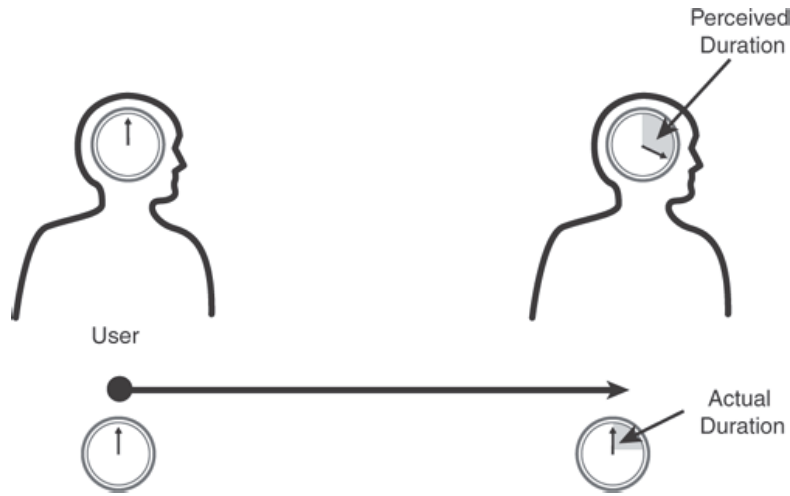


Figure 4.4: Perceived duration vs. objective duration from [191].

how they relate to fundamental relationships discussed in Section 4.3.1.

A recurring characteristic of human perception in general is the difference between the subjectively perceived nature of a signal or stimulus and its objectively measured value. Also in human time perception it is widely acknowledged that (subjective) perception of a duration should never be assumed to be accurate with respect to the actual duration. Whereas actual duration reflects objective time, perceived duration reflects subjective psychological time, which is susceptible to varying degrees of distortion. When users do gauge durations, they are more likely to rely on mental estimations rather than objective measurements [55, 95, 191] as depicted in Figure 4.4. The reasons for this mismatch between subjectively perceived duration of an event and its actual duration are manifold and a selection of influence factors is discussed below.

By its nature, time cannot be a direct stimulus, it is a certain duration between electrical stimuli signals of the nervous system. This requires the transformation from physical signals into a electrical signals in the nervous system via a sensory organ. Due to the different (temporal) properties of different sensory organs, the temporal resolution differs for stimuli of different modalities. E.g. auditory stimuli are more precisely processed on a temporal level compared to visual or tactile stimuli [94]. Regarding stimuli duration perception it has been found that auditory marked intervals are perceived longer than visually marked ones [91, 151]. In addition to the modality of the stimulus, its complexity has also a certain impact on the perceived duration, with highly complex stimuli being perceived longer.

Another characteristic of temporal stimuli is the effect that there are instances in which the second of two identical intervals is perceived as being much shorter than the first one, an effect known as the time-shrinking illusion [58]. These differences and characteristics of temporal stimuli have to be considered in the relation between waiting times and QoE and are especially interesting in the context of browser based applications. While for simple file downloads or low-complex web-sites the stimulus, or the duration perceived by the user until delivery is straightforward attributable, higher complex web-pages do inherit several of the above addressed characteristics, e.g. their loading behaviour constitutes a complex stimulus due to the number of different objects rendered in parallel after issuing a request for a new page. Also the interactive nature of web browsing and the numerous request-response patterns can be associated with the time-shrinking illusion as mentioned above.

In the context of interactive applications, system response times do not only contribute to the user's perceived quality of the system but also add to the felt interactive nature of the system. Regarding this perceived interactivity, three important limits for subjective response times (i.e. waiting times) are distinguished in [167], with response times < 0.1 s giving an instantaneous feeling of system reaction, for response times up to 1.0s the user's flow is not interrupted and for response times > 10 s user's attention is lost (cf. Section 2.2.2). However, user satisfaction or user perceived quality is not automatically linked with these times as there are also other influencing factors to be considered such as service or application used, related expectations etc.

For analysing user satisfaction based on perceived duration, [191] states that this is only meaningful when the perceived duration is compared to a tolerance threshold. If the perceived duration is shorter than the tolerance threshold, the user interprets that as fast and decent. Conversely, if the duration is perceived as longer than the tolerance threshold, the user interprets the duration as slow and insufficient. The value of this tolerance threshold is influenced by the context, personal factors, past experiences etc. (cf. [191]). Putting that into the QoE context it is obvious that this is congruent with the formation of subjectively perceived quality as described in [1]. An example for the context influence of the duration threshold reads as follows [191]: *A ten-minute wait for a person who is already 15 minutes late for an important meeting is excruciating. The same ten-minute wait for a person who has already waited three days for a package to arrive is trivial.* This also shows the importance of the relation between stimulus and stimulus change for user satisfaction with a service, and bridges to the principles of psychophysics and human perception

and its relation to QoE as described in Section 4.3.1 and Section 4.3.2. Therefore, the following section discusses psychophysical principles in human time perception.

4.3.5 Fundamental Relationships in Human Time Perception

Similar to the approaches for the relation between QoS and QoE discussed above, psychophysical principles in human time perception have been studied in [83]. The author identified a ratio between the magnitude of the time estimation errors and the duration of the sample length to be estimated, and attributed this finding to Steven's Power Law [195]. Successive work by [55] extended these results and added other models including the Weber-Fechner-law, while [93,139] set out to identify the minimal achievable error for time estimation based on the aforementioned models. They came to the conclusion that the relationship between estimation error and stimulus length is constant, which is essentially a version of Weber's law where the estimation error (termed *Weber Fraction*) is equivalent to the just noticeable difference already discussed in Section 4.3.1. Extension of these results to time related problems in other disciplines such as medicine [200] or consumer behavior research [56, 218] has proven that these logarithmic relations can be successfully transferred from psychological lab studies to real world problems. Of particular interest to our problem is the work of [56], which shows that for the subjective evaluation of waiting times on a linear scale a logarithmic relationship does apply.

These results prove that fundamental relationships of a logarithmic form, as discussed in Section 4.3.1, are also prevalent in human time perception. Therefore, the following section will postulate a related hypothesis in the context of browser based applications.

4.4 Verifying the WQL Hypothesis in Browser Based Applications⁵

As previous sections have shown that waiting time is the key determinant of QoE for browser based applications, and the aim of this chapter is the identification

⁵This section is based on original work from the author with adaptations as published in [5, 8, 29, 35] where the author actively contributed text, was involved in the test execution and performed the data analysis and modelling.

of a relationship between QoE and waiting time, the following “WQL hypothesis” quantifying the relationship between waiting time and QoE is postulated:

WQL: The relationship between **W**aiting time and its **Q**oE evaluation on a linear ACR scale is **L**ogarithmic.

This section sets out to verify this hypothesis with the data gathered in the subjective experiments described in Section 4.2. In a first step the WQL postulated above is applied to relatively simple scenarios of file downloads and simple web browsing in Section 4.4.1 and data from related work in Section 4.4.2. In a second step more natural and complex web browsing is analysed with respect to the WQL and related challenges with other upcoming influence factors in the complex browsing case in Section 4.4.3.

4.4.1 File Downloads and Simple Web Browsing

In order to prove the WQL hypothesis, a logarithmic fitting has to be applied to the MOS data of the respective results. This is achieved as follows: The mean opinion score (MOS), i.e. the average over the subjective ratings for the same test condition, is plotted depending on the pre-set waiting time t with markers, while the logarithmic curve fitting $QoE(t)$ according to the WQL is plotted as solid or dashed line. For the fitting, we use a logarithmic function with two parameters a and b which are derived by minimising the least square errors between the fitting function and the MOS values.

$$QoE(t) = \alpha \cdot \ln(t) + \beta \quad (4.4)$$

Figure 4.5 depicts the results of the file download tasks from studies A and B in which a 2.5 MB and a 10 MB file were downloaded by the users. In addition the results from the field trial (study C) are shown for 5.5 MB. The measurement studies were conducted in 2009, 2010 and 2011, respectively. It can be seen the same waiting time results in significantly different MOS scores depending on the file size. For example, a waiting time of 38s for the 2.5 MB files yields a MOS of 2.75 whereas the MOS of the 10 MB files was 3.58 for the same waiting time. This can be explained by the fact that the expectation dimension of QoE (cf. [79]) interferes here. If people do know that the file size is large, they have different expectations regarding the respective download time to expect. As this expected time is longer in case of the 10MB files compared to the 2.5MB files, the ratings for the 10MB

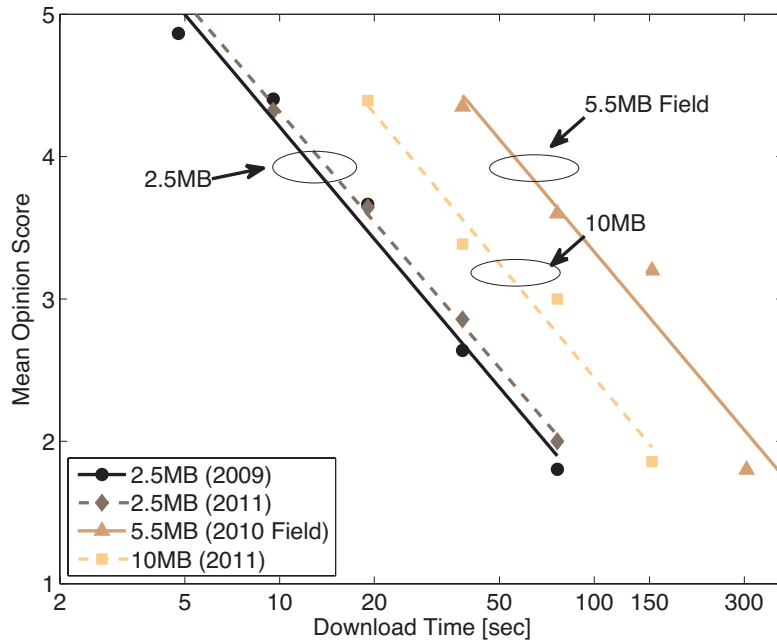


Figure 4.5: Download of files of various sizes obtained in three subjective user studies conducted in 2009 (study A), 2010 (study C) and in 2011 (study B), respectively (DL task).

files are better. A further discussion on expectations and their influence on waiting time evaluation can be found in [56]. Another influence of altered expectations, due to the context the user was situated in, is visible for the 5.5 MB slope of the field trial (study C). In this case, the users issued better ratings than users did for the 10 MB files and same download times in the lab environment. This can be explained by the user being more relaxed on download times in his normal environment and possible distraction side activities. Nevertheless, also in this example the logarithmic relationship holds true with $RMSE = 0.1506$, just the parameters a and b as used in Equation (4.4) are different. Details of the the logarithmic fitting and its goodness of fit in terms of coefficient of determination can be found in Table 4.4.

For the case of simple web-page views the stimulus for the WQL is not the waiting time until the file has downloaded but the page-load-time (PLT) which was manipulated for these tests. Figure 4.6 shows the result for manipulated page-load-times (PLT task). The subjects were asked to view several pictures or to perform certain Google search queries. In both cases the request for the next picture and the search result were delayed for a certain time, respectively. The user study for the 'picture load' task was repeated in study A and B. In addition, a 'photo' task has

file size	year	coeff. D	logarithmic fitting function
2.5 MB	2009	0.98	$QoE(t) = -1.14 \ln(t) + 6.83$
2.5 MB	2011	1.00	$QoE(t) = -1.12 \ln(t) + 6.89$
5.5 MB	2010	0.97	$QoE(t) = -1.14 \ln(t) + 8.58$
10 MB	2011	0.98	$QoE(t) = -1.68 \ln(t) + 9.61$

Table 4.4: DL task: Logarithmic fitting parameters and coefficient of determination (D) for download of files (see Figure 4.5).

been conducted which differs from the 'picture load' task in the technical realisation of the instrumented waiting time. For the 'picture load' (and the 'search') task, the HTTP requests were delayed, while for the 'photo' task the HTTP response instead of the HTTP request was delayed. However, this does not lead to observable differences from the end user's point of view. These results show that the ratings do coincide with the logarithmic fitting pretty well – except for the lowest load time $t = 0.18$ s for the 'picture load 1' task in Figure 4.6 (the data point marked with the red arrow). We explain this by the fact, that the two shortest time settings (0.18 s and 0.44 s) are already sufficiently convenient, such that the lower value does not lead to a far better waiting time evaluation. This means that QoE reaches saturation for small waiting times and that the WQL hypothesis only applies above the saturation point, i.e. for noticeable waiting times. This is in line with psychological time perception literature stating that waiting times below 0.5 s of waiting time are differently evaluated in term of user satisfaction [95]. Therefore, the parameters of the logarithmic curve fitting are derived without considering user ratings for waiting times below 0.5 s. Then, the RMSE is about $R = 0.0446$. All logarithmic fittings and goodness of fit values are reported in Table 4.5.

task	coeff. D	logarithmic fitting function
Picture Load 1	1.00	$QoE(t) = -0.80 \ln(t) + 3.77$
Picture Load 2	1.00	$QoE(t) = -0.63 \ln(t) + 3.58$
Search Task	0.98	$QoE(t) = -0.88 \ln(t) + 4.72$
Picture Load Task	0.99	$QoE(t) = -1.00 \ln(t) + 4.73$

Table 4.5: PLT task: Logarithmic fitting parameters and coefficient of determination (D) for loading times of pages (see Figure 4.6).

For all logarithmic fittings shown in this section it has to be noted that the maximum value that can be reached by mean opinion scores is $MOS = 5$, therefore

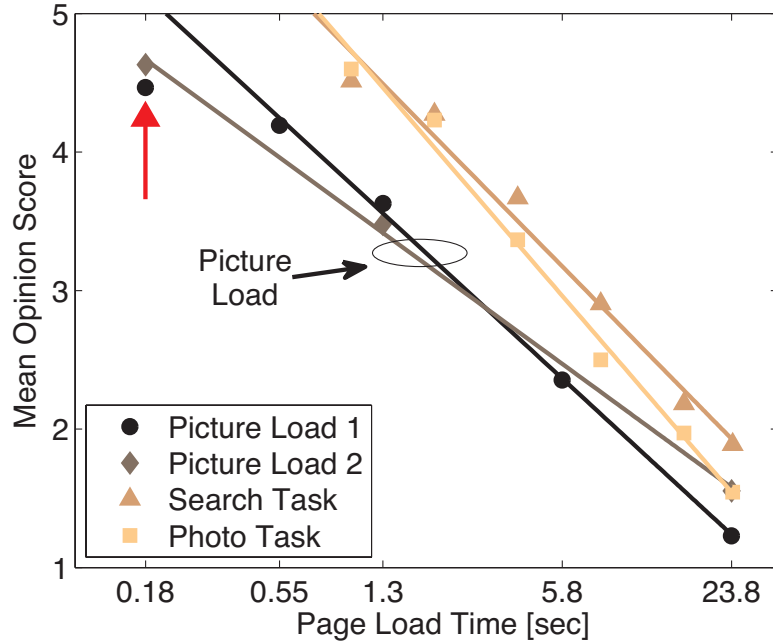


Figure 4.6: User satisfaction for various constant page-load-times (PLT task).

the predictions can only range between $MOS = 1$ as lower bound and $MOS = 5$ as upper bound.

4.4.2 Data from Related Work

Figure 4.7 shows results from related work reported in [169] where waiting times were used as stimuli for different applications and respective QoE ratings on an ACR scale were gathered.

Task	coeff. D	logarithmic fitting function
Web	1.00	$QoE(t) = -1.19 \ln(t) + 5.23$
Voice	0.99	$QoE(t) = -1.61 \ln(t) + 5.90$
E-Mail Text	0.99	$QoE(t) = -1.42 \ln(t) + 5.64$
E-Mail Attached	0.99	$QoE(t) = -1.27 \ln(t) + 6.03$
Download	1.00	$QoE(t) = -1.14 \ln(t) + 5.57$

Table 4.6: Logarithmic fitting parameters and coefficient of determination (D) for data from [169] as depicted in Figure 4.7.

The tasks used in this study were instrumented as follows:

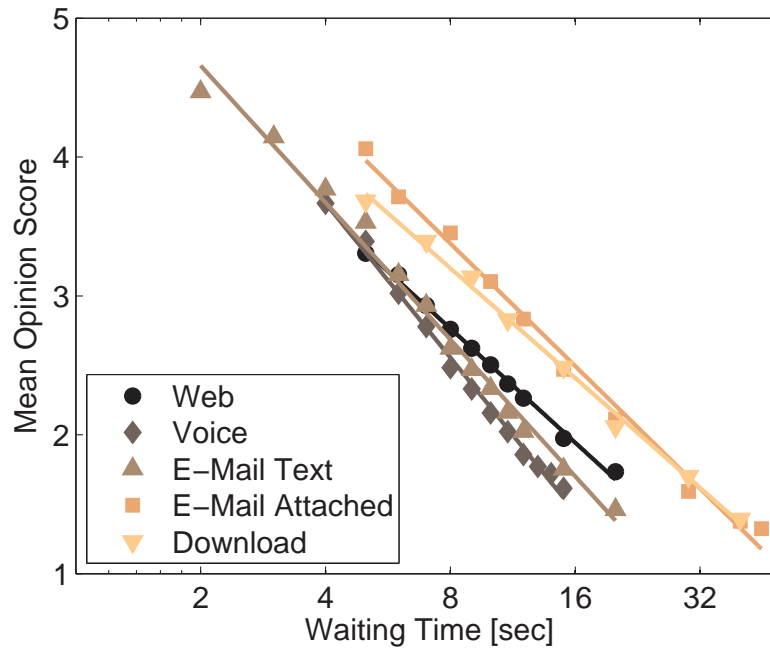


Figure 4.7: Results from [169] with logarithmic fittings applied as described in Table 4.6 supporting the WQL hypothesis, i.e. logarithmic relationship between MOS and waiting times, for several waiting time impaired services.

Web: Subjects were asked to press the start button for a web site access, end the time was manipulated until the top page was displayed on the screen. Waiting times were: 5, 6, 7, 8, 9, 10, 11, 12, 15, 20 [s].

Voice: In this task the call setup time (CST) was manipulated and the subjects were a posteriori asked for their perceived satisfaction with the CST. The CST's used were: 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 [s].

E-Mail Text / Attached: The subjects were asked to send a plain text E-Mail or an E-Mail with an attached file, and the time from pressing the send button until the E-Mail transmission was completed was manipulated. The instrumented waiting times were: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 15, 20 [s] for the plain text E-Mails and 5, 6, 8, 10, 12, 15, 20, 30, 40, 45 for the E-Mails with attachment, respectively.

Download: For this scenario the time from pressing the downlink button until the file download finished was instrumented. The used waiting times were: 5, 7, 9, 11, 15, 20, 30, 40 [s].

The download and web scenarios are comparable to our previously shown results, whereas voice and e-mail related waiting times were not used in our studies. In order to compare their results, logarithmic fittings adhering to Equation 4.4 have been included. Also these results can be closely approximated by the shown logarithmic fitting as described in Table 4.6 with a RMSE = 0.0669, hence verifying the WFL as well.

Summarising, our results show that the relationship between waiting time evaluation on a linear ACR scale and its respective waiting time can be predicted via the proposed logarithmic function, with goodness of fit values ranging from $D = 0.97$ to $D = 1.00$ and root mean square errors from RMSE = 0.1506 to RMSE = 0.0446 (for a scale ranging from 1 to 5), which represents very good prediction performance. Based on these results, the validity of the WQL hypothesis can be safely claimed.

4.4.3 Complex Web Browsing

After successfully applying WQL to file downloads and simple web browsing as shown in the preceding section, the aim in this section is to apply WQL to data from complex web browsing from study B (cf. Section 4.2.3). The difference to the application of the WQL in the preceding section is that the stimulus in case of complex web browsing is not the actual waiting time but the downlink bandwidth. The assumption was that the relationship between downlink bandwidth and websites of a certain size (the subjects were using the same websites for their tasks throughout the different test conditions) is linear across different downlink bandwidths, hence the stimulus would be related to the waiting time or page-load-time the subject experiences.

Figure 4.8 shows the acquired MOS and the corresponding logarithmic fitting in dependence of the downlink bandwidth. However, it can be seen that the logarithmic fitting does not match the MOS values very well with a goodness of fit of $D = 0.89$ and an RMSE = 0.2856. Also from a graphical impression it can be seen that the logarithmic fitting does not coincide well with the actual MOS (=measurement) values. With these results the WQL can not be verified. Recalling the conclusion from related work derived in Section 4.3.2, that an exponential fitting as postulated by the IQX hypothesis delivers better fitting results in case of not directly sensible stimuli (like the download bandwidth), such a fitting was applied to above results. This is shown by the dashed line in Figure 4.8. Obviously such a fitting does also perform pretty well on this data.

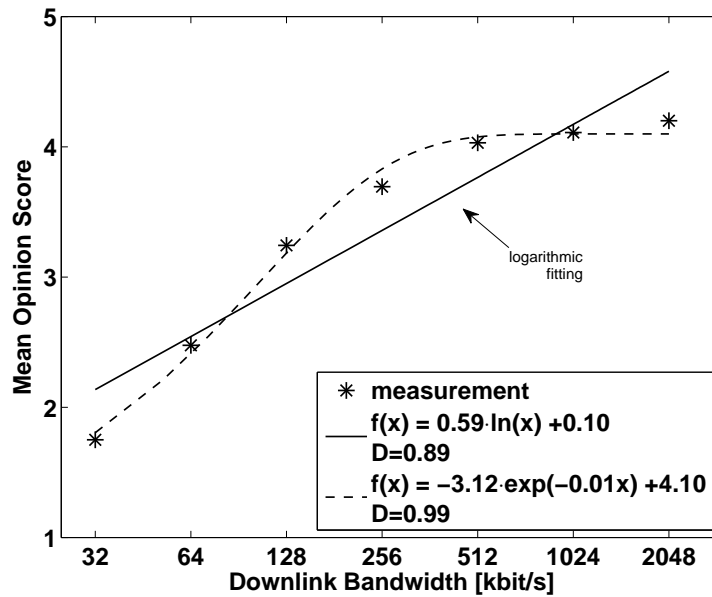


Figure 4.8: Web browsing with downlink bandwidth limitation instead of instrumented constant page-load-times.

In order to understand why the WQL was not holding for the above example we analysed a number of additional information gathered throughout the subjective test such as application level measurement of the page-load-times and network monitoring traces from passive monitoring. These results together with observations made throughout the subjective tests revealed a number of perceptual challenges and practical issues that come into play with more complex web browsing and are discussed in the following section.

4.4.4 Perceptual Challenges and Practical Issues for the Application of the WQL to Complex Web Browsing⁶.

Within this section several challenges are discussed which contribute to the complexity of (close to) real world web browsing and therefore interfere with the WQL relation between waiting time and QoE.

⁶This section is based on original work from the author with adaptations as published in [8, 29, 35, 51].

4.4.4.1 From Pages to Page Views and to Web Sessions

From a technical perspective, a web page is an HTML (Hyper Text Markup Language) text document with references to other objects embedded in it, such as images, scripts, etc. While HTTP (Hyper Text Transfer Protocol) constitutes the messaging protocol of the Web, the HTML describes the content and allows content providers to connect other web pages through hyperlinks. Typically, users access other pages or new data by clicking on links or submitting forms. Within this basic paradigm, each clicked link (or submitted form) results in loading a new web page in response to the respective HTTP request issued by the user, resulting in a new *page view* whose QoE is characterised by the time the new content takes to load and render in the browser. Furthermore, the surfing user typically clicks through several pages belonging to a certain web site and of course also occasionally changes sites as well. In this respect, a user's web *session* can be characterised by a series of page view events and the related timings of the stream of interactions as already shown in Figure 2.7 in Section 2.3.2.

As an example, Figure 4.9 shows the distribution (cumulative density functions) of PLTs measured on application-level during study B for four different web sites. In each condition, users browsed through the given website for approx 180s at a predefined downlink bandwidth. The results show that even at constant downlink bandwidths, the user experiences a wide range of PLTs that deviate from each other by a factor of 10 and more within one session of 180s. This holds true even for lean web sites accessed at fast network speeds (cf. Figure 4.9(c)). The main explanation for this phenomenon is that different page content structures, caching or other first-time effects (like DNS lookups) cause fluctuations of load patterns and thus waiting times, even when a client repeatedly accesses the same page. In addition it has to be mentioned that the content itself has an influence: depending on the actual weight and complexity of the web content, average PLTs for the same downlink bandwidth levels can differ considerably. However, end-users typically are aware of the "heaviness" of a website and tend to adjust their expectations accordingly.

Further, the subject's experiencing of differing PLT's throughout a web session results in a different QoE formation process on a psychological level. In web sessions the subject's QoE evaluation is not based on the experience of a momentary event (like it has been the case for the earlier discussed simple web browsing and download tasks) but rather on a retrospective rating of a series of events throughout a web session (also referred to as episode in [212]). Following [135], the retrospective rating

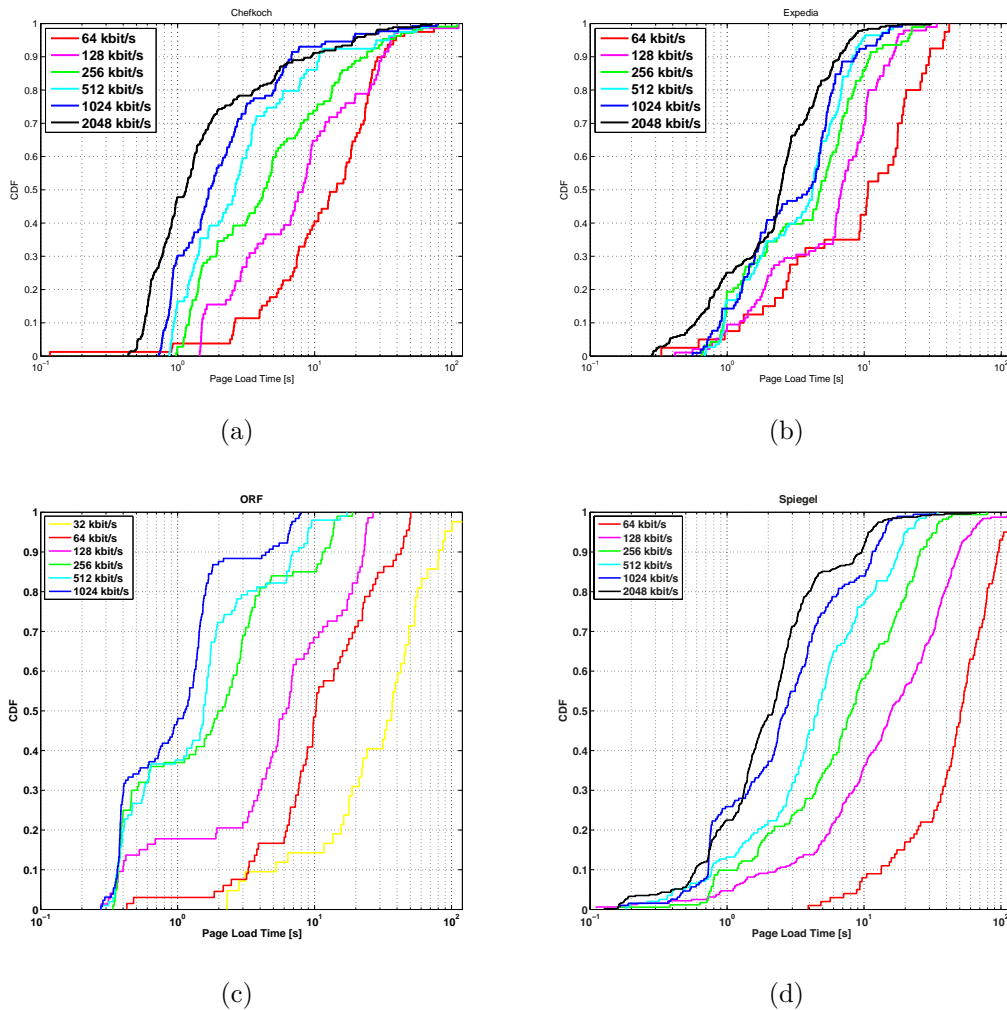


Figure 4.9: The cumulative distribution function of application-level page-load-times over one browsing session (approx. 180s) and several downlink bandwidths for four different websites [Chefkoch (a), Expedia (b), ORF (c) and Spiegel (d)].

of episodic QoE utilises a memory-based approach, resulting in a remembered experience. The formation of such remembered experiences is based on an integration process of momentary QoE values. This integration process involves several cognitive processes such as primacy, recency or memory effects [104, 201, 212]. Generally speaking, these effects describe the influence that momentary QoE events within an episode exert on the retrospective QoE rating (cf. [212]). For web sessions this means that (plain) averaging across PLT's and applying the logarithmic relationship to the averaged PLT does not consider these effects.

In [35] we have shown that, memory effects do affect QoE in web browsing contexts, where the perception of the current page-load-time is strongly biased by preceding page-load-times. E.g., if several preceding PLT's were better than the current PLT (which is only slightly higher), it might be rated considerably worse due to the influence of the memory effect. Such memory effects are also connected to flow-experiences where a single occurrence of a very unpleasant experience can lead to a particularly negative impression of the overall process. In the upcoming section the relevance of such flow experiences in the context of web browsing is discussed.

Time vs. Bandwidth. Another issue related with aforementioned application-level PLT's and their measurement is the assumption taken at the beginning of this study. We assumed that the relationship between downlink bandwidth and related download time (or PLT) is linear. However, this is not fully applicable due to the complexity and interactions of the HTTP and TCP protocols with the network performance (e.g. impact of high bandwidth-delay product on TCP performance, impact of TCP's slow start, congestion and flow control on loading times of small pages, HTTP pipelining etc., cf. [59]). Although this effect is not strongly pronounced for the bandwidths used in the studies (described in Section 4.2), it nevertheless represents a challenge to reliably calculate PLT's when downlink bandwidth is manipulated instead of directly manipulating application-level PLT. This is of particular interest when real websites are used for subjective tests as direct PLT manipulation is often not feasible.

Additional to this mapping problem, another technical problem has to be considered when downlink bandwidth is manipulated or PLT's are measured on a network level: Due to the necessary processing of the delivered content through the application, in terms of rendering, the network level content delivery time (or network-level PLT) can differ from the application-level PLT due to the web technology used⁷

⁷e.g. plain HTML vs. Javascript based rendering.

or the use of additional plugins, which is also depicted by the two different times t_{TPLT_1} (network-level) and t_{TPLT_2} (application level) in Figure 4.10. For displaying a requested web page to the user, in addition to the network page-load-time, the local machine rendering and displaying the web page requires a certain amount of time. Hence, the application-level page-load-time differs from the network PLT and may vary dramatically for different types of web pages, e.d. due to the actual implementation, the used plugins, etc.

4.4.4.2 From Request-Response to Flow Experience

While the former section was devoted to problems on the technical PLT of web pages, this section addresses the immersive experience dimension of web browsing. Examples given in Chapter 2.4 and reports from related work as [210] confirm that web browsing is a highly interactive activity. As a result of the interactions while web browsing, even new pages with plentiful information and many links tend to be often viewed only for a brief period. Thus, users do not perceive web browsing as sequence of single isolated page retrieval events but rather as an immersive *flow* experience (cf. [193]). In general, the flow state is characterised by positive emotions (enjoyment) and focused attention [76] and as a result, heightened human performance and engagement [209]. Hence, flow related experiences are perceived more positive. The notion of flow implies that the quality of the web browsing experience is determined by the timings of multiple page-view events that occur over a certain time frame, during which the user interacts with a website and forms a quality judgment. This is inline with comments from the preceding section where it was shown that web sessions consists of several page-view events. This has a dual influence on the relationship between waiting times and QoE: on the one hand, flow experiences cause users to 'lose their sense of time', resulting in distorted time perception [76] in the way that people underestimate waiting times. On the other hand, a sudden instance of overly long waiting times (and thereby interrupting the flow) tends to be perceived particularly negatively [193,215], which links again to the memory effects discussed in the preceding section.

4.4.4.3 Perceived vs. Application-Level Page-Load-Time

In Section 4.4.4.1 it has already been shown that the objectively measured page-load-time on an application level fluctuates strongly for constant downlink bandwidths (cf. Figure 4.9). However, this is not only the case on a technical level (=application

level PLT). Even if the application-level PLT would be constant throughout one web session the perceived page-load-time might still differ between the page views within the session due to the following reasons:

First, page elements are typically displayed progressively (by the rendering engine) before the page has been fully loaded (cf. t_{PRs} has already taken place although t_{TPLT_1} or t_{TPLT_2} are not yet reached in Figure 4.10), thus the user's information processing activity overlaps with the page load phase, resulting in a rather complex stimulus. As a consequence, the user's perception of waiting time and latencies becomes blurred by the rendering process itself (which in turn is strongly influenced by page design and programming) [54, 140, 193]. Second, on a perceptual level, the duration from request submission until the rendering of the new page starts, i.e. when the user receives the first visual sign of progress [77, 171] (cf. t_{PRs} in Figure 4.10) is another relevant factor for the perceived speed of information delivery that is not covered by the application-level PLT. And third, a page might appear to the end-user to be already loaded (t_{PPLT}) although page content is still being retrieved (until t_{TPLT_1} is reached), due to the progressive rendering of the browser, asynchronous content loading (AJAX) and the fact that pages are often larger than the browser window itself.

In order to better understand the relation between subjectively perceived events and measurable application-level events, Figure 4.10 depicts these events which are related with above mentioned concept of perceived PLT. The upper timeline in black describes perceptual events from an end-user point of view, whereas the lower timeline in blue describes technical events on application or network level. The different events, using the same color scheme as in Figure 4.10, are defined as follows:

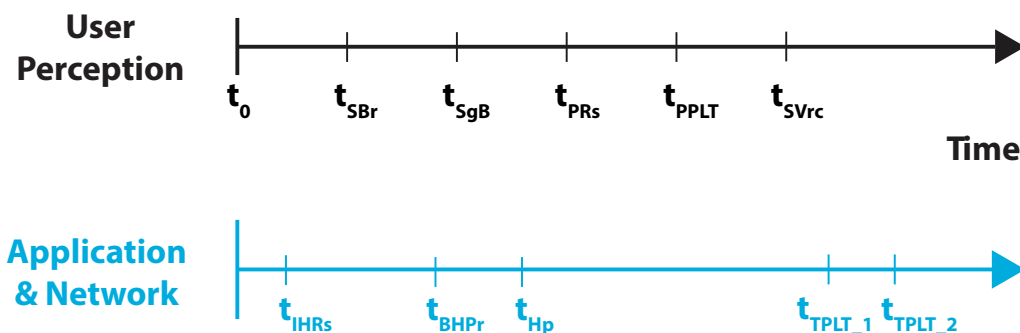


Figure 4.10: Perceptual events in a web page view cycle from the end-user point of view. The lower timeline (blue) displays related technical events on application or network level.

t_0 : Is the moment in time when the user requests a new web page (typically by clicking or pressing enter after having entered the URL of the web page in the browser's address bar).

t_{IHRs} : The moment in time when the initial HTTP request is sent by the browser.

t_{SBr} : The moment in time when a change in the status bar happens (usually a progress bar becomes visible at this moment).

t_{BHP_r} : The moment in time when the first HTML <head> element is received.

t_{SgB} : The moment in time when the previously viewed web page vanishes and the content of the requested page has not yet started to render.

t_{Hp} : The moment in time when the HTML page is processed by the browser (Can only be observed on the application level).

t_{PRs} : the moment in time when the first element of the requested page appears on the screen, independent of the type of element.

t_{PPLT} : The moment in time when from the point of view of the user the page is sufficiently rendered such that he can access the information he is seeking for.

t_{VSrc} : The moment of time where the visible portion of the web page (as determined by screen or browser windows size) is fully rendered.

t_{TPLT_1} : The moment of time when all objects of the page are downloaded from the server at the browser's device.

t_{TPLT_2} : The moment of time when the page is completely rendered and displayed by the browser.

Together, all of the aforementioned three factors contribute to the divergence of perceived PLT (t_{PPLT} in Figure 4.10) and application-level or technical PLT (t_{TPLT_1} or t_{TPLT_2} in Figure 4.10). In order to prove the perceived page-load-time concept and analyse its consistency across different subjects and to identify the (potential) relationship between t_{PPLT} and t_{TPLT_1} , a dedicated study was set up with the participants from study B. In this study the subjects were asked to mark the point in time, by pressing a dedicated button, when they considered a page to be loaded, i.e. the subjectively perceived PLT was reached. Figure 4.11 shows the results with the application-level PLT in yellow and the subjectively perceived PLT in cyan for different website types (and three different pages within each type, e.g. front page, search results and article detail page for Amazon). It can be seen that there are large differences between technical and perceived PLT time, with ratios $\frac{t_{TPLT_1}}{t_{PPLT}}$ ranging

from 1.3 up to 3 (where 1 would be the exact match between subjectively perceived and application level PLT). In addition it can be seen that the confidence intervals are pretty small, which can be explained by the fact that the subjects were very consistent in their judgment on perceived page-load-time for the used web pages.

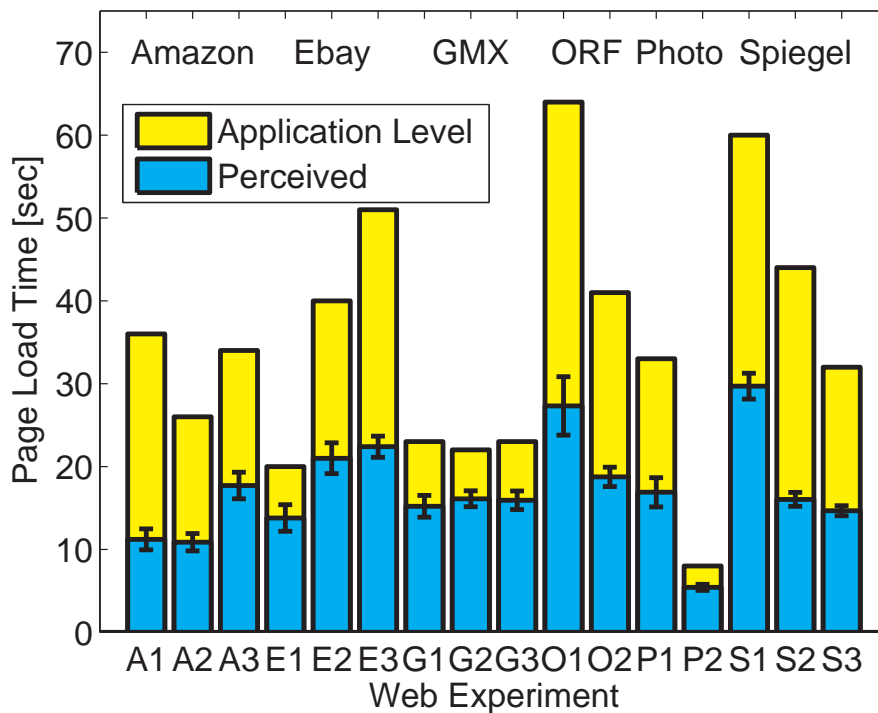


Figure 4.11: Perceived subjective vs. application-level PLT for different pages. The bar labels on the x-axes can be interpreted as follows: *A1* stands for "Amazon" and "Page 1" (i.e. Amazon landing page), *A2* stands for "Amazon" and "Page 2" (i.e. Amazon article page), *E1* stands for "Ebay" and "Page 1" (i.e. ebay landing page) etc.

Revisiting the reflections in Section 2.2 regarding the interactional process and related interactional metrics, it can be concluded that the perceived page-load-time does represent an interaction related measure. Although it is not distorted by transmission delays (as e.g. the unintended interruptions as shown in Section 3.1.4), it is strongly related to the interactive nature of web browsing as the (intrinsic or given) task execution, and hence the resulting interaction process, is strongly determined by the way the visited web pages behave and how they present information (that the user is searching for), and to a lesser extent by the respective (application-level) page-load-times. In terms of relation to QoE, the results reported in [78] have shown that subjects do rate conditions of identical application-level PLT's for a given

task considerably better if the task-relevant element appears earlier. In their study subjects had to find certain information via navigating through four web pages of identical application-level PLT's. On each of these web pages one clickable element was task-relevant. The appearance of this element was varied within the page load process between appearing in the middle of the application-level PLT or at the end. The better results for conditions where the relevant element appeared earlier can be interpreted such that the subjects were perceiving shorter page-load-times for these cases.

Summarising, all these different aspects on the perceptual and technical level lead to practical issues and challenges to measure or estimate the waiting time as input for the WQL (even if the WQL hypothesis would be valid for web browsing too). However, through analysis of the results, concepts from related work, and an additional study, the subjectively perceived page-load-time could be identified as an interactional metric to be used as potential input for QoE modelling approaches.

4.5 Conclusion and Lessons Learned

4.5.1 Novel QoE Assessment Methodology for Web Browsing

Based on the identified requirements and challenges for realistic web browsing, a novel QoE assessment methodology has been proposed. This new methodology establishes an interaction process between the test subject and the website under test leading to a web sessions that are comparable to interaction patterns in real world web browsing.

Data acquired within two lab studies that utilised the proposed assessment methodology for file downloads and web browsing have been compared to data from a related field trial. The results prove that the data acquired with this methodology is capable to deliver reliable and externally valid results for both services across different contexts. By comparing rating data from a lab study with rating data acquired in a field trial it was shown that the novel assessment methodology is able to deliver reliable and externally valid results for web browsing across different contexts.

4.5.2 QoE Modelling for Browser Based Applications

In a first step waiting times have been identified as the key influence parameter in browser based applications working on TCP/IP networks. Furthermore, fundamental relationships from psychophysics and their application to QoE modelling in related work have been reviewed. Together with results from related work on human time perception from psychology domains that has shown that Weber-Fechner's law does actually apply to a number of human time perception problems, this has led to the formulation of the WQL hypothesis, which quantifies the relationship between waiting time and QoE as:

WQL: The relationship between **W**aiting time and its **Q**oE evaluation on a linear ACR scale is **L**ogarithmic.

With data from the preceding studies in lab and field contexts, the WQL hypothesis could be verified for pure waiting tasks that are typical for simple web usage scenarios (e.g. file downloading, picture downloads, simple search queries). In a second step, the WQL was applied to the more complex case of interactive web browsing, which revealed that the hypothesis does not hold true for this application as practical issues and challenges prevent proper estimation of the subjective waiting time as input for the WQL. Hence, research question RQ5 can only answered partially.

In addition, the subjectively perceived page-load-time was identified as an interactional measure that provides information of the relevant waiting time users connect with their (intrinsic or given) task on a web page. The consistency of this perceived PLT across different subjects could be shown through results from a dedicated study. Therefore, the perceived PLT is a promising interactional metric that can serve as input for further modelling attempts in the context of web browsing.

4.5.3 Challenges and Practical Issues for Modelling QoE Based on Waiting Times for Complex web browsing

The failed verification of the WQL for complex web browsing has shown that identification of subjectively perceived waiting time is not straightforward, and is rooted in several challenges and practical issues on a perceptual and technical level which have been identified as follows:

Perceptual level: On a perceptual level three different challenges appear: First, throughout a web session the user's experience is characterised by a series of page view events rather than single page views and respectively, a series of different waiting times rather than one single waiting time. Second, web browsing is an immersive flow experience, that causes the users to 'lose their sense of time' and different page rendering throughout a session blurs the users sense of 'real waiting time'. And third, the subjectively perceived page-load-time and the application level measured page-load-time deviate strongly. Summarising, all these different aspects on the perceptual level lead to practical issues and challenges to measure or estimate the waiting time as input for the WQL

Technical level: When real websites are used for subjective testing directly manipulating of the application-level page-load-time is not feasible, therefore manipulating the downlink bandwidth can be used as indirect route to manipulate the PLT (when file sizes of the web pages are known). However, this leads to two practical issues: First, the relationship between downlink bandwidth and resulting download time is not linear. Hence, measuring throughput and extracting waiting time from the throughput measurement and a given file size cannot be applied. Second, even if an object is delivered on a network level in a certain (network-level load) time additional processing steps have to be passed until the object is rendered to the user on the application level. Therefore, network-level load time and application-level load time are not identical and further complicate deriving the proper waiting time as input to the logarithmic relationship described by the WQL.

Chapter 5

QoE Formation for Interactive Internet Applications¹

As already discussed within Section 2.1, QoE as it is defined in [1] is a pluridimensional concept that puts the experiencing subject into the focus. QoE related research is therefore concerned with measuring or assessing the quality as experienced or perceived by a subject, identifying factors that influence this perception, and finally developing models that are able to predict or estimate QoE. In order to accomplish this in a proper way, it is essential to understand how the quality formation process within the human subject is organised. There have been several attempts to describe the quality formation process, like [132] and [176], which formed the basis of the quality formation model in its most recent versions as presented in [1] and [9].

By reviewing the quality formation model of [1] in the following section, it is shown that this model does not properly consider the interaction process. It considers only a static input signal but not the request-response cycle which is inherent to interactive applications, as shown in Section 2.2. In a second step, a taxonomy is discussed that considers interaction performance aspects and their relation to quality influence factors and quality aspects. Then, a perceptual model is proposed that shows how such interaction performance aspects, arising from a request-response cycle between two interacting entities, can be identified. Finally, this perception model is used to update the QoE formation model of [1] to include the interaction

¹This section is based on original work from the author with adaptations as published in [1] where he was the lead author for section 2 contributing large portions of the text and drawing the quality formation process figure, [49] where he was acting as the lead author of the chapter and [9] where he was acting as co-author contributing substantial parts of the text and figures therein.

process and related perceived interactivity features in the quality formation process.

5.1 Quality Formation Process for Static Media Experiences

In order to understand how experience quality or QoE emerges within human subjects, it is necessary to understand how the (media) input signal is processed into a QoE score. Based on psychological and neurological knowledge, the perception of the input signal by the human sensory system and its further processing by higher cognitive processes into a perceived quality and a respective QoE rating is explained.

Figure 5.1 depicts the elements involved in the quality formation process as described in [1]. Boxes denote external inputs to the process, circles represent perceptual processes and two parallel lines represent storages for different types of representations. This process consists of two paths: a perception path and a reference path.

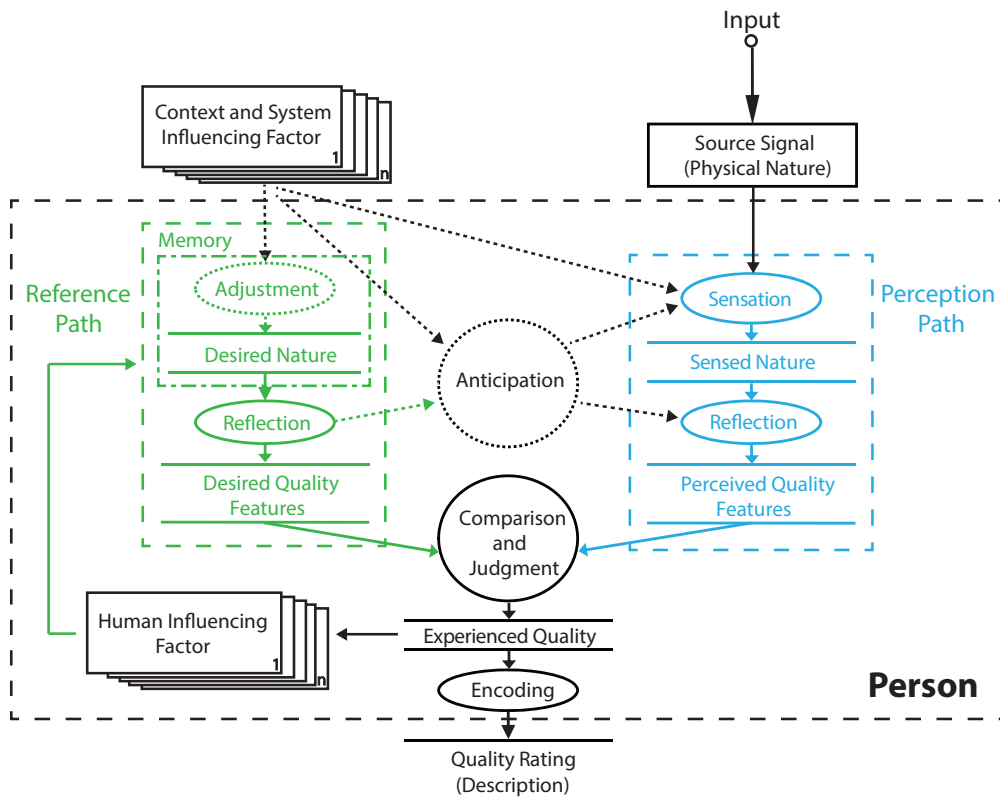


Figure 5.1: Quality formation process as depicted in [1].

The reference path (green, left hand side in Figure 5.1) reflects the temporal and contextual nature of the quality formation process and also inherits a *memory* of former experienced qualities, indicated by the dot-dashed box in the reference path. In this *memory*, perceptual references are stored at different levels: sensory memory, short term memory and long term memory, with all of them adhering to different time constants of information retention time. E.g., the short term memory might provide perceptual references of the primary stimuli in a double stimulus quality evaluation (cf. [114]). The reference path is influenced by two inputs: 1) the experienced quality via the *human influencing factors*, that assigns certain attributes to the experienced quality, which is then stored in the memory, and 2) the *context and system influencing factors*. The latter input influences the adjustment process which selects an appropriate perceptual reference for the given situation which leads to a *desired nature*. In the following *reflection* process respective *desired quality features* of this *desired nature* are identified.

The quality perception path (blue, right hand side in Figure 5.1) takes a physical event, triggered e.g. by a physical signal reaching our sensory organs, as an input. This physical event is converted into neural representations that include characteristic electric signals. These neural representations are transmitted to the brain via neural transmission. Throughout this transmission they are further converted into symbolic representations by the neural system (cf. [57]). At this level, internal references and rules are created in the reference path and linked through the *anticipation* process in the human brain to the neural representations of the input signal, which leads to the formation of the *sensed nature*. In Figure 5.1 these processes are summarised in the *sensation* process. Further, this *sensation* process can be also directly influenced by *contextual and system factors*. In such a case certain signal processing features of the sensory organs are activated in order to react to (probable) emergency signals. The following *reflection* of the *sensed nature* is linked with the identification of emotional, sensory or conceptual quality features of the experience, and is additionally influenced by the reference path through the *anticipation* process. The outcome of the *reflection* process are *perceived quality features* of the input stimulus.

Finally, the *desired quality features* resulting from the reference path and the *perceived quality features* originating from the quality perception path are then translated into the experienced quality on behalf of the *comparison and judgment* process (cf. [1]). If a quality rating is demanded from the person, then the *experienced quality* has to be described. This is achieved by the *encoding* process, which assigns

a certain code, in the form of a verbal description or a numerical descriptor, to the experienced quality, leading to a *quality rating*. In terms of external influences, context, system and human factors are considered².

However, this quality formation process is targeted rather towards media experiences on a single and static (in the sense of interactivity) input signal and does not consider actions by the experiencing person nor does it consider interaction performance aspects. As a result, this approach does not account for recurring (inter-)actions between two or more entities³, which happens in the interactive request-response cycle as described in Section 2.2 and its related signals. In the context of the above described QoE framework, this can not be overcome by multiple (subsequent) iteration steps of the quality formation process. Such multiple iterations would consider multiple signals, but still neglect the interaction process between involved entities (as only one entity is considered), and the deterioration of interaction cues due to system properties. This emphasises the novel challenge of including interaction related influence factors into QoE frameworks.

5.2 Interaction Performance Aspects

An approach to overcome the lack of missing interaction related aspects is outlined in the taxonomy proposed by [161, 164]. It incorporates the influence of the interaction process on the overall quality formation process by introducing an additional layer of interaction performance aspects which acts as mediation layer between quality influence factors and perceived quality features. Each of these layers spans over several stages of the quality formation process, therefore the relationships between these layers are not one to one and can vary in their strength, depending on the system, user, or context (cf. [164]). Naturally, such a mediation layer does of course not fully integrate (inter-)action between entities into the quality formation process, however it is a simple and efficient way to consider the influence interactivity exerts on QoE.

These interaction performance aspects result from the process of interaction between two or more entities and their perception of this process on several dimensions as depicted in Figure 5.2 and are described as follows:

²A detailed discussion of influencing factors related to QoE can be found in [79]

³Thereby running several times through the respective perception and judgement processes

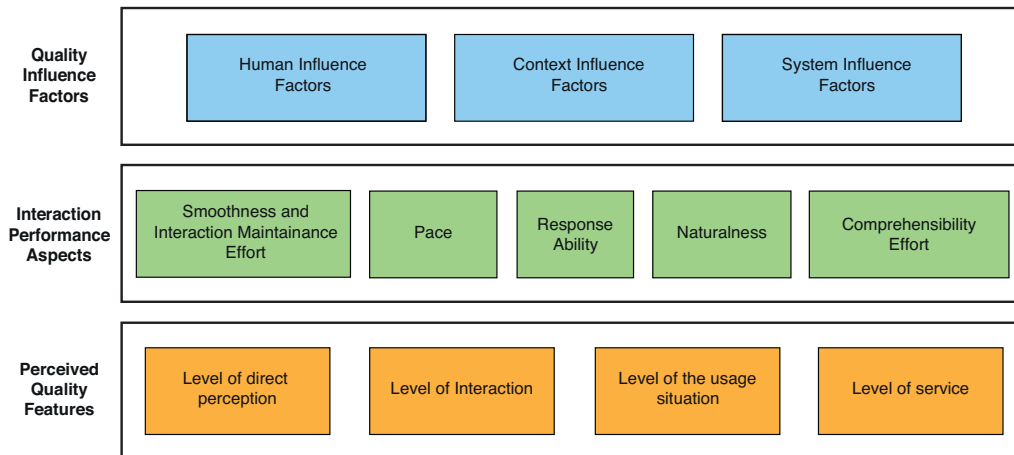


Figure 5.2: Taxonomy of influence factors, interaction performance aspects and quality features, from [164] adapted with terminology from [1].

Smoothness and Interaction Maintenance Effort: describes how fluent and effortless the users experience the conversational flow. If normal interaction behaviour has to be adapted as a result of bad system performance in order to maintain the ongoing interaction as smooth as possible, the interaction will usually also be perceived as being less smooth. Typically, an interaction has an inherent pace it establishes, thereby keeping the maintenance efforts of the interaction parties minimal. However, due to system impairments, the interaction pace can be changed, thereby accordingly demanding additional user effort in order to adapt to the changed pace. For H2M interaction this can severely impact the *flow experience* or the experienced smoothness, whereas for H2H interaction the conversational rhythm can be impaired (cf. [52]).

Pace : is the users perceived promptness of the interactional acts and respective actions by the other entity.

Response Ability: denotes if it is possible to issue a response following a prior message (or request) from the system or the other user. Response abilities based on interruptions in H2H interactions can be severely obstructed by transmission delays, as interruptions may not arrive in time and are not able to interrupt in the way originally intended. In terms of browser based applications the necessary information can already be rendered to the user, but the element relevant for issuing the response (e.g. clickable link or a forward button) might not be rendered yet. Hence, the response can not yet be issued

in this case.

Naturalness: is related to the inherent knowledge about how an interaction takes place in a non-mediated or ideal case.

Comprehension Effort: is required to understand either the other interlocutor (in case of H2H interaction), or is needed to interpret the response from the machine. Comprehension can be distorted by e.g. double talk or not rendered portions of the webpage which might be needed for navigation or information retrieval.

It has to be noted that the above aspects can not be seen as disjunct factors, hence overlaps of the different concepts are possible. E.g., naturalness coincides with smoothness, or response ability can be interrelated with comprehensibility effort.

In terms of quality formation, the output from this interaction performance aspects layer is further translated into interaction quality features and constitutes then an additional input to the comparison and judgement stage (cf. Figure 2.3 in Chapter 2), where these interaction quality features are further processed in conjunction with the other (more media related) quality features. Metrics for measuring these interaction performance aspects have been derived for Internet telephony in Chapter 3 with the I³R, the UIR and the $\frac{I^3R}{AIR}$ ratio. In terms of perceived quality features, the perceived page-load-time has been identified as a key quality feature, not contained in current web QoE prediction models, that is strongly influenced by the interaction process in Chapter 4.

The information provided by the *interaction performance aspects* layer is an addition to the perceived (static) quality features considered in Figure 5.1. What is still left open is the question how these interaction performance aspects can be derived or perceived in an interaction process between two or more entities. Therefore, the following section proposes a perception model that outputs such interaction performance aspects.

5.3 Perception Model for Interacting Entities

The perception model proposed in this section incorporates (inter-)actions between two (or more) interacting entities and is depicted in Figure 5.3. It achieves that by adding an additional output to the perception model and thereby considers responses of one entity as a reaction to a request by the other entity. The relation

between the input and output as well as the derivation of perceived interaction performance aspects is described as follows: In the first, *sensory processing* step, the input signal is automatically processed and converted into a neural representation, which is further processed in the *perceptual event formation* process. This process is already influenced by remembered perceptual events (reflected in the person's state in Figure 5.3) and combines different modalities into perceived events such as utterances, interruptions or other interaction cues. At this stage also the *sensed nature* is available as output (for further processing into perceived interaction features by subsequent processing steps cf. Figure 5.4). It includes not only information of the input stimulus but also information from the interaction process itself. This information stems either from the *perceptual event formation* process or from the *anticipation* process, whereas in the case of the latter also the *higher level cognitive processes* can participate in the formation of interactional information.

The following stage of *anticipation* connects information from different processes and storages, and then decides if certain (*inter-*)actions should be performed (as reaction to the input signal). This decision can be twofold: the presence of certain stimuli (or perceived events) may lead to a direct and unconscious⁴ (*inter-*)action or the decision is based on further processing by *higher level cognitive processes*. This differentiation is also related to the different processing of temporal stimuli as discussed in Section 2.2.2. When *higher level cognitive processes* are included in the processing of the input signal, e.g. extraction of semantic information that was searched for by the user, then conscious processing, and hence the interval timing system, applies. Contrary, when unconscious processing (within the *anticipation* process) causes a reaction then the millisecond timing system applies. The *person's state* includes the physiological state of the person as well as its cognitive state, whereas *assumptions* refer to the person's attitude and concepts (which can be influenced by e.g. the given or intrinsic task).

This perception model explains the formation of action or interaction between interacting entities, which then can be measured. Such (inter)actions between two or more entities form the basis of the *request-response cycle* introduced in Figure 2.3 in Section 2.2.1, and depicted in blue in Figure 5.3. Within such a request-response cycle several (inter)actions mutually take place between the interactants, thereby running several times through each perception and reaction process.

In the setting with two interacting human entities, as depicted in Figure 5.3, the

⁴Based on the discussion on human perception in [90], processing on this stage is still accomplished by unconscious brain processes

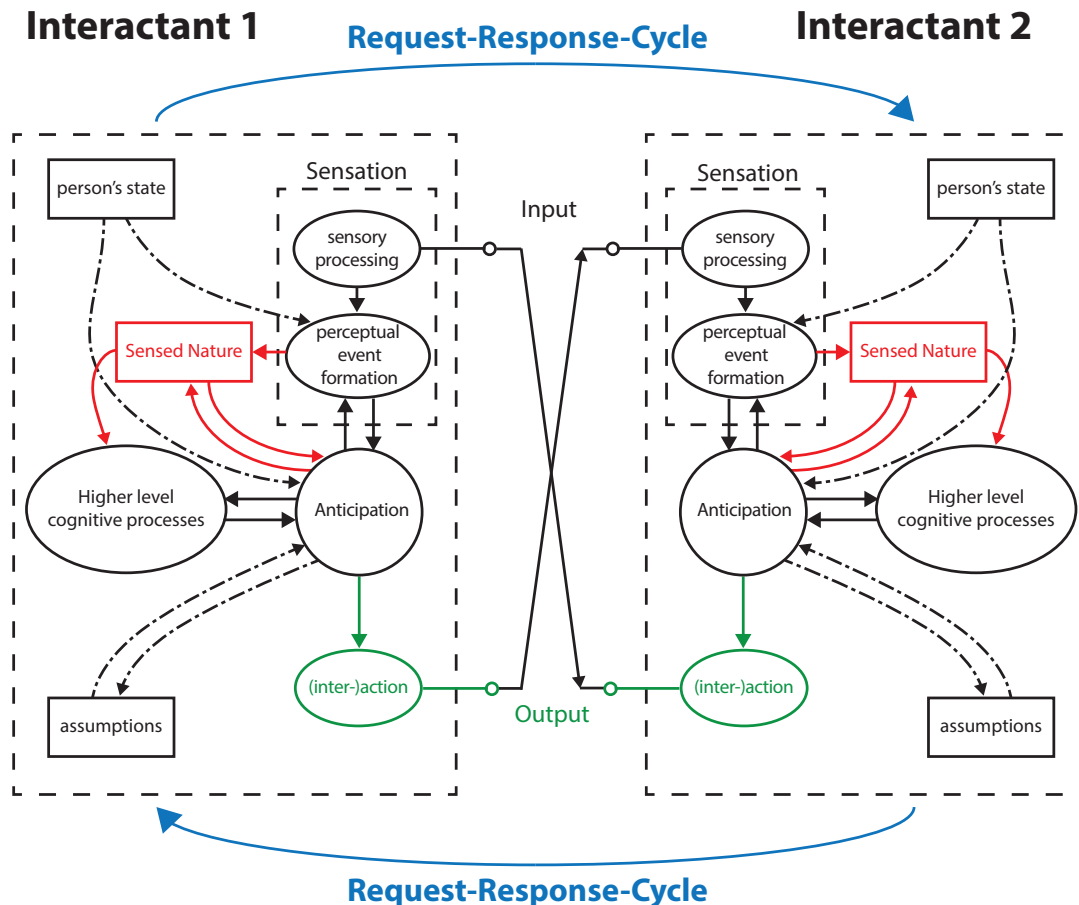


Figure 5.3: Perception model that allows to capture interaction performance aspects from an interaction between two or more entities based on [9]. Circles represent perceptual processes and boxes represent storages for different types of representations. Note that continuous lines represent direct input to (or output of) the perceptual processes and the control of (inter-)actions, whereas the dashed lines are influences on the respective processes from higher level cognitive constructs. The person's state refers to both the cognitive as well as the physiological, current state of the person. In turn, assumptions here refer to the person's attitude and concepts.

reciprocal (inter) action as result of a certain input stimulus is obvious. However, the perception model also holds true for cases where only one human entity interacts with non human entities, such as a computer in computer gaming or speech dialogue systems. In these cases, machine reactions on certain user (inter)actions are determined by the underlying algorithms, whereas user (inter)actions are formed according to the same processes as in the human to human interaction setting. An integration into the quality formation model described in Section 5.1 is performed in the following section.

5.4 Quality Formation Process for Interactive Media Experiences

The updated quality formation process is depicted in Figure 5.4 and integrates the perception model proposed in the previous section. For better illustration the integrated parts are marked in red. Slight adaptations from the proposed model as described in Figure 5.3 are: The higher level cognitive processing is contained in the dashed box in the reference path (left hand side Figure 5.4), the person's state and assumptions are reflected in the *human influencing factors*.

Similar to the process described in Figure 5.1, the first sensory processing step of *sensation* converts the input signal into a neural representation. This process is already influenced by remembered perceptual references from the memory in the reference path through the *anticipation* process. Additionally, it combines different modalities into perceived events such as utterances, interruptions or other interaction cues. As outcome of this process the *sensed nature* is available, based not only the input stimulus but also based on information from the interaction process itself. This information from the interaction process arises from running through several interactions within a request response cycle and is stored in the working memory (reference path). Via the *anticipation* process it is fed back, indicated by the red arrow to the *sensed nature* storage. The *sensed nature* is then further processed into perceived interaction features as part of the *perceived quality features* storage by the subsequent reflection process.

In terms of (inter)actions of each interactant the stage of *anticipation* connects information from different processes and storages, and then decides if certain (inter)actions should be performed (as reaction to the input signal). In addition, the presence of certain stimuli (or perceived events) may lead to a direct and uncon-

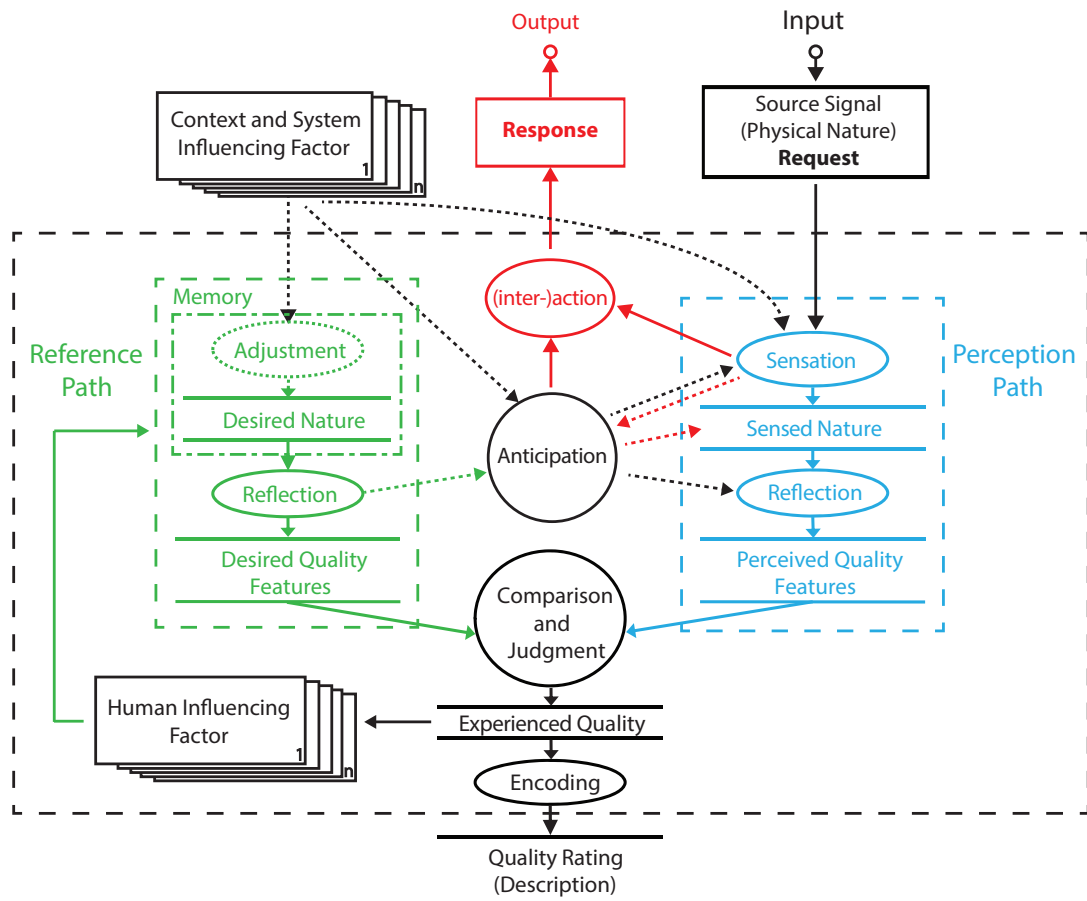


Figure 5.4: Integration of the proposed perception model for interacting entities into the quality formation process as described in [1], with changes due to the proposed perceptual model for interacting entities as introduced in Section 5.3, in red.

scious triggering of *(inter-)action*, indicated by the red arrow from *sensation* to *(inter)action*.

This updated model of the quality formation process now fully integrates the interaction process between two or more entities (only one depicted in Figure 5.4), which can be extended easily by adding more entities and connecting their inputs and outputs accordingly. Also interaction performance aspects are reflected as they can exert influence on the *sensed nature*, which is then further processed through the reflection process (perception path, right hand side in Figure 5.4) into perceived interactivity features (as a sub-feature of the perceived quality features). In the context of this thesis, this updated model can now be used to describe QoE formation for the two prototypical services.

For quality formation in human-to-human conversations, conversational (inter-)actions of both interactants can be described by using two instances of the model. In terms of the E-model modification proposed in Section 3.3 the additional inputs sT and mT, which are derived from $mean(I^3R)$, can be related to the perceived quality features in Figure 5.4. Two human interactants connected through a delay affected VoIP system will, after a while, sense that they are not able interrupt the other person. Hence, their internal quality formation process will sense low interaction performance in terms of their response abilities. As a result the perceived interaction features will deviate strongly from the desired interaction features as stored in the reference path and lead to bad experienced quality.

Regarding browser based applications, it allows to describe the interaction between a human user, which can be represented by the updated model, and a machine in form of an interactive website or a download portal. When the user requests a certain page, the page starts to load and the counter in the higher level cognitive processing within the user is started. Next, the elements that appear on the screen (as response to the request) are processed, by the quality formation process, into certain representations within the user's brain. When the perceived nature of one of these representations matches the desired element (=nature) the user was searching for (e.g. a link to further navigate on the page), the user takes (inter)action and clicks this link. Now the running counter is stopped and its content represents the subjectively perceived page load time. Finally, this time is then compared to the expected load time (desired quality feature) and the outcome of this comparison is a certain experienced quality.

Chapter 6

Conclusions and Future Work

Quality-of-experience (QoE) of interactive applications transmitted over TCP/IP networks has recently gained considerable attention, and is influenced by transmission delays due to TCP/IP's retransmission characteristic. This network-induced delays (and respective waiting times) are particularly critical for interactive Internet applications. Interactive applications typically establish a request-response cycle between two or more entities. In case of transmission delays, this interactive process is deteriorated. The impact of the temporal impairment depends on the application type (e.g. Internet telephony or browser based applications) and the interactivity level of the application. Recent QoE concepts and related models fall short in considering the influence of the impaired interaction process, as they are targeted towards signal fidelity of static media signals as main factor determining QoE. Therefore, appropriate QoE assessment methodologies considering interactivity related impairments and their measurement are needed. Datasets acquired through such methodologies could then be used to derive QoE models that incorporate interactivity metrics and thereby enhance their prediction performance.

In this thesis, the impact of transmission delays on interactive Internet applications has been analysed. As a starting point, Chapter 2.4 reviewed existing QoE concepts with respect to their integration of interactivity related aspects. This review revealed that, despite the claimed multidimensionality of QoE and the broad consideration of several different influencing factors, interactivity related influencing factors are often neglected in existing QoE models. In order to better address this dimension, a commonality of interactive processes in such applications has been identified, the request-response cycle. As this request-response cycle is inherent to interactive applications in general, it represents an ideal object of study, as con-

clusions regarding its deterioration and the relation to QoE would be, in return, applicable across interactive applications in general. Based on transmission delay as most important impairment in TCP/IP networks, the influence of delays and waiting times on the request-response cycle was chosen as primary focus for the analysis of QoE for interactive services. In terms of temporal stimuli processing within the human sensory system, related work from human time perception psychology revealed the existence of two different timing systems. These two systems differ strongly in their stimulus processing characteristics. In the context of interactive Internet applications, the millisecond timing system is applied for delays < 1 s, and the interval timing system is used for delays > 1 s. To study the delay impact for both these timing systems, two prototypical applications have been selected: Internet telephony where typical delay ranges are below 1 s, and browser based applications where the related delays are often longer than 1 s. The chapter is concluded by an analysis of existing QoE assessment methodologies and its consideration of the request-response cycle. Based on that analysis, related challenges and requirements for QoE assessment methodologies have been derived that have to be met, to properly assess QoE of interactive services and interaction related metrics.

Chapter 3 addressed the influence of transmission delays on human-to-human conversation and its relation to conversational quality. As a starting point, a communication theoretical discussion revealed interruptions as a key interaction cue used for controlling smooth interaction in human conversations. Hence, new conversational metrics, which are able to identify the influence of delay on human-to-human conversations, and which consider interruptions, are proposed: The I³R metric, the UIR metric and the $\frac{I^3R}{ATR}$ ratio. In order to prove the applicability of these metrics for capturing delay induced conversational problems in human-to-human VoIP conversations, two subjective studies were conducted. The data acquired with these studies has then been used to first prove the applicability of the proposed metrics to capture the delay influence on conversational behaviour, and second to derive a conversational prediction model with improved prediction performance compared to state-of-the-art models. Based on these results, updated delay thresholds for high interactive and low interactive conversations were proposed.

The second prototypical application, browser based applications, was analysed with respect to QoE in Chapter 4. In order to address the lack of QoE assessment methodologies enabling proper interaction for browser based applications, a novel test methodology was proposed. Two lab studies and a field study were used to verify the ability of this methodology to deliver reliable and externally valid results.

In order to derive a QoE model for these services, the WQL hypothesis postulates that the relationship between "Waiting time and resulting QoE is Logarithmic", which is a form of the Weber-Fechner law used in psychophysics. Using the data acquired by the three studies, the WQL was verified for file downloads and simple web browsing. Contrary, in the context of complex web browsing it was shown that the WQL hypothesis had to be rejected. A thorough analysis of interactions and respective perceptual events revealed a number of challenges and practical issues on a perceptual and technical level. In this analysis, the subjectively perceived page-load-time was identified as an interaction based measure of waiting time, that is influenced by the content and the task the user follows in a web browsing session.

Finally, Chapter 5 identified five interaction performance aspects that should be considered in a QoE perception model to capture interaction related impairments. Consequently, a perceptual model was proposed that allows to detect these interaction performance aspects for interactions between two or more entities. In addition, it was shown that this model is also able to describe (re-)actions to (conversational) input signals in the form of output signals, which then serve as a news input signal for the other interacting entity and vice versa. Thereby, interactional processes between two or more entities can be explained. This perception model was then integrated into an existing model of the quality formation process that was initially proposed for static input signals. By this modification, the updated model, can be used to describe also interactive quality formation processes, and considers interaction performance aspects for the formation of its QoE output.

Considering future work, related work and results acquired in Chapter 3 indicate that casual human conversations are not severely impacted by transmission delays in terms of conversational quality. However, all of these results have been acquired in dyadic interaction settings with conversations not lasting longer than three minutes. Therefore, future work on the topic of conversational quality should address the following questions: 1) How do transmission delays in human-to-human conversations impact conversations of different duration in terms of conversational quality? E.g. are longer conversations more prone to be disturbed by transmission delays, as the (mental) adaptation load fatigues the participants? 2) What influence has the number of interlocutors on interactional problems and conversational quality, respectively, in delay impaired conversations? E.g. Will a larger number of interlocutors demand a higher degree of control (inter-) actions by the participant to maintain a stable conversation? This maintenance can be severely disturbed by the transmission delays. 3) Do interlocutors sense the impact of the transmission

delay on other dimensions than conversational QoE? E.g. how do the participants perceive the personality of the other interlocutor for different transmission delays? Finally, future work on conversational quality should address interactional metrics as input parameters to a new model that does not carry the legacy of the E-model, and therefore allows for more reasonable relationships between interactional metrics and resulting QoE.

The results presented in Chapter 4 as well as results from related work revealed empirical evidence that in case of not directly perceivable impairments, the exponential model postulated in the IQX hypothesis performed pretty well in terms of fitting performance. However, a theoretical foundation why this is the case is missing, as well as an identification of the relationship between the observed parameters (= not directly perceivable impairments) and the actual psychophysical stimulus that is processed by the human sensory system. In terms of complex web browsing, the WQL hypothesis had to be rejected due to numerous reasons. The major reason identified was the difference between the technical page-load-time, that was used as input to the logarithmic model postulated by the WQL, and the subjectively perceived page-load-time. Further studies that utilise this subjectively perceived page-load-time could help to identify if the WQL holds true for complex web browsing with this input parameter or if other factors not considered in the results of this thesis do influence the QoE perception for that application.

On an application overarching QoE level, the updated quality formation model in Chapter 5 provides guidance and a starting point for the development of novel QoE models, that incorporate interaction performance aspects for different interactive services.

In the larger context of QoE research, there are several topics which will be of high importance in the near future. The assessment, understanding and modelling of QoE for highly interactive services is certainly one of them. As interactive and time variant services are growing by large, QoE prediction models that consider interaction performance aspects and stimuli of longer durations, will be of particular interest to ensure proper prediction performance for such services. Due to the proliferation of resource intensive services, network and service providers struggle to enhance and optimise their networks in terms of customer satisfaction. Accurate and actionable QoE models will enable these stakeholders to better manage their high-performance infrastructure in an active way. Such an active management guarantees that scarce resources can be allocated to services and customers that momentarily demand these resources, and thereby assures high quality experiences.

Satisfied customers will in return be less annoyed, show increased loyalty and will be less inclined to churn.

Bibliography from the Author

- [1] P. L. Callet, S. Möller, and A. Perkis (eds), “Qualinet White Paper on Definitions of Quality of Experience,” Lausanne, Switzerland, Jun. 2012.
- [2] S. Egger, R. Schatz, and S. Scherer, “It Takes Two to Tango - Assessing the Impact of Delay on Conversational Interactivity on Perceived Speech Quality,” in *INTERSPEECH*, 2010, pp. 1321–1324.
- [3] S. Egger, R. Schatz, K. Schoenenberg, A. Raake, and G. Kubin, “Same but Different? - Using Speech Signal Features for Comparing Conversational VoIP Quality Studies,” in *IEEE ICC 2012 - Communication QoS, Reliability and Modeling Symposium (ICC’12 CQRM)*, Ottawa, Ontario, Canada, Jun. 2012.
- [4] A. Raake, K. Schoenenberg, J. Skowronek, and S. Egger, “Predicting Speech Quality based on Interactivity and Delay,” in *Interspeech 2013*, 2013.
- [5] P. Reichl, S. Egger, R. Schatz, and A. D’Alconzo, “The Logarithmic Nature of QoE and the Role of the Weber-Fechner Law in QoE Assessment,” in *Proceedings of the 2010 IEEE International Conference on Communications*, May 2010, pp. 1 –5.
- [6] R. Schatz and S. Egger, “Vienna Surfing - Assessing Mobile Broadband Quality in the Field,” in *Proceedings of the 1st ACM SIGCOMM Workshop on Measurements Up the Stack (W-MUST)*, N. Taft and D. Wetherall, Eds. ACM, 2011.
- [7] R. Schatz, S. Egger, and A. Platzer, “Poor, Good Enough or Even Better? Bridging the Gap between Acceptability and QoE of Mobile Broadband Data Services,” in *Proceedings of the 2011 IEEE International Conference on Communications*, June 2011, pp. 1 –6.

- [8] S. Egger, P. Reichl, T. Hossfeld, and R. Schatz, “ ’Time is Bandwidth’? Narrowing the Gap between Subjective Time Perception and Quality of Experience,” in *IEEE International conference on communications (ICC)*, Ottawa, Ontario, Canada, 2012.
- [9] A. Raake and S. Egger, “Quality and Quality of Experience,” in *Quality of Experience: Advanced Concepts, Applications and Methods*, S. Möller and A. Raake, Eds. Springer, Jun. 2014.
- [10] S. Egger, R. Schatz, T. Hoßfeld, and W. Müllner, “ITU-T SG 12 CONTRIBUTION C-033: Perceptual Events in a Page View Cycle, outcome from the interim meeting in Berlin 11/2012,” ITU, Vienna, Austria, Tech. Rep., November 2013.
- [11] —, “ITU-T SG 12 CONTRIBUTION C-034: Relevant Factors and Use Cases for Web QoE,” ITU, Geneva, Switzerland, Tech. Rep., November 2013.
- [12] —, “ITU-T SG 12 CONTRIBUTION C-046: Draft Test Plan for P.STMWeb,” ITU, Geneva, Switzerland, Tech. Rep., November 2013.
- [13] —, “ITU-T SG 12 CONTRIBUTION C-049: Web page categorization for P.STMWeb,” ITU, Geneva, Switzerland, Tech. Rep., November 2013.
- [14] —, “ITU-T SG 12 CONTRIBUTION C-336: P.863 performance on GSM handover impaired speech samples compared to P.862,” ITU, Geneva, Switzerland, Tech. Rep., November 2012.
- [15] S. Egger, “Interactive content for subjective studies on web browsing qoe: A kepler derivative,” in *Workshop on Selected Items on Telecommunication Quality Matters*. Vienna: ETSI, November 2012.
- [16] K. Hoeldtke, A. Raake, S. Möller, S. Egger, R. Schatz, and N. Rohrer, “ITU-T SG 12 CONTRIBUTION 189: How the Need for fast Interaction affects the Impact of Transmission Delay on the overall Quality Judgment,” FTW, Geneva, Switzerland, Tech. Rep., November 2011.
- [17] S. Egger and R. Schatz, “Perceptual Events in a Page View Cycle,” FTW, Vienna, Austria, Tech. Rep., November 2012.
- [18] S. Egger, R. Schatz, D. Strohmeier, and A. Raake, “Shortcomings of G.1030 Annex A,” FTW, Berlin, Germany, Tech. Rep., November 2012.

- [19] S. Egger and P. Reichl, "A Nod Says More than Thousand Uhmms: Towards a Framework for Measuring Audio-Visual Interactivity," in *COST 298 Conference: THE GOOD, THE BAD AND THE CHALLENGING*, COST298. COST298, May 2009.
- [20] S. Egger, "Why Videotelephony (currently) Fails: An Interactional Perspective," in *First International Conference on 'What makes Humans Human'*, Ulm, March 2010.
- [21] R. Schatz, T. Hossfeld, L. Janowski, and S. Egger, "From Packets to People: Quality of Experience as New Measurement Challenge," in *TMA Book*. Springer LNCS, Apr. 2013.
- [22] A. Sackl, P. Zwickl, S. Egger, and P. Reichl, "The Role of Cognitive Dissonance for QoE Evaluation of Multimedia Services," in *Proceedings of IEEE Workshop on Quality of Experience for Multimedia Communications - QoEMC2012, Anaheim, California*. IEEE, Dec. 2012.
- [23] A. Sackl, S. Egger, P. Zwickl, and P. Reichl, "The QoE Alchemy: Turning Quality into Money. Experiences with a Refined Methodology for the Evaluation of Willingness-to-pay for Service Quality," in *Proc. QoMEX (Quality of the Multimedia Experience) 2012, Yarra Valley, Australia*. IEEE, Jul. 2012.
- [24] K. Masuch, M. Muehlegger, A. Sackl, S. Egger, R. Schatz, E. Oberzaucher, and K. Grammer, "What you get is what you see? Pretending facts in applied user ratings studies," in *Proc. XXI Biennial Conference on Human Ethology, Austria, Vienna*, Aug. 2012.
- [25] P. Casas, A. Sackl, and S. Egger, "YouTube & Facebook Quality of Experience in Mobile Broadband Networks," in *Proceedings of the IEEE Globecom 2012, Anaheim, California*. IEEE, Dec. 2012.
- [26] P. Casas, M. Seufert, S. Egger, and R. Schatz, "Quality of Experience in Remote Virtual Desktop Services," in *Proc. IFIP/IEEE Workshop on QoE-Centric Management (QCMan 2013), Ghent, Belgium*. IEEE, May 2013.
- [27] A. Sackl, S. Egger, and R. Schatz, "Where's the Music? Comparing the QoE impact of temporal impairments between music and video streaming," in *Proc. The fifth international workshop on Quality of Multimedia Experience (QoMEX), Klagenfurt, Austria*, Jul. 2013.

- [28] R. Schatz and S. Egger, "The Impact of Terminal Performance and Screen Size on QoE," in *Proceedings of ETSI Workshop on Selected Items on Telecommunication Quality Matters, Vienna, Austria*, P. Pocta and J. Pomy, Eds. ETSI, Nov. 2012.
- [29] S. Egger, T. Hossfeld, R. Schatz, and M. Fiedler, "Tutorial: Waiting Times in Quality of Experience for Web based Services," in *IEEE QoMEX 2012, Yarra Valley, Australia*, 2012.
- [30] P. Fröhlich, S. Egger, R. Schatz, M. Muehlegger, K. Masuch, and B. Gardlo, "QoE in 10 Seconds: Are Short Video Clip Lengths Sufficient for Quality of Experience Assessment?" in *Proc. QoMEX (Quality of the Multimedia Experience) 2012, Yarra Valley, Australia*, Jul. 2012.
- [31] T. Hossfeld, S. Egger, R. Schatz, M. Fiedler, K. Masuch, and C. Lorentzen, "Initial Delay Vs. Interruptions: Between The Devil And The Deep Blue Sea," in *Proc. QoMEX 2012, Yarra Valley, Australia*, Jul. 2012.
- [32] A. Sackl, K. Masuch, S. Egger, and R. Schatz, "Wireless vs. Wireline Shootout: How user expectations influence Quality of Experience," in *Proc. QoMEX (Quality of the Multimedia Experience) 2012, Yarra Valley, Australia*, Jun. 2012.
- [33] R. Schatz, S. Egger, and K. Masuch, "The Impact of User Fatigue and Test Duration on the Reliability of Subjective Quality Ratings," *JAES - Journal of the Audio Engineering Society*, 2012.
- [34] T. Hossfeld, R. Schatz, and S. Egger, "SOS: The MOS is not enough!" in *QoMEX 2011*, Mechelen, Belgium, Sep. 2011.
- [35] T. Hossfeld, R. Schatz, S. Biedermann, A. Platzer, S. Egger, and M. Fiedler, "The Memory Effect and Its Implications on Web QoE Modeling," in *23rd International Teletraffic Congress (ITC 2011)*, San Francisco, USA, Sep. 2011.
- [36] R. Schatz, S. Egger, and T. Hossfeld, "Understanding Ungeduld – Quality of Experience Assessment and Modeling for Internet Applications," in *Proc. EuroView 2011, Würzburg, Germany*, Aug. 2011.
- [37] K. Masuch, R. Schatz, S. Egger, I. Holzleitner, E. Oberzaucher, and K. Grammer, "The duration effect in rating studies - quantity instead of quality?" in *Human Behavior and Evolution Society (HBES)*, 2011.

- [38] R. Schatz, S. Egger, K. Masuch, and S. Scherer, "Gain from Strain? Measuring the Influence of User Fatigue on the Quality of Subjective Ratings," in *Third International Workshop on Perceptual Quality of Systems 2010*, U. Jekosch, E. Altinsoy, M. Sebastian, and A. Raake, Eds., vol. 3. ISCA, September 2010.
- [39] S. Egger, P. Reichl, and M. Ries, "Quality-of-Experience Beyond MOS: Experiences with a Holistic User Test Methodology for Interactive Video Services," in *21st ITC Specialist Seminar on Multimedia Applications - Traffic, Performance and QoE*, Miyazaki, Japan, 3 2010, pp. 13–18.
- [40] R. Schatz, L. Baillie, P. Fröhlich, S. Egger, and T. Grechenig, "What Are You Viewing?" Exploring the Pervasive Social TV Experience," in *Mobile TV: Customizing Content and Experience*, Marcus, A., Roibás, A. C., & Sala, R., Ed. Springer, 2010, pp. 255–+.
- [41] A. Baer, A. Berger, S. Egger, and R. Schatz, "A Lightweight Mobile TV Recommender," in *Changing Television Environments*, ser. Lecture Notes in Computer Science, M. Tscheligi, M. Obrist, and A. Lugmayr, Eds. Springer Berlin / Heidelberg, 2008, vol. 5066, pp. 143–147.
- [42] R. Schatz, L. Baillie, P. Fröhlich, and S. Egger, "Getting the Couch Potato to Engage in Conversation: Social TV in a Converging Media Environment," in *EuroITV 2008*, July 2008.
- [43] P. Reichl, F. Hammer, S. Egger, and M. Ries, "The Well-Tempered Conversation: On Quality of Experience of Interactive VoIP," in *Workshop on Socio-Economic Aspects of Future Generation Internet*. Blekinge Institute of Technology, May 2008.
- [44] R. Schatz and S. Egger, "Social Interaction Features for Mobile TV Services," in *Proc. IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, March 31 2008–April 2 2008, pp. 1–6.
- [45] R. Schatz, S. Wagner, S. Egger, and N. Jordan, "Mobile tv becomes social - integrating content with communications," in *Proc. 29th International Conference on Information Technology Interfaces ITI 2007*, 25–28 June 2007, pp. 263–270.

- [46] M. S. et al., “To Pool or not to Pool”: A Comparison of Temporal Pooling Methods for HTTP Adaptive Video Streaming,” in *Proc. of the QoMEX , Klagenfurt, Austria.* IEEE, 2013.
- [47] K. Schoenenberg, A. Raake, S. Egger, and R. Schatz, “On Interaction Behaviour in Telephone Conversations under Transmission Delay,” *Speech Communication*, 2014.
- [48] B. Gardlo, S. Egger, and M. Seufert, “Crowdsourcing 2.0: Enhancing Execution Speed and Reliability of Web-based QoE Testing,” in *Proc. IEEE ICC, Sydney, Australia*, Jun. 2014.
- [49] S. Egger, P. Reichl, and K. Schoenenberg, “Quality of experience and interactivity,” in *Quality of Experience: Advanced Concepts, Applications and Methods*, S. Möller and A. Raake, Eds. Springer, Jun. 2014.
- [50] S. Egger, R. Schatz, and W. Karner, “ITU-T SG 12 CONTRIBUTION C-248: Web Browsing QoE Subjective Testing Methodology,” ITU, Geneva, Switzerland, Tech. Rep., November 2011.
- [51] S. Egger, “ITU-T SG 12 Temporary Document TD-GEN-0272: Draft Recommendation ITU-T G.QoE-Web,” ITU-T, Geneva, Switzerland, Tech. Rep., December 2013.
- [52] K. Schoenenberg, A. Raake, S. Egger, and R. Schatz, “On Interaction Behaviour in Telephone Conversations under Transmission Delay,” *Speech Communication*, Jun. 2014.
- [53] S. Egger, “MAVIA - Mediated Audio-Visual Interaction Analysis,” Master’s thesis, Institute of Sociology, University of Graz, Graz, Austria, September 2008.

Bibliography

- [54] A. Matthew et al., “Measuring human satisfaction in data networks,” in *Proceedings of INFOCOM 2006*. IEEE, 2006.
- [55] L. G. Allan, “The perception of time,” *Attention, Perception, & Psychophysics*, vol. 26, no. 5, 1979.
- [56] G. Antonides, P. C. Verhoef, and M. van Aalst, “Consumer perception and evaluation of waiting time: A field experiment,” *Journal of Consumer Psychology*, vol. 12, no. 3, 2002.
- [57] J.-N. Antons, S. Arndt, R. Schleicher, and S. Möller, “Brain Activity Correlates of Quality of Experience,” in *Quality of Experience: Advanced Concepts, Applications and Methods*, S. Möller and A. Raake, Eds. Springer, Jun. 2014.
- [58] H. Arao, D. Suetomi, and Y. Nakajima, “Does time-shrinking take place in visual temporal patterns?” *Perception*, vol. 29, no. 7, pp. 819–830, 2000. [Online]. Available: <http://www.perceptionweb.com/abstract.cgi?id=p2853>
- [59] M. Belshe, “More Bandwidth does not Matter (much),” Google, Tech. Rep., 2010.
- [60] N. Bhatti, A. Bouch, and A. Kuchinsky, “Integrating User-Perceived quality into web server design,” in *9th International World Wide Web Conference*, 2000, pp. 1 – 16.
- [61] A. Bouch, M. A. Sasse, and H. G. DeMeer, “Of packets and people: a user-centered approach to quality of service,” in *Proceedings of IWQoS 2000*, 2000.
- [62] A. Bouch, A. Kuchinsky, and N. Bhatti, “Quality is in the eye of the beholder: meeting users’ requirements for internet quality of service,” in *CHI ’00: Proceedings of the SIGCHI conference on Human factors in computing systems*. New York, NY, USA: ACM, 2000, pp. 297–304.

- [63] F. Boulos, B. Parrein, P. Le Callet, D. Hands *et al.*, “Perceptual Effects of Packet Loss on H. 264/AVC encoded Videos,” 2009.
- [64] P. T. Brady, “A Technique for Investigating On-Off Patterns of Speech,” *Bell System Technical Journal*, vol. 44, no. 1, pp. 1–22, Jan. 1965.
- [65] —, “A Statistical Analysis of On-Off Patterns in 16 Conversations,” *Bell System Technical Journal*, vol. 47, no. 1, pp. 73–91, Jan. 1968.
- [66] —, “A model for generating on-off patterns in two-way communications,” *Bell System Technical Journal*, vol. 48, pp. 2445–2472, Sep. 1969.
- [67] —, “Effects of transmission delay on conversational behavior on echo-free telephone circuits,” *Bell System Technical Journal*, vol. 50, no. 1, pp. 115–134, Jan. 1971.
- [68] P. Brooks and B. Hestnes, “User measures of quality of experience: why being objective and quantitative is important,” *Network, IEEE*, vol. 24, no. 2, pp. 8–13, Apr. 2010.
- [69] C. V. Buhusi and W. H. Meck, “What makes us tick? functional and neural mechanisms of interval timing,” *Nature Reviews Neuroscience*, vol. 6, no. 10, pp. 755–765, Sep. 2005. [Online]. Available: <http://www.nature.com/doi/10.1038/nrn1764>
- [70] M. Butkiewicz, H. V. Madhyastha, and V. Sekar, “Understanding website complexity: Measurements, metrics, and implications,” in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM, 2011, pp. 313–328.
- [71] M. C. Chan and R. Ramjee, “Tcp/ip performance over 3g wireless links with rate and delay variation,” *Wireless Networks*, vol. 11, no. 1-2, pp. 81–97, 2005.
- [72] K. T. Chen, C. J. Chang, C. C. Wu, Y. C. Chang, and C. L. Lei, “Quadrant of euphoria: a crowdsourcing platform for QoE assessment,” *Network, IEEE*, vol. 24, no. 2, pp. 28 – 35, 2010.
- [73] Cisco, “Cisco visual networking index: Forecast and methodology 2012 to 2017,” Cisco, Tech. Rep., 2013.

- [74] D. Collange and J.-L. Costeux, "Passive estimation of quality of experience," *Journal of Universal Computer Science*, vol. 14, no. 5, pp. 625–641, 2008.
- [75] E. Crawley, R. Nair, B. Rajagopalan, and H. Sandick, "RFC 2386: A Framework for QoS-based Routing in the Internet," IETF, Tech. Rep., Aug. 1998. [Online]. Available: <http://www.ietf.org/rfc/rfc2386.txt>
- [76] M. Csikszentmihalyi and I. S. Csikszentmihalyi, *Optimal experience: Psychological studies of flow in consciousness*. Cambridge University Press, 1992.
- [77] H. Cui and E. Biersack, "Trouble shooting interactive web sessions in a home environment," in *Proceedings of the 2nd ACM SIGCOMM workshop on Home networks*, ser. HomeNets '11. New York, NY, USA: ACM, 2011, pp. 25–30. [Online]. Available: <http://doi.acm.org/10.1145/2018567.2018574>
- [78] D. Strohmeier et al., "Toward task-dependent evaluation of web-QoE: Free exploration vs. Who Ate What?" in *Globecom Workshops*. IEEE, 2012, pp. 1309–1313.
- [79] K. De Moor, L. De Marez, T. Deryckere, W. Joseph, and L. Martens, "Bridging troubled water: Quality of experience in a mobile media context," in *Terena Networking Conference*, Bruges, May 2008.
- [80] B. G. Dellaert and B. E. Kahn, "How tolerable is delay?: Consumers' evaluations of internet web sites after waiting," *Journal of interactive marketing*, vol. 13, no. 1, pp. 41–54, 1999.
- [81] F. Dobrian, V. Sekar, A. Awan, I. Stoica, D. Joseph, A. Ganjam, J. Zhan, and H. Zhang, "Understanding the impact of video quality on user engagement," in *Proceedings of the ACM SIGCOMM 2011 conference*, ser. SIGCOMM '11. New York, NY, USA: ACM, 2011, pp. 362–373. [Online]. Available: <http://doi.acm.org/10.1145/2018436.2018478>
- [82] P. C. Earley, G. B. Northcraft, C. Lee, and T. R. Lituchy, "Impact of process and outcome feedback on the relation of goal setting to task performance," *Academy of Management Journal*, vol. 33, no. 1, pp. 87–105, 1990.
- [83] H. Eisler, "Subjective duration and psychophysics." *Psychological Review*, vol. 82, no. 6, pp. 429–450, 1975.

- [84] J. Emling and D. Mitchell, "The effects of time delay and echos on telephone conversations," *Bell System Technical Journal*, vol. 42, no. 2, pp. 2869–2891, Nov. 1963.
- [85] M. Fiedler, "Deliverable D.WP.JRA.6.1.1: state of the art with regards to user perceived quality of service and quality feedback," EuroNGI, Tech. Rep., May 2004. [Online]. Available: <http://eurongi.enst.fr>
- [86] M. Fiedler, T. Hoffeld, and P. Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," *Netw. Mag. of Global Internetwkg.*, vol. 24, pp. 36–41, March 2010.
- [87] Forrester Research, "eCommerce Web Site Performance Today," Akamai, Tech. Rep., 2009.
- [88] N. Freed and N. Borenstein, "Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types," RFC 2046 (Draft Standard), Internet Engineering Task Force, November 1996, updated by RFCs 2646, 3798, 5147. [Online]. Available: <http://www.ietf.org/rfc/rfc2046.txt>
- [89] D. F. Galletta, R. M. Henry, S. McCoy, and P. Polak, "When the wait isn't so bad: The interacting effects of website delay, familiarity, and breadth," *Information Systems Research*, vol. 17, no. 1, pp. 20–37, 2006.
- [90] E. B. Goldstein, *Sensation and perception*. Cengage Learning, 2013.
- [91] S. Goldstone and W. T. Lhamon, "Studies of auditory-visual differences in human time judgment. 1. sounds are judged longer than lights." *Perceptual and motor skills*, vol. 39, no. 1, pp. 63–82, 1974.
- [92] G. J. Gorn, A. Chattopadhyay, J. Sengupta, and S. Tripathi, "Waiting for the Web: How Screen Color Affects Time Perception," *Journal of Marketing Research*, vol. 41, no. 2, pp. 215–225, 2004.
- [93] S. Grondin, "From physical time to the first and second moments of psychological time." *Psychological Bulletin*, vol. 127, no. 1, pp. 22–44, 2001.
- [94] ———, "Sensory modalities and temporal processing," in *Time and mind II: information processing perspectives*, H. Helfrich, Ed. Hogrefe & Huber, 2003.

- [95] —, “Timing and time perception: a review of recent behavioral and neuroscience findings and theoretical directions.” *Attention perception psychophysics*, vol. 72, no. 3, pp. 561–582, 2010.
- [96] L. Gros and N. Chateau, “The impact of listening and conversational situations on speech perceived quality for time-varying impairments,” in *Proceedings of the 1st International Conference on Measurement of Speech and Audio Quality in Networks (MESAQIN '02)*. Prague, Czech Republic: MESAQIN, January 2002.
- [97] T. I. R. Group, “The network providers business case for internet content delivery,” Akamai, Tech. Rep., 1999.
- [98] M. Guéguin, R. Le Bouquin-Jeannès, V. Gautier-Turbin, G. Faucon, and V. Barriac, “On the Evaluation of the Conversational Speech Quality in Telecommunications,” *EURASIP J. Adv. Signal Process*, vol. 2008, pp. 1–15, 2008.
- [99] D. Guse and S. Möller, “Macro-temporal development of qoe: Impact of varying performance on qoe over multiple interactions,” in *DAGA 2013*, 2013.
- [100] F. Hammer, P. Reichl, and A. Raake, “The well-tempered Conversation: Interactivity, Delay and perceptual VoIP Quality,” *Communications, 2005. ICC 2005. 2005 IEEE International Conference on*, vol. 1, pp. 244–249 Vol. 1, May 2005.
- [101] F. Hammer, “Quality aspects of packet-based interactive speech communication,” Ph.D. dissertation, Signal Processing and Speech Communication Laboratory, Faculty of Electrical and Information Engineering, University of Technology Graz, Graz, Austria, June 2006.
- [102] F. Hammer, P. Reichl, and A. Raake, “Elements of Interactivity in Telephone Conversations,” in *Proc. in Proc. 8th International Conference on Spoken Language Processing (ICSLP/INTERSPEECH 2004)*, Jeju Island, vol. 3, Oct. 2004, pp. 1741–1744.
- [103] D. Hands and M. Wilkins, “A Study of the Impact of Network Loss and Burst Size on Video Streaming Quality and Acceptability,” in *Interactive Distributed Multimedia Systems and Telecommunication Services*, ser. Lecture

- Notes in Computer Science, M. Diaz, P. Owezarski, and P. Senac, Eds. Springer Berlin / Heidelberg, 1999, vol. 1718, pp. 45–57. [Online]. Available: <http://www.springerlink.com/content/21u2413r58534152/abstract/>
- [104] D. S. Hands and S. E. Avons, “Recency and duration neglect in subjective assessment of television picture quality,” *Applied Cognitive Psychology*, vol. 15, no. 6, pp. 639–657, Nov. 2001. [Online]. Available: <http://doi.wiley.com/10.1002/acp.731>
- [105] H. D. Höhne, *Influence of long transmission delays and reverberation on telephone conversations of Testpersons*. VDE-Verlag GmbH, 1970. [Online]. Available: <http://books.google.at/books?id=SgyTMwAACAAJ>
- [106] J. Holub and O. Tomiska, “Delay effect on conversational quality in telecommunication networks: Do we mind?” *Wireless Technology*, 2009.
- [107] T. Hoßfeld, D. Hock, P. Tran-Gia, K. Tutschku, and M. Fiedler, “Testing the IQX Hypothesis for Exponential Interdependency between QoS and QoE of Voice Codecs iLBC and G.711,” University of Wuerzburg, Tech. Rep. 442, mar 2008.
- [108] T. Hoßfeld, R. Schatz, M. Seufert, M. Hirth, T. Zinner, and P. Tran-Gia, “Quantification of YouTube QoE via Crowdsourcing,” in *IEEE MQoE Workshop*, Dana Point, CA, USA, Dec. 2011.
- [109] T. Hoßfeld, P. Tran-Gia, and M. Fiedler, “Quantification of quality of experience for edge-based applications,” in *20th International Teletraffic Congress (ITC20)*, Ottawa, Canada, jun 2007.
- [110] Q. Huynh-Thu, M. Garcia, F. Speranza, P. Corriveau, and A. Raake, “Study of rating scales for subjective quality assessment of High-Definition video,” *Broadcasting, IEEE Transactions on*, vol. 57, no. 1, pp. 1–14, 2011.
- [111] S. Iai, T. Kurita, and N. Kitawaki, “Quality requirements for multimedia communication services and terminals-interaction of speech and video delays,” in *Global Telecommunications Conference, 1993, including a Communications Theory Mini-Conference. Technical Program Conference Record, IEEE in Houston. GLOBECOM '93., IEEE*, 1993, pp. 394–398 vol.1.

- [112] E. Ibarrola, F. Liberal, I. Taboada, and R. Ortega, “Web qoe evaluation in multi-agent networks: Validation of itu-t g.1030,” in *ICAS '09: Proceedings of the 2009 Fifth International Conference on Autonomic and Autonomous Systems*. Washington, DC, USA: IEEE Computer Society, 2009, pp. 289–294.
- [113] International Telecommunication Union, *Handbook on Telephony*. ITU-T, July 1992.
- [114] —, “Methods for Subjective Determination of Transmission Quality,” *ITU-T Recommendation P.800*, Aug. 1996.
- [115] —, “Subjective performance evaluation of network echo cancellers,” *ITU-T Recommendation P.831*, Dec. 1998.
- [116] —, “End-user multimedia QoS categories,” *ITU-T Recommendation G.1010*, November 2001.
- [117] —, “One-way transmission time,” *ITU-T Recommendation G.114*, Aug. 2003.
- [118] —, “Talker echo and its control,” *ITU-T Recommendation G.131*, November 2003.
- [119] —, “Vocabulary and effects of transmission parameters on customer opinion of transmission quality, amendment 2,” *ITU-T Recommendation P.10/G.100*, 2006.
- [120] —, “Subjective Evaluation of Conversational Quality,” *ITU-T Recommendation P.805*, July 2007.
- [121] —, “Quality of telecommunication services: Concepts, models, objectives and dependability planning. terms and definitions related to the quality of telecommunication services,” *ITU-T Recommendation E.800*, Sep. 2008.
- [122] —, “Subjective video quality assessment methods for multimedia applications,” *ITU-T Recommendation P.910*, April 2008.
- [123] —, “The E-model, a computational model for use in transmission planning,” *ITU-T Recommendation G.107*, April 2009.

- [124] —, “Reference Guide to Quality of Experience Assessment Methodologies,” *ITU-T Recommendation G.1011*, June 2010.
- [125] —, “Methodology for the Subjective Assessment of the Quality of Television Pictures,” *ITU-R Recommendation BT.500*, 2012.
- [126] —, “Estimating end-to-end performance in ip networks for data applications,” *ITU-T Recommendation G.1030*, Feb. 2014.
- [127] —, “QoE factors in web-browsing,” *ITU-T Recommendation G.1031*, Feb. 2014.
- [128] —, “Subjective testing methodology for web browsing,” *ITU-T Recommendation P.1501*, Feb. 2014.
- [129] Internet Society, “Bandwidth Management: Internet Society Technology Roundtable,” Internet Society, Geneva, Switzerland, Tech. Rep., November 2012.
- [130] J. Issing and N. Farber, “Conversational quality as a function of delay and interactivity,” in *Software, Telecommunications and Computer Networks (SoftCOM), 2012 20th International Conference on*, 2012, pp. 1–5.
- [131] ITU-T Study Group 12, *Practical procedures for subjective testing 2011*. Geneva: ITU, 2011.
- [132] U. Jekosch, *Voice And Speech Quality Perception: Assessment And Evaluation*, ser. Signals And Communication Technology. Springer, 2005. [Online]. Available: <http://books.google.at/books?id=Ef3lHiSzq1QC>
- [133] E. Jones, C. Gallois, V. Callan, and M. Barker, “Strategies of accommodation: Development of a coding system for conversational interaction,” *Journal of Language and Social Psychology*, vol. 18, no. 2, pp. 123–151, 1999.
- [134] S. Jumisko-Pyykkö, V. K. Malamal Vadakital, and M. M. Hannuksela, “Acceptance threshold: A bidimensional research method for user-oriented quality evaluation studies,” *International Journal of Digital Multimedia Broadcasting*, 2008.
- [135] D. Kahneman, D. Kahneman, A. Tversky *et al.*, “Experienced utility and objective happiness: A moment-based approach,” *The psychology of economic decisions*, vol. 1, pp. 187–208, 2003.

- [136] D. Karis, "Evaluating transmission quality in mobile telecommunication systems using conversation tests," in *Human Factors Society 35th Annual Meeting*, vol. 1, Santa Monica, CA, 1991, pp. 217–221.
- [137] R. Khare and I. Jacobs, "W3C Recommendations Reduce 'World Wide Wait'," <http://www.w3.org/Protocols/NL-PerfNote.html>, July 1999, accessed: 2013-05-10.
- [138] S. Khirman and P. Henriksen, "Relationship between Quality-of-Service and quality-of- experience for public internet service," in *Proceedings of the 3rd Workshop on Passive and Active Measurement*, Fort Collins, Colorado, USA, Mar. 2002.
- [139] P. R. Killeen and N. A. Weiss, "Optimal timing and the weber function." *Psychological Review*, vol. 94, no. 4, pp. 455–468, 1987.
- [140] A. King, *Speed Up Your Site: Web Site Optimization*. Indianapolis: New Riders, 2003.
- [141] N. Kitawaki and K. Itoh, "Pure delay effects on speech quality in telecommunications," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 4, pp. 586–593, 1991.
- [142] J. Klein, *The Sociology of Behaviour and Psychology: In 18 Volumes. The study of groups*, ser. The international library of sociology. Routledge, 1956, no. v. 16. [Online]. Available: <http://books.google.at/books?id=HFz7kssDbbUC>
- [143] E. Klemmer, "Subjective evaluation of delay in Telephone Communications," *Bell System Technical Journal*, vol. 46, pp. 1141–1147, Sep. 1967.
- [144] H. Knoche, H. De Meer, and D. Kirsh, "Utility curves: mean opinion scores considered biased," in *Quality of Service, 1999. IWQoS '99. 1999 Seventh International Workshop on*, 1999, pp. 12–14.
- [145] H. O. Knoche, "Quality of experience in digital mobile multimedia services," <http://discovery.ucl.ac.uk/1322706/>, Jul. 2011. [Online]. Available: <http://discovery.ucl.ac.uk/1322706/>
- [146] R. Krauss and P. Bricker, "Effects of transmission delay and access delay on the efficiency of verbal communication," *Journal of the Acoustical Society of America*, vol. 42, pp. 286–292, 1967.

- [147] R. Kubey and M. Csikszentmihalyi, *Television and the Quality of Life: How Viewing Shapes Everyday Experience*, ser. A Volume in the Communication Series. L. Erlbaum Associates, 1990. [Online]. Available: http://books.google.at/books?id=zk_Zg5fJSVwC
- [148] R. Kumar and A. Tomkins, "A characterization of online browsing behavior," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 561–570.
- [149] K. Laghari, N. Crespi, and K. Connelly, "Toward total quality of experience: A QoE model in a communication ecosystem," *Communications Magazine, IEEE*, vol. 50, no. 4, pp. 58–65, Apr. 2012.
- [150] T. Lakshman and U. Madhow, "The performance of tcp/ip for networks with high bandwidth-delay products and random loss," *Networking, IEEE/ACM Transactions on*, vol. 5, no. 3, pp. 336–350, 1997.
- [151] W. T. Lhamon and S. Goldstone, "Studies of auditory-visual differences in human time judgment. 2. more transmitted information with sounds than lights." *Perceptual and motor skills*, vol. 39, no. 1, pp. 295–307, 1974.
- [152] E. A. Locke, K. N. Shaw, L. M. Saari, and G. P. Latham, "Goal setting and task performance: 1969-1980," DTIC Document, Tech. Rep., 1980.
- [153] I. Luengo, E. Navas, I. Odriozola, I. Saratxaga, I. Hernaez, I. Sainz, and D. Erro, "Modified LTSE-VAD Algorithm for Applications requiring reduced Silence Frame Misclassification," in *Proceedings of the LREC*, May 2010.
- [154] R. L. Mandryk, K. M. Inkpen, and T. W. Calvert, "Using psychophysiological techniques to measure user experience with entertainment technologies," *Behaviour & IT*, vol. 25, no. 2, pp. 141–158, 2006.
- [155] A. P. Markopoulou, F. A. Tobagi, and M. J. Karam, "Assessing the quality of voice communications over Internet backbones," *IEEE/ACM Transactions On Networking*, vol. 11, no. 5, Oct. 2003.
- [156] S. McMillan, "Exploring models of interactivity from multiple research traditions: users, documents and systems," *Handbook of new media*, vol. 2, pp. 205–29, 2005.

- [157] R. B. Miller, “Response time in man-computer conversational transactions,” in *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*. ACM, 1968, pp. 267–277.
- [158] S. Möller, *Quality of Telephone-Based Spoken Dialogue Systems*. Springer, 2006. [Online]. Available: <http://books.google.at/books?id=ZVMHOvS8LlcC>
- [159] S. Möller, C. Bang, T. Tamme, M. Vaalgamaa, and B. Weiss, “From single-call to multi-call quality: A study on long-term quality integration in audio-visual speech communication,” in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.
- [160] S. Möller, W. Chan, N. Cote, T. Falk, A. Raake, and M. Wältermann, “Speech quality estimation: Models and trends,” *Signal Processing Magazine, IEEE*, vol. 28, no. 6, pp. 18–28, Nov. 2011.
- [161] S. Möller, K.-P. Engelbrecht, C. Kühnel, I. Wechsung, and B. Weiss, “A taxonomy of quality of service and quality of experience of multimodal human-machine interaction,” in *Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on*, 2009, pp. 7–12. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5246986
- [162] S. Möller, *Assessment and Prediction of Speech Quality in Telecommunications*, 1st ed. Springer, August 2000.
- [163] —, *Quality Engineering - Qualität kommunikationstechnischer Systeme*. Springer, 2010.
- [164] S. Möller, K.-P. Engelbrecht, C. Kühnel, I. Wechsung, and B. Weiss, “Evaluation of multimodal interfaces for ambient intelligence,” *Human-Centric Interfaces for Ambient Intelligence*, pp. 347–370, 2009.
- [165] S. Möller and A. Raake, “Telephone speech quality prediction: towards network planning and monitoring models for modern network scenarios,” *Speech Communication*, vol. 38, pp. 47–75, Sep. 2002, ACM ID: 638082. [Online]. Available: <http://portal.acm.org/citation.cfm?id=638078.638082>
- [166] K. D. Moor, I. Ketyko, W. Joseph, T. Deryckere, L. D. Marez, L. Martens, and G. Verleye, “Proposed framework for evaluating quality of experience in a mobile, testbed-oriented living lab setting,” *Mobile Networks and*

- Applications*, vol. 15, no. 3, pp. 378–391, 2010. [Online]. Available: <http://www.springerlink.com/index/10.1007/s11036-010-0223-0>
- [167] J. Nielsen, *Usability Engineering*. San Francisco, California: Morgan Kaufmann Publishers, October 1993.
- [168] J. Nielsen and J. Levy, “Measuring usability: Preference vs. performance,” *Commun. ACM*, vol. 37, no. 4, pp. 66–75, Apr. 1994. [Online]. Available: <http://doi.acm.org/10.1145/175276.175282>
- [169] S. Niida, S. Uemura, and H. Nakamura, “Mobile services,” *Vehicular Technology Magazine, IEEE*, vol. 5, no. 3, pp. 61–67, sept. 2010.
- [170] T. P. Novak and D. L. Hoffman, “Measuring the flow experience among web users,” *Interval Research Corporation*, vol. 31, 1997. [Online]. Available: <http://www.whueb.com/whuebiz/emarketing/research/m031121/m031121c.pdf>
- [171] D. Olshefski and J. Nieh, “Understanding the management of client perceived response time,” in *Proceedings of the joint international conference on Measurement and modeling of computer systems*, 2006, pp. 240–251.
- [172] K. Papamiltiadis, H. Zisimopoulos, M. Gasparroni, and A. Liotta, “User quality of service perception in 3g mobile networks,” *Telecommunications Quality of Services: The Business of Success, 2004. QoS 2004. IEE*, pp. 64–69, March 2004.
- [173] A. Parasuraman, V. A. Zeithaml, and L. L. Berry, “A conceptual model of service quality and its implications for future research,” *The Journal of Marketing*, pp. 41–50, 1985.
- [174] A. Raake, M. Garcia, S. Möller, J. Berger, F. Kling, P. List, J. Johann, and C. Heidemann, “T-V-model: parameter-based prediction of IPTV quality,” in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, Apr. 2008, pp. 1149–1152.
- [175] A. Raake, “Predicting Speech Quality under Random Packet Loss: Individual Impairment and Additivity with other Network Impairments,” *ACUSTICA/Acta Acustica*, vol. 90, no. 6, pp. 1061–1083, 2004.

- [176] ———, *Speech Quality of VoIP: Assessment and Prediction*. John Wiley & Sons, 2006.
- [177] S. Rafaeli, “Interactivity: from new media to communication,” in *Advancing Communication Science: Merging Mass and Interpersonal Processes*, R. P. Hawkins, J. M. Wiemann, and S. Pingree, Eds. Sage Publications, 1998, pp. 110–135. [Online]. Available: http://gsb.haifa.ac.il/~sheizaf/interactivity/Interactivity_Rafaeli.pdf
- [178] P. Reichl, “How to Define Conversational Interactivity: A Game-Theoretic Approach and Its Application in Telecommunications,” *Journal of Information Technologies and Control (JITC)*, no. No. 3-4/2006, pp. 18–24, Feb. 2007.
- [179] ———, “From charging for quality of service to charging for quality of experience,” *Annales des Télécommunications*, vol. 65, no. 3-4, pp. 189–199, 2010.
- [180] P. Reichl and F. Hammer, “Hot Discussion or Frosty Dialogue? Towards a Temperature Metric for Conversational Interactivity.” in *Proc. in Proc. 8th International Conference on Spoken Language Processing (ICSLP/INTERSPEECH 2004)*, Jeju Island, Oct. 2004, pp. 317–320.
- [181] P. Reichl, B. Tuffin, and R. Schatz, “Logarithmic laws in service quality perception: where microeconomics meets psychophysics and quality of experience,” *Telecommunication Systems*, pp. 1–14, 2011.
- [182] J. Research, “Retail Web Site Performance: Consumer Reaction to a Poor Online Shopping Experience,” Akamai, Whitepaper, Jun. 2006.
- [183] D. L. Richards, *Telecommunication by speech: The transmission performance of telephone networks*. Butterworths, 1973.
- [184] D. Richards and J. Hutter, “Echo suppressors for telephone connections having long propagation times,” *Electrical Engineers, Proceedings of the Institution of*, vol. 116, no. 6, pp. 955–964, 1969.
- [185] E. Riesz and E. Klemmer, “Subjective evaluation of delay and Echo-suppressors in Telephone Communications,” *Bell System Technical Journal*, vol. 42, pp. 2919–2941, Sep. 1963.

- [186] G. Rubino, “Quantifying the quality of audio and video transmissions over the internet: the PSQA approach,” in *Design and Operations of Communication Networks: A Review of Wired and Wireless Modelling and Management Challenges*. Imperial College Press, 2005.
- [187] H. Sacks, E. Schegloff, and G. Jefferson, “A simplest systematics for the organization of turn-taking for conversation,” *Language*, vol. 50, pp. 696–735, 1974.
- [188] J. Saliba, A. Beresford, M. Ivanovich, and P. Fitzpatrick, “User-perceived quality of service in wireless data networks,” *Personal Ubiquitous Comput.*, vol. 9, no. 6, pp. 413–422, 2005.
- [189] B. Sat and B. W. Wah, “Analyzing voice quality in popular voip applications,” *IEEE MultiMedia*, vol. 16, no. 1, pp. 46–59, 2009.
- [190] P. R. Selvidge, B. S. Chaparro, and G. T. Bender, “The World Wide Wait: Effects of Delays on User Performance,” *International Journal of Industrial Ergonomics*, vol. 29, no. 1, pp. 15 – 20, 2002.
- [191] S. C. Seow, *Designing and Engineering Time: The Psychology of Time Perception in Software*. Addison-Wesley Professional, 2008.
- [192] J. Shaikh, M. Fiedler, and D. Collange, “Quality of experience from user and network perspectives,” *Annals of Telecommunications*, vol. 65, pp. 47–57, 2010, 10.1007/s12243-009-0142-x. [Online]. Available: <http://dx.doi.org/10.1007/s12243-009-0142-x>
- [193] Y. X. Skadberg and J. R. Kimmel, “Visitors’ flow experience while browsing a Web site: its measurement, contributing factors and consequences,” *Computers in Human Behavior*, vol. 20, pp. 403–422, 2004.
- [194] D. Soldani, M. Li, and R. Cuny, *QoS and QoE management in UMTS cellular systems*. John Wiley and Sons, Aug. 2006.
- [195] S. S. Stevens, “On the Psychophysical Law,” *Psychology Review*, vol. 64, no. 3, pp. 153–181, 1957.
- [196] T. Stockhammer, “Dynamic adaptive streaming over http–: standards and design principles,” in *Proceedings of the second annual ACM conference on Multimedia systems*. ACM, 2011, pp. 133–144.

- [197] D. Strohmeier, S. Jumisko-Pyykkö, and K. Kunze, “Open profiling of quality: A mixed method approach to understanding multimodal quality perception,” *Advances in Multimedia*, vol. 2010, pp. 1–28, 2010. [Online]. Available: <http://www.hindawi.com/journals/am/2010/658980/abs/>
- [198] J. Stromer-Galley, “Interactivity-as-product and interactivity-as-process,” *The Information Society*, vol. 20, no. 5, pp. 391–394, 2004. [Online]. Available: <http://www.ingentaconnect.com/content/routledg/utis/2004/00000020/00000005/art00008>
- [199] A. Takahashi, A. Kurashima, and H. Yoshino, “Objective assessment methodology for estimating conversational quality in voip,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 6, pp. 1984–1993, nov. 2006.
- [200] T. Takahashi, H. Oono, and M. Radford, “Psychophysics of time perception and intertemporal choice models,” *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 8-9, pp. 2066–2074, 2008.
- [201] S. L. Thompson-Schill, K. J. Kurtz, and J. D. Gabrieli, “Effects of semantic and associative relatedness on automatic priming,” *Journal of Memory and Language*, vol. 38, no. 4, pp. 440–458, 1998.
- [202] A. Van Moorsel, “Metrics for the internet age: Quality of experience and quality of business,” in *Fifth International Workshop on Performability Modeling of Computer and Communication Systems, Arbeitsberichte des Instituts für Informatik, Universität Erlangen-Nürnberg, Germany*, vol. 34, no. 13. Citeseer, 2001, pp. 26–31.
- [203] M. Varela, “Pseudo-subjective quality assessment of multimedia streams and its applications in control,” PhD Thesis, University of Rennes 1, France, 2005.
- [204] B. W. Wah and B. Sat, “The design of voip systems with high perceptual conversational quality,” *Journal of Multimedia*, vol. 4, no. 2, pp. 49–62, 2009.
- [205] P. Watzlawick, J. H. Beavin, and D. D. Jackson, *Menschliche Kommunikation. Formen, Störungen, Paradoxien*. Bern: Verlag Hans Huber, 1969.
- [206] P. Watzlawick, D. D. Jackson, and J. B. Bavelas, *Pragmatics of human communication: a study of interactional patterns, pathologies, and paradoxes [by]*

- Paul Watzlawick, Janet Helmick Beavin [and] Don D. Jackson.* Faber, London,, 1968.
- [207] E. H. Weber, *De Pulsu, Resorptione, Auditu Et Tactu. Annotationes Anatomicae Et Physiologicae.* Leipzig: Koehler, 1834.
- [208] I. Weber and A. Jaimes, “Who uses web search for what: and how,” in *Proceedings of the fourth ACM international conference on Web search and data mining.* ACM, 2011, pp. 15 – 24. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1935839>
- [209] J. Webster and J. S. Ahuja, “Enhancing the design of web navigation systems: the influence of user disorientation on engagement and performance,” *MIS Quarterly*, vol. 30, no. 3, pp. 661–678, 2006.
- [210] H. Weinreich, H. Obendorf, E. Herder, and M. Mayer, “Not quite the average: An empirical study of web use,” *ACM Transactions on the Web (TWEB)*, vol. 2, no. 1, pp. 1–31, 2008.
- [211] B. Weiss, S. Moller, A. Raake, J. Berger, and R. Ullmann, “Modeling call quality for time-varying transmission characteristics using simulated conversational structures,” *In: Acta Acustica united with Acustica*, vol. 95, no. 6, pp. 1140–1151, 2009. [Online]. Available: <http://www.ingentaconnect.com/content/dav/aaau/2009/00000095/00000006/art00018>
- [212] B. Weiss, D. Guse, S. MÄšller, A. Raake, A. Borowiak, and U. Reiter, “Temporal development of quality of experience,” in *Quality of Experience*, ser. T-Labs Series in Telecommunication Services, S. MÄšller and A. Raake, Eds. Springer International Publishing, 2014, pp. 133–147. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-02681-7_10
- [213] G. M. Wilson and M. A. Sasse, “Do Users Always Know What’s Good For Them? Utilising Physiological Responses to Assess Media Quality,” in *In: The Proceedings of HCI 2000: People and Computers XIV - Usability or Else! (HCI 2000).* Springer, 2000, pp. 327–339.
- [214] S. Winkler and P. Mohandas, “The evolution of video quality measurement: From PSNR to hybrid metrics,” *IEEE Transactions on Broadcasting*, vol. 54, no. 3, pp. 660–668, Sep. 2008. [Online]. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4550731>

- [215] A. K. Wong, “A literature review of the impact of flow on human-computer interactions (hci)—the study of a fundamental ingredient in the effective use of computers,” in *Proceedings of the IAMB 2006 conference*, 2006.
- [216] L. Yamamoto and J. Beerends, “Impact of network performance parameters on the end-to-end perceived speech quality,” in *In Proceedings of EXPERT ATM Traffic Symposium*, 1997.
- [217] Zona Research, “The economic impacts of unacceptable web-site download speeds,” Zona Research, Tech. Rep., April 1999.
- [218] H. Zourrig and J.-C. Chebat, “Waiting in a queue with strangers and acquaintances: An integrative model of customer-to-customer interactions effect on waiting time evaluation,” *International Journal of Quality and Service Sciences*, vol. 1, pp. 145 – 159, 2009.