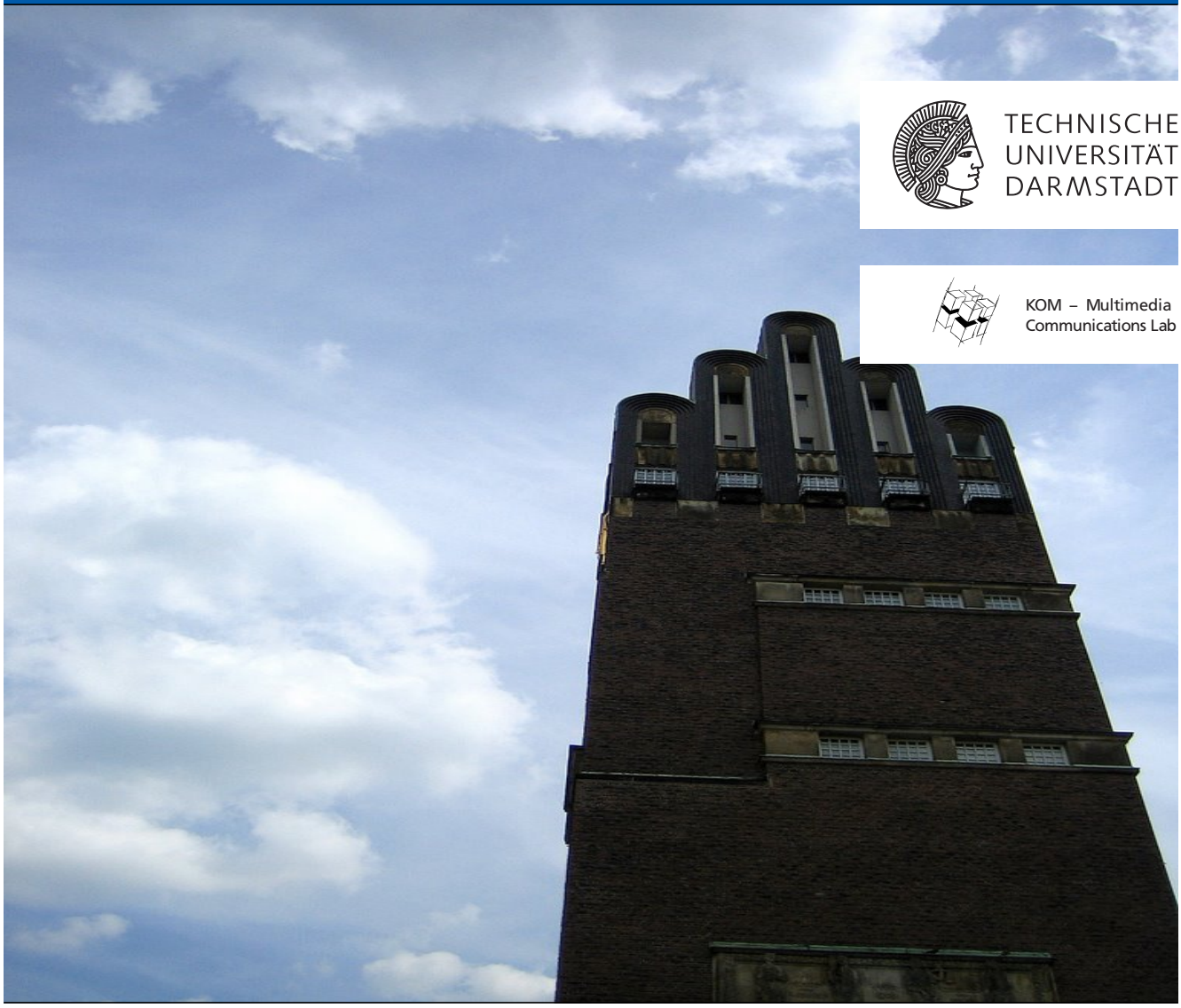
Maxim Graubner, Parag Mogre, Ralf Steinmetz:
A New QoE Model, Evaluation Method and Experiment Methodology for Broadcast Audio Contribution over IP. no. KOM-2010-2, March 2010.
ftp://ftp.kom.tu-darmstadt.de/pub/TR/KOM-TR-2010-02.pdf KOM-TR-2010-02.

A New QoE Model, Evaluation Method and Experiment Methodology for Broadcast Audio Contribution over IP

Maxim Graubner, Parag S. Mogre, Ralf Steinmetz
Technical Report – KOM-TR-ACIP



The documents distributed by this server have been provided by the contributing authors as a means to ensure timely dissemination of scholarly and technical work on a non-commercial basis. Copyright and all rights therein are maintained by the authors or by other copyright holders, not withstanding that they have offered their works here electronically. It is understood that all persons copying this information will adhere to the terms and constraints invoked by each author's copyright. These works may not be reposted without the explicit permission of the copyright holder.

A New QoE Model, Evaluation Method and Experiment Methodology for Broadcast Audio Contribution over IP

Maxim Graubner, Parag S. Mogre, Ralf Steinmetz

Technical Report – KOM-TR-ACIP
http://www.kom.tu-darmstadt.de

First published: April 1, 2010
Last revision: March 29, 2010

For the most recent version of this report see
ftp://ftp.kom.tu-darmstadt.de/TR/KOM-TR-ACIP.pdf

Technische Universität Darmstadt
Department of Electrical Engineering and Information Technology
Department of Computer Science (Adjunct Professor)

Multimedia Communications Lab (KOM)
Prof. Dr.-Ing. Ralf Steinmetz

Contents

1 Introduction

Recently, the *European Broadcasting Union (EBU) Network Technology Management Committee (NMC)* published a technical specification for professional audio contribution over networks using *Internet Protocol (IP)* technology with the aim of achieving interoperability between devices of different manufactures and enabling the sharing of infrastructure among member broadcasters [57] [81]. These are necessary requirements for the functional capability of *Audio Contribution over IP (ACIP)* [49]. The framework developed by the EBU N/ACIP project group includes well known coding algorithms (i.e. MPEG1/2 Layer2 [59]), signaling standards (SIP/SDP) and transport protocols (RTP/UDP) for real-time communication services over IP [90] [60].

Up to now, public broadcasters established their audio contribution links mostly over synchronous circuit-switched ISDN systems or using connection-oriented but packet-switched *Asynchronous Transfer Mode (ATM)* systems [53] with *Quality of Service (QoS)* guarantees [8]. These are to be substituted by IP-based networks in the future and certain QoS provisioning will be required [6] [11]. In general, modern radio broadcasters wish for versatile and reliable ways of professional audio transport [87], capable of carrying out interactive communications which are explicitly dedicated for broadcasting to the radio listener.

Especially for the most challenging audio contribution application, an ad-hoc connection to remote places, mostly no IP-service with ISDN/ATM-equivalent QoS (i.e. guaranteed bandwidths and delay bounds) will be available in the near future or some service guarantees will be too costly in comparison to widely available Internet access possibilities like standard xDSL or wireless solutions (i.e. UMTS, WiMAX). Primary users such as journalists tend therefore to also use best-effort services without QoS guarantees, accepting some possible degradation of service quality with respect to well-managed networks [49]. As such, the impact of missing or restricted service guarantees should be compensated as much as possible with intelligent network parameter trade-offs and application layer approaches. Thereby the user satisfaction in terms of *Quality of Experience (QoE)* is the final measure of interest and therefore it should be maximized [76].

In this work, the challenge of finding a QoE prediction framework for professional audio communication will be addressed. With building a QoS model incorporating evaluation-based objective QoE estimates and defining relations of adequate quantitative parameters with respect to qualitative statements, a non-intrusive application-dependent assessment of the transmission quality is made possible (as with the *E-model* for telephone networks [41]). This enables the rating of the potentials of different network and application implementations and principles; as well as providing an analysis framework for cross-layer optimization [78] of suitable parameters for gains in transmission quality (i.e. for pre-transmission set-up optimization). Furthermore, such a framework (ideally simplified as much as possible for reducing the computational effort) can also be used in real-time quality monitoring. This enables the network adaptive selection of *perceptually optimal* parameter values on different layers. For example a variable playout buffer size, adapted to assessed network jitter [84], must take maximally tolerated latency into account and allow for a certain audio quality degradation (depending on the importance of a low-delay connection). This already depicts a perceptually motivated multivariate cross-layer optimization problem, whose solution results in a *perceptually optimal* trade-off between delay and audio quality, respective the associated conversational and listening QoE, while other facts such as costs also need to be be considered.

To the author's knowledge no work on a quality prediction for ACIP applications in conjunction with a QoS optimization nor the approach taken in this thesis exists. Similar perceptually driven approaches for *Voice communication over IP (VoIP)* or general multimedia streaming can not be adapted directly because of the different characteristics and requirements of the applications (see fig. 1.1 for a preliminary general

classification). Nevertheless, approaches from other multimedia applications can generally give design proposals and important hints, while the dedicated ACIP approach can also be useful for the expected evolution of narrowband VoIP to broadband *audio* communication over IP. This trend was predicted i.e. in [24].

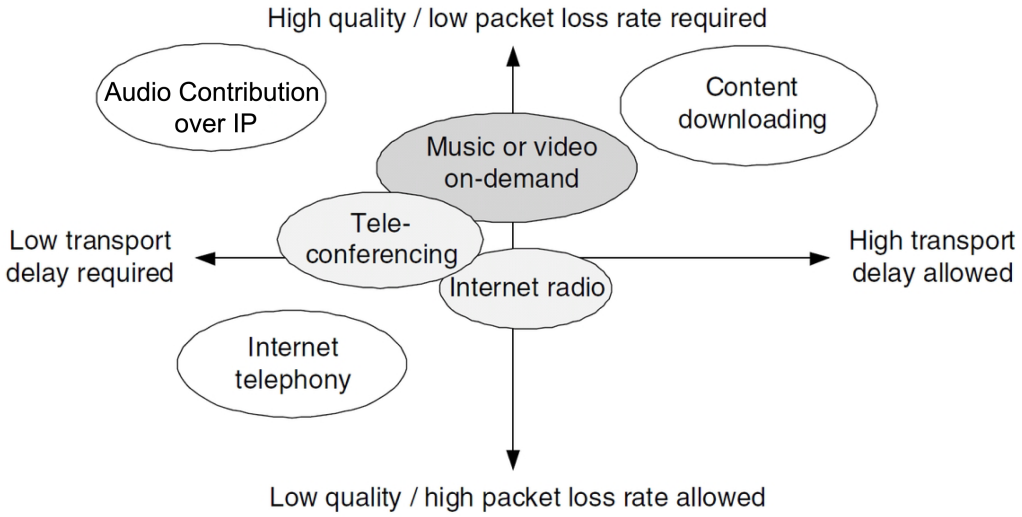


Figure 1.1: Audio content delivery types and their principal quality requirements on IP networks, based on [48].

The main difference here in terms of listening quality between well known voice communication over IP and audio communication over IP is the used audio bandwidth. The audio bandwidth for telephony/VoIP of 3 kHz narrowband and 7 kHz wideband extends to above 15 kHz for broadband audio communication. In terms of conversational quality, professional audio communication has stricter demands on latency. Only in exceptional cases would a one-way delay of up to 400 ms as allowed for VoIP be accepted, i.e. for a catastrophe report.

Unlike most other multimedia services, i.e. television broadcasting, radio broadcasting has a problem with this delay value. For radio broadcasting, a high latency in remote conversations results in silence on the transmission channel while today's radio always tries to have a mostly uninterrupted "flow" in its final audio transmission. This means, a distortion of the "rhythm" of audio elements (i.e. music, reports, jingles) needs to be avoided. In television broadcasting, this silence is more accepted because if they can deliver a picture, the end-user is willing to accept waiting, i.e. in a news program a remote reporters answer can be waited upon because the viewer realizes the upcoming continuation.

In the following, the main use cases for professional audio contribution over IP are discussed and professional audio quality, also referred to as *studio quality*, is explained before an initially classification of ACIP with respect to VoIP is examined. There after, the contributions of the presented thesis are surveyed. The introductory chapter finishes with an outline of the thesis.

1.1 Audio Contribution Use Cases

In order to formulate the requirements of an application as well as the QoS grading and its improvement on the different layers of the transmission system, a clear classification of the application use-cases have to be specified. Only then an application-specific QoS model can be derived and an application-oriented traffic and performance analysis is made possible.

Professional audio contribution has a variety of different use cases and possible configurations. In general, a contribution in this context refers to an exchange of broadcasting audio material, normally from remote sites or local offices to main studio centers and reverse [81]. Two main methods exist to accomplish these tasks, either

1. *file transfer* (non-real-time store and forward) or
2. *live contributions* (streamed).

In this work, only live contributions are regarded. As in [58], on the *operational level* this method can primarily be classified into the main use cases of audio contribution, which are

- *outside broadcasts* (i.e. news contribution, sports commentary, concerts live-feeds),
- *interviews* (i.e. two people conversing between two separate locations),
- *discussions, talk shows, roundtable* (i.e. many contributors, multiple locations).

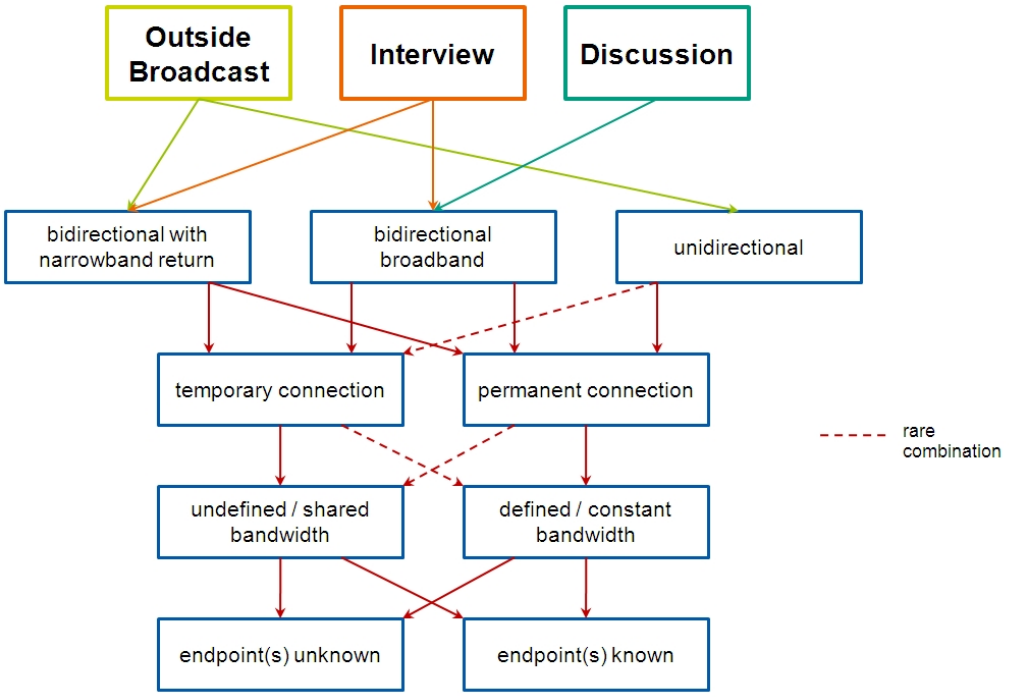


Figure 1.2: Overview on audio contribution use cases and their classification.

These can be used to make a differentiation in terms of user demands on the Quality of Service of an ACIP infrastructure, i.e. the amount of interactivity determines the end-to-end delay bounds on the system, while the particular audio quality requirements are related to network bandwidth demands and the choice of audio coding algorithm. The general requirements of the different applications on one-way latency and audio bandwidth (as one of the main indicators for audio quality) can be surveyed in table 1.1.

The different use cases on the operational level operate in different types of *audio contribution modes* [50]. The modes can be identified as

- *unidirectional*: transmission with no return channel (i.e contribution by satellite),
- *bidirectional with narrowband return*: transmission where the return channel is narrowband (i.e. telephone quality) and for the purposes of cueing the contribution (i.e. for sports reports) and

QoS Parameter	Outside Broadcasts			Interviews	Discussions
	Concerts	Sports	News		
One-way Delay [ms]	<500 contribution, <50 talkback			<100	<100
Audio Bandwidth [kHz]	15-20	7-15	3.5-12	7-12	12-20

Table 1.1: Operational requirements for different audio contribution use cases as to [50].

- *bidirectional broadband*: transmission with bidirectional broadcast quality audio (i.e. interview).

For these operational modes two types of connections can be used [58], either

- *permanent connections* (generally on managed networks with QoS guarantees) or
- *temporary connections* (may be based on previously-unknown networks without QoS guarantees).

It is of general importance that the audio coding devices at the endpoints are *known* by their IP address. But if the coding devices are *unknown*, they must ask a central entity (normally a SIP-Server) to deliver the actual IP address of their counterpart. The described relations can be regarded in fig. 1.2.

In this work the focus is on the most challenging application: a bidirectional broadband connection to remote places (i.e. a temporary connection). Finally it has to be mentioned for the sake of completeness, that there is also the operational class *multicast* (point-to-multipoint transmission) quoted in [58], which is dedicated to the distribution of programs but not an ordinary contribution application. Therefore, this category is neglected here.

1.2 Studio Quality

Here an insight in general professional audio quality requirements will be given. The fulfillment of these requirements results in so called *studio quality*. Thus, in order to reach the desired best possible sound reproduction, an optimal audio quality must be kept throughout the professional sound processing chain. This begins with the recording of a sound source. The quality of this gathered audio signal then needs to be maintained over all processing steps including digitalization and delivery to storage media or to remote places. For broadcasting, this includes the final distribution to the end-user.

Because the digitalization is among the most important steps in maintaining professional audio quality, the best possible digital representation of an analog signal is desired in professional audio engineering. Even though today's broadcasting standard for digital audio is definitely a sufficient resolution for almost all users, many audio engineers affirm that they nevertheless perceive an impairment in the current digital representation with 48 kHz sample rate and up to 24 bit quantization resolution. They would prefer an even better *Signal-to-noise Ratio (SNR)* for an even higher dynamic range, which is in the digital domain directly related to the digitization resolution [95]. For example, today's concert recordings are often A/D-converted using samplerates of 96 kHz or 192 kHz with up to 28 bit quantization. Thereby the computational effort and storage capacity is no more a big issue but rather the capabilities of the

analog components preceding the A/D-converter.

In general, normative professional audio quality can be defined as the absence of linear as well as non-linear distortions of the audio source signal. They can be introduced by a processing system or transmission system. *Linear distortion* is normally related to a non-flat frequency response in the audible range from 20 Hz to 20 kHz, while *non-linear distortion* is related to harmonic distortion, i.e. due to non-ideal amplifiers. The most important metrics for quantitative assessment of the audio quality can therefore be summarized as

- the *frequency response*,
- *Total Harmonic Distortion (THD)*, and
- *Signal-to-noise Ratio (SNR)*.

The definitions of the measures can be found in [89]. The broadcasting requirements on the metric values with respect to a system are a flat frequency response with a small deviation (i.e. up to 0.5 dB), a THD below a certain value and an SNR higher than a certain bound. THD and SNR are usually combined to a unique THD+N measure (i.e. at least -90 dBFS/dBr)[1].

In contrast, inside today's digital broadcasting systems, almost all final program material is data compressed for cost, transmission and storage efficiency using *transparent* coding such as MPEG-1/2 Layer 2 with a coding bitrate of 384 kBit/s at a samplerate of 48 kHz. Here, the term transparent refers to a perceived audio quality within which even sophisticated audio experts can not detect distortions of the original signal. The output is therefore *perceptually undistorted*. An assessment of the perceived audio quality is done with appropriate QoE metrics (see chapter **??**). Hence, the focus in audio quality assessment is relocated from quantitative measures to qualitative measures and the discussion of the needed quality for public broadcasting and related costs is deepened. This is an ongoing process because different people involved have different opinions on this issue, including journalists, audio engineers, musicians, administrators, politicians, and also the ordinary citizens enjoying the public program.

In general it can be stated that today's public broadcasters have to ensure the best possible quality at the lowest possible costs (see also chapter 2.1.3). The key in terms of quality is the absence of audible degradations, which could otherwise bring a radio listener to switch radio channels due to shoddy audio quality - this would be the worst case. With synchronous circuit-switched communication systems this was not a big problem, because dedicated lines could be specified and used. In packet-based transmissions this is more of an issue and additional questions on possible quality impairments, such as packet loss, previously unknown to broadcasting have to be addressed.

It need to be stated that audio experts such as audio engineers are traditionally aware of processing which can not directly be controlled. Therefore, automatic and adaptive approaches theoretically optimizing the audio quality are taken with scepticism and must be carefully designed, ideally considering QoE metrics derived from the subjective opinion of the professional audio experts additionally to a general optimization with respect to the characteristics of the human ear as applied in coding algorithm design.

Furthermore, broadcasting engineers are always careful to ensure processing and transmission without perceivable interruption or failure. Therefore the most important elements in broadcasting always have a physical hardware redundancy.

[1] The exemplary values are taken from the German public broadcasters rec. HFBL-K 20 RBT/AKAS: "Richtlinien für die Messung von digitalen Geräten und Anlagen in der Audio-Technik", Aug. 2001.

1.3 VoIP versus ACIP

Although a strong relationship exists between ACIP and other real-time applications like voice and video over IP, it has different characteristics and requirements. For example, call durations for audio contributions are generally longer than with *Voice over IP (VoIP)* telephony [87]. In terms of audio bandwidth, VoIP normally uses narrowband speech transmission, while ACIP requires up to the entire bandwidth ($\geq$ 20 kHz) and coding artifacts are less tolerated.

Therefore, ACIP usually also has greater network bandwidth demands in order to provide better audio quality. Furthermore, ACIP often uses different coding algorithms, normally dedicated to audio signals whereby unified speech and audio coding algorithms still under development should also be used. In contrast, VoIP uses dedicated voice coding algorithms operating at very low bitrates.

The intelligibility optimization throughout the development of VoIP systems as well as the efficiency optimization of the network service with respect to costs and resources destroys the suitability of public telephony for professional audio contribution. So far, the ISDN technology was able to provide guaranteed services as desired for broadcast audio contribution, i.e. a delay below 50 ms and virtual no losses as well as a very high availability. With VoIP, all these guarantees are lowered or may even disappear.

Failures like packet loss are accepted within a certain degree for VoIP communications as long as the intelligibility is maintained. In contrast, professional audio over IP requires a “perfect” transmission quality (without interruptions) and loss of information would have to be concealed as well as possible in terms of perception.

Delay Source	**Typical Range (ms)**
Recording	10-40
Coding	10-20
Internet delivery	70-120
Playout buffer	50-200
Decoding	10-20
Total	150-400

Table 1.2: VoIP processing steps contributing to the overall transmission delay, values taken from [64].

Especially for bidirectional live contribution a low latency is needed. Hence the difference between (time-)critical professional audio contribution tasks and related applications such as professional audio distribution or conventional multimedia streaming becomes apparent: for the latter ones a higher latency is generally tolerated and the quality requirements are not as strict as for ACIP applications. In table 1.2 an overview of the principal delay budget is given for VoIP. For ACIP, similarly optimized values do not exist so far.

2 General Issues

2.1 Quality of Service and User Perceived Quality of Service

Applications such as those based on the professional *Audio Contribution over IP (ACIP)* principles have specific requirements on the quality of the service of a communication system. The presentation of the use cases in chapter 1.1 already denoted that in principle there should be an upper limit on the delay of the transmission system as well as a specific concept for the audio quality. Both are factors which can be used for the requirement formulation on the service quality of the system based on user's expectations on the behavior of the overall communication system with respect to its required use. Furthermore, the factors indicate that a *guaranteed service* [75] is desired. The means of accomplishing this and other requirements is by applying the concepts of *Quality of Service (QoS)* and *Quality of user Experience (QoE)*.

In the following, the notion of QoS and the respective user perceived QoS (the QoE) are discussed in general, before an overview on IP QoS provisioning mechanisms is given. After the more detailed definition of real-time audio content delivery over IP, the application-oriented QoS assurance for ACIP is also treated in further detail in chapter **??**.

2.1.1 General Definitions

A useful general definition of *Quality of Service (QoS)* is given in [73] as

"QoS is the welldefined and controllable behavior of a system with respect to quantitative parameters."

This definition indicates, that for providing a particular QoS, definable and measurable parameters describing the performance of the system need to be identified. An example of one such parameter is the system delay mentioned above. Therewith concepts to ensure that a parameter does not exceed a deterministic maximum value or a statistical mean over the long term, dependent on the desired type of service, can be developed and applied. If these QoS provisioning mechanisms do a good job meeting the requirements, the end-user will certainly be satisfied. Therefore, the objective system's QoS factors are directly related to the subjective end-user's perceived QoS [94]. The initial challenge is then to determine these parameters on the different layers of a communication system (see chapter **??** for an comprehensive overview). Jointly with appropriate QoE metrics, the relationship between the user's experience and the system performance can be quantified (see chapter **??**).

Before principal performance parameters are presented, the concept of *Quality of Experience (QoE)*, the mentioned satisfaction of the user with respect to the system's Quality of Service, shall be defined, using the definition in [76]:

"QoE is the user's perceived experience of what is being presented by the application layer, where the application layer acts as a user interface front-end that presents the overall result of the individual Quality of Services."

This statement already includes the layer concept used in modern transmission systems. Traditional QoS was defined only for the network layer of a communication system. An enhancement to the QoS concept then was to induce QoS in transport services. However, for networked multimedia systems such as ACIP, the QoS system must be extended further because the end-to-end quality is of particular importance here and many services at different layers contribute to it [80]. If appropriate quantitative QoE metrics are available, the layer concept can be further extended by defining the *user perception pseudo*

layer [76] at the top of the application layer. In fig. 2.1 an overview on the different layers of a QoS layered model for real-time applications is given.

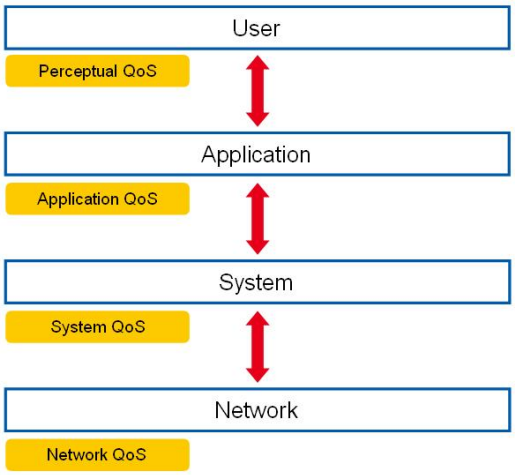


Figure 2.1: General QoS layered model for real-time applications as in [80].

2.1.2 Who is the User in Broadcast Audio Contribution?

So far, the user experiencing an audio contribution service was often mentioned without a clear definition. Generally, a telecommunications user is a human being employing a particular service. In Broadcasting, two different kinds of user need to be considered,

1. the *primary user* (i.e. a journalist or audio engineer working at a broadcasting station) and
2. the *secondary end-user* (i.e. a person listening to a radio program).

The specific audio contribution application is directly used by the primary user who applies the service, for example in order to perform an interview. His needs regarding the Quality of Experience have to be addressed, first because only with this premise can the satisfaction of the secondary user listening to the radio program be ensured. Herein, the focus of the different user levels is on different performance parameters.

A moderator and his interview partner(s) place a particularly high importance on the intelligibility and a minimal end-to-end delay for their conversation, for example partaking in a live discussion. The end-user benefits from the primary user's requirements because he is thereby able to enjoy a flowing conversation with potentially fast interposed questions and reactions which in turn increase the merit of the broadcast communication.

The listening quality in such a situation is of particular importance for the end-user's satisfaction, however it can also increase the conversational quality for the primary users if they have a more realistic impression of the voice of the other people involved. Further issues are the potential archiving requirements of public broadcasters, i.e. if a historically relevant discussion was transmitted and needs to be conserved for the future. For this, they rely on storing the best possible quality, preferably better then the usually bandlimited final audio broadcasting signal (i.e. FM transmission is limited to 15 kHz audio bandwidth). Another user level could theoretically be introduced at this stage, but for the sake of convenience, this can be neglected.

Finally it needs to be mentioned, that for political reasons deterministic bounds on the performance factors may be important for public broadcasters. For example, in a transmitted dispute between political opponents located at separate remote places, the discrimination of a participant due to missing QoS needs to be avoided. As a one-sided higher delay could result in the reduction of auditory response speed of only one of the contenders, thereby disregarding the legal requirements of equal opportunity in free speech as laid out for public broadcasters.

2.1.3 QoS for Audio Communication over IP

Until now, there was a lot of discussion about quality aspects which now need to be addressed in more detail with respect to audio communication over IP. Therefore, a general definition on quality is introduced as in [55]:

"Quality can be regarded as a point where the perceived characteristics and the desired or expected ones meet."

The discussion of the different user levels also introduce different attributes of end-to-end *conversational quality*. Following [69], they can principally be differentiated into

- *listening quality*,
- *interaction quality* and
- *talking quality*.

The listening quality and interaction quality are the most important attributes to be considered because for talking quality the main influence is a possible acoustic echo which is negligible. This is justified due to the fact that mainly professional audio applications are addressed which allows the assumption that ideal communication end-devices are used. Additionally, the echo problem is no longer as important as it was, especially since adequate *Acoustic Echo Cancellation (AEC)* is ever increasingly used. Following from this, the talking quality is henceforth neglected.

In order to define the network QoS parameters in the next section, the general factors which affect the conversational quality in digital audio communication needs to be determined now. In conventional telephony noise was the main problem to consider. Today's packet-switched networks deploy new coding technologies and new transmission technologies [85], in which many more and differentiated impairments are present, such as larger coding distortions, packet loss and increased delay. The general factors affecting the conversational quality of audio communication over IP can therefore be summarized as follows,

- the *audio coding algorithm configuration* (i.e desired audio bandwidth and coding bitrate),
- *transmission channel impairments* (i.e. packet loss on an IP network), and
- *end-to-end delay* (i.e. coding and network transmission delay).

While the audio coding algorithm configuration and possible transmission channel impairments mainly affect the listening quality, the end-to-end delay is the main indicator for interaction quality.

Above, the most important quality elements were defined. While the quantitative assessment of the perceived quality by appropriate QoE metrics with respect to the quality factors is addressed in chapter **??**, the great variation in the perceived satisfaction of the user with audio communication needs to be

mentioned here, also addressing the influence of the service costs on the QoE. As obtained from [12], the user demand on the conversational quality in audio communication can vary anywhere between

- the *minimum required quality* to allow intelligible voice communication and
- the *highest possible quality* towards audio fidelity for aesthetic enjoyment,

while besides the communication requirements, the user satisfaction with respect to audio communication can be influenced by many more facts, i.e. also the costs. If an Internet telephony application is freely usable without additional costs to an flat rate, the user would be with high probability satisfied if he can use it with the minimum requirements to have an principal intelligibility. On the other side, if someone is paying much for enjoying a cinema movie with an realistic experience of the sound scenery, he will have the expectation that at least a multi-channel reproduction of the sound is offered to satisfy the demands .

In public broadcasting, normally the end-users directly paying the offered broadcasting services over dues or taxes. Therefore, public broadcasters are forced to provide the best possible quality on adequate costs to ensure their acceptability and usability for the general public.

2.1.4 IP Network QoS Parameters

Now, the specific general network QoS parameters of an IP-based communication system are identified before introducing IP QoS provisioning mechanisms. The network QoS parameters are derived from the general QoS demands of audio communication above, following the definition for network QoS stated in [94] as

"A network can be said to provide QoS if it is is able to meet the need of the end users' applications in a satisfactory way."

Finally, the main IP network QoS parameters for audio communication applications can be quantified as

- *available network bandwidth*,
- *delay* and *delay variation (jitter)*,
- *IP packet loss ratio*.

Thereby the network bandwidth is chosen because it limits the maximal usable coding bitrate of the audio coding algorithm, the delay and jitter determines the network contribution to the end-to-end delay while packet loss is the main transmission channel impairment in IP networks. Therewith, the claimed relation of network parameters to application factors based on user demands presented in the previous chapter 2.1.3, is established. The specific importance and assessment of QoS parameters for audio contribution over IP is discussed later in chapter **??**. Besides the above mentioned parameters, the service availability, reliability and security shall also be mentioned as important factors generally affecting the perceived QoS.

2.2 Assessment of User Perceived Quality of Service

The end-to-end quality of service of a transmission system for audio communication over IP can be described by the appropriate assessment of quantitative QoS parameters. In contrast, the perceptual QoS

with respect to the subjective impression respective satisfaction of a user is only assessable by subjective auditory testing or instrumental approaches which model the estimated subjective opinion [55]. Thereby, the challenge of assessing the audio quality of user experience (QoE) can be "as comprehensive as human needs and imagination" [12] and the selection of a specific method need to be done carefully because of the high dependence of QoE metrics on the application and environmental conditions as well as user expectations.

For communication applications in general, metrics for describing the user perceived quality of service were developed. They usually refer to two different quality categories (see chapter 2.1):

- the *listening-only quality* or
- the *conversational quality* at all.

The most commonly used measure is the *Mean Opinion Score (MOS)* which was standardized in the ITU-T recommendation P.800 [33] for the subjective determination of transmission quality. It is defined as to [38] in the following way:

"The mean of opinion scores, i.e., of the values on a predefined scale that subjects assign to their opinion of the performance of the telephone transmission system used either for conversation or for listening to spoken material."

Apart from subjective opinion, the mean opinion score is also used for ratings that originate from instrumental models as objective ones or network planning models (see below for further explanation). To better identify to which quality category and assessment method a MOS value refers, a MOS terminology was defined in ITU-T recommendation P.800.1 [38]. In table 2.1 the different identifiers are listed. The different MOS types get suffixes with respect to the quality category ("LQ" for listening, "CQ" for conversational quality) and assessment method ("S" for subjective assessment, "O" for objective assessment and "E" for an estimate based on a quality prediction model as the E-model presented in chapter 2.2.3).

	Listening-only	**Conversational**
Subjective	MOS LQS	MOS CQS
Objective	MOS LQO	MOS CQO
Estimated	MOS LQE	MOS CQE

Table 2.1: Mean Opinion Score (MOS) terminology from ITU-T rec. P.800.1 [38].

The different metrics in table 2.1 can be further discriminated with respect to the audio bandwidth of the transmission system. This is important, because a user would rate the quality of a narrowband transmission generally different, dependent on the availability to use a larger bandwidth instead. For example, a narrowband codec would get a poorer rating if it is directly compared to wideband transmission (see also chapter 2.2.3). Therefore, the ITU-T recommends to use another MOS suffix to discriminate between a narrowband ("n") or wideband ("w") transmission system under test. These additional suffix is neglected in this work for convenience even more because mostly the quality assessment with respect to fullband transmission systems is regarded, which is not standardized yet.

For the *opinion rating scale* of a QoE metric, different approaches exist. In table 2.2 the two most important five-grade scales are depicted. Herewith, the left one (a) is based on the philosophy that someone can easily rate the perceived *quality* between "excellent" and "bad", more dedicated to consumer perception. The right one (b) instead refers to the amount of *impairment* perceived by the user mostly with

respect to a reference which is more used for experts analysis.

(a) Quality Rating		(b) Impairment Rating	
5	Excellent	5.0	Imperceptible
4	Good	4.0	Perceptible but not annoying
3	Fair	3.0	Slightly annoying
2	Poor	2.0	Annoying
1	Bad	1.0	Very annoying

Table 2.2: Opinion scales defined by (a) ITU-T recommendation P.800 [33] and (b) ITU-R recommendation BS.562-3 [25].

The quality rating scales presented in table 2.2 are defined in the ITU-T recommendation P.800 [33] (table 2.2 (a)) for speech and in the ITU-R recommendation BS.562-3 [25] (table 2.2 (b)) for audio quality rating. Furthermore, many other application-specific rating scales exist. For example, television picture quality rating is often performed on the *Continuous Quality Scale (CQS)* defined in ITU-R recommendation BT.500 [28]. This is also a five-grade scale but with a numerical representation in the range of $[0, 100]$ (100 best), which is divided into five equal intervals with the same descriptions as for the common quality rating. These was also chosen for subjective auditory tests based on the MUSHRA method (see below) developed by the EBU and standardized in ITU-R recommendation BS.1534-1 [29], because the developers expected a more precise rating using CQS.

For subjective tests based on the methodology defined in ITU-R Recommendation BS.1116-1 [26], the impairment rating is normally expressed in terms of the *Subjective Difference Grade (SDG)*, which can be categorized as a *Difference Mean Opinion Score (DMOS)*[1] [86]. The SDG scale is defined in the ITU-R Recommendation BS.1284-1 [30].

Impairment Rating	Grade	SDG / ODG
Imperceptible	5.0	0
Perceptible but not annoying	4.0	-1.0
Slightly annoying	3.0	-2.0
Annoying	2.0	-3.0
Very annoying	1.0	-4.0

Table 2.3: Relation of Subjective Difference Grade (SDG) respective Objective Difference Grade (ODG) to the absolute impairment rating [30].

The counterpart of the SDG used as score for the instrumental perceptual evaluation of audio quality (PEAQ [27]) is the *Objective Difference Grade (ODG)*. The relation of the absolute impairment rating to the DMOS metrics can be seen in 2.3. With the absolute impairment rating $MOS^{(i)}$ both metrics are defined as

$$ODG = MOS^{(i)}_{SuT} - MOS^{(i)}_{ref} \ , \tag{2.1}$$

[1] For avoiding confusion: the term DMOS is also used for the degradation mean opinion score in the literature [33] while here it shall always refer to the difference mean opinion score as in [86] or [45].

where $MOS_{SuT}^{(i)}$ is the impairment rating with respect to the *signal under test (SuT)* and $MOS_{ref}^{(i)}$ is the impairment rating with respect to an *reference signal*. For convenience in eq. (2.1) only the ODG is mentioned.

Another concept is exploited for the *R*-factor opinion scale of the E-model (see chapter 2.2.3). Here also a range of $[0, 100]$ is chosen but the user perceived quality is directly associated with the user *satisfaction*, i.e. a value of 100 indicates the theoretically best perceived quality by the term "users are very satisfied". The relation of *R*-factor, MOS CQE (in terms of speech transmission quality) and user satisfaction are depicted in table 2.4 following the definition in ITU-T Recommendation G.109 [34]. In general, Different quality scales are hard to compare or to translate. Only in some cases it is possible to clearly define a transformation relation.

R-factor Range	MOS CQE	User satisfaction
90 < R < 100	Best	Very satisfied
80 < R < 90	High	Satisfied
70 < R < 80	Medium	Some users dissatisfied
60 < R < 70	Low	Many users dissatisfied
50 < R < 60	Poor	Nearly all users dissatisfied

Table 2.4: Relation of *R*-factor, MOS CQE and user satisfaction as to ITU-T Recommendation G.109 [34].

In the following, first methodologies for the subjective evaluation of audio quality are surveyed before actual instrumental methods are discussed. Thereby, it is differentiated between objective methods for speech or audio quality assessment which analyze a signal under test by comparison with a reference signal attempting to predict the perceived quality of the test signal and methods which use knowledge of the transmission system to predict a system-level quality [91].

2.2.1 Subjective Listening Tests

Normally, a MOS value is obtained from subjective listening tests. Thereby an appropriately trained group of people listens to specific test signals possibly impaired by a system under test. The test subjects give gradings of their perceived quality impressions on a quality or impairment scale. Then, the arithmetic mean of all the opinion scores collected with respect to a specific system configuration is the mean opinion score per definition (see above) [83]. The tests have to be carried out under controlled conditions, i.e. by using calibrated presentation equipment such as loudspeakers or headphones [69].

In general, two different approaches exist for the subjective QoE assessment of speech [85] as well as audio [13]. For speech signals, the opinion rating defined in ITU-T recommendation P.800 [33] is most widely used. The two different methodologies defined for speech quality rating are

- *Absolute Category Rating (ACR)* and
- *Degradation Category Rating (DCR)*,

while with ACR an absolute rating with respect to one single test signal is obtained and with DCR the rating is performed relative to the knowledge of a reference signal. So called *single-stimulus methods* [13] as ACR are extensively used for speech quality rating while for audio tests mostly paired or *multi-stimulus tests* comparable to DCR are used.

The most famous methodologies for subjective audio quality assessment are the ITU-R Recommendation BS.1116-1 [26] for the rating of small impairments and ITU-R recommendation BS.1534-1 [29] for intermediate quality rating, i.e. low-bitrate coding algorithm evaluation [18]. BS.1534-1 results from a proposal of the EBU project group B/AIM (*Audio In Multimedia*) [82] and was named MUSHRA (*MUlti Stimulus Test with Hidden Reference and Anchors*) while the name already includes the description of the test method. Besides some differences in the test procedure, it shall be mentioned that the two ITU-R recommendations use different opinion rating scales. While MUSHRA adapted the Continuous Quality Scale (CQS), BS.1116-1 uses the five-grade impairment scale (see above). More detailed information on subjective listening tests can be found in [3].

Until now, only the subjective quality assessment by listening tests for rating equipment impairments was discussed. For the sake of completeness, also the subjective assessment of the whole conversation quality (CQS) shall be mentioned. The ITU-T specifies also MOS CQS assessment methods in ITU-T recommendation P.800 [33]. Thereby, conversation quality tests are even more complicated and extensive than listening-only tests [20].

2.2.2 Objective Evaluation

Because listening tests are time-consuming, expensive and limited in scope, instrumental models for assessing the listening quality of speech and audio signals as well as the conversational quality were derived. Nevertheless, they always shall rely on careful calibration against listening test results.

These objective models predict the user perceived quality with respect to communication applications from physical quality parameters and psycho-acoustical metrics, taking into account the subjective nature of human perception. Traditionally, the *Signal-to-noise Ratio (SNR)* was a good indicator for the impairment of a signal because noise was the main problem. But nowadays using packet-switched networks, much more different impairments are present, as packet-loss and higher coding distortions. These are assessed by the psycho-acoustical based measures which are combined in a computational model to get a MOS-like rating output. Objective models have the big advantage to be automated and that their evaluations are repeatable [71].

In the following, instrumental models for speech and audio quality assessment shall be categorized and then the most important approaches, also used in this thesis, are surveyed in further detail. In fig. 2.2 an overview on QoE assessment methods for speech and audio is given accounting for their calibration with subjective tests as to [7]. The survey is restricted to the most actual methods respective the ones used in this thesis. The classification seen there is now further explained.

In general, objective methods can be distinguished in *intrusive* and *non-intrusive* approaches. Thereby signal-based methods as the *Perceptual Evaluation of Speech Quality (PESQ)* algorithm and the *Perceptual Evaluation of Audio Quality (PEAQ)* algorithm are classified as intrusive because they operate *full-reference (FR)* and need as input for the evaluation the possibly degraded output signal of the system under test as well as the undistorted reference signal. Then, they perform a sequence comparison between the signals and result a MOS-like score. Both algorithms are explained in further detail below.

In contrast, non-intrusive models perform their QoE estimation without knowledge about the reference signal (*non-reference, NR*). Alternatively, they can have obtained some information about the character of the signal (i.e. simply the SNR value) and operate therefore with *reduced-reference (RR)*. Non-intrusive approaches are normally parameter-based as the E-model for transmission planning (see below). They estimate the quality using quantitatively measurable parameters, i.e. ordinary packet-header informa-

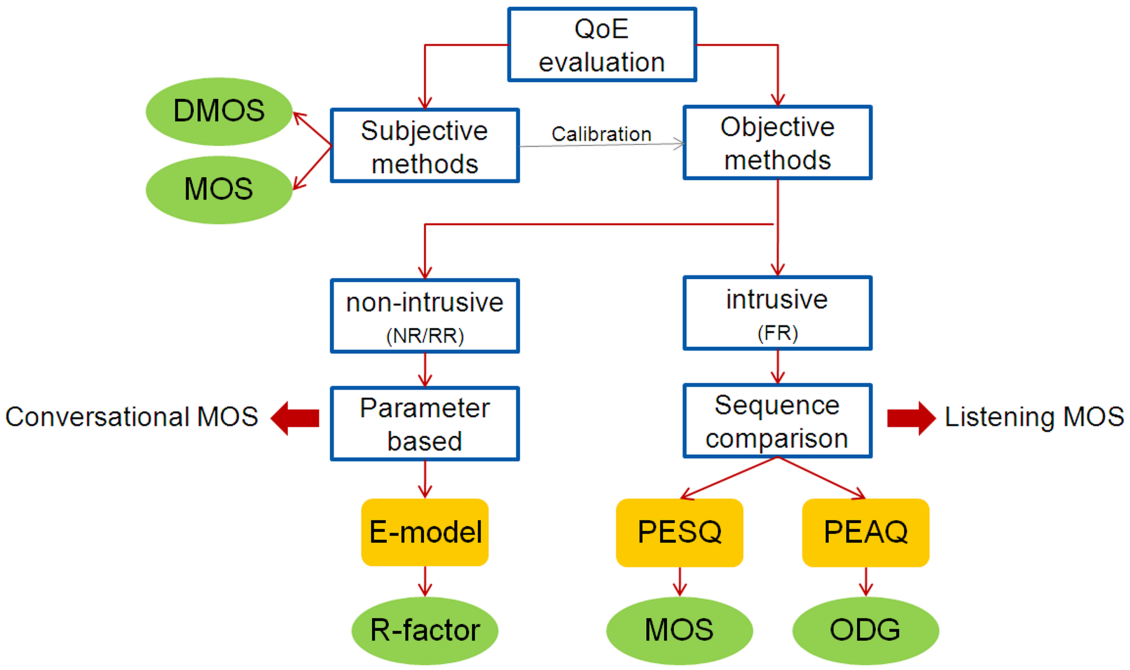


Figure 2.2: Classification of instrumental QoE assessment methods for speech and audio.

tion.

While signal-based methods are restricted to the listening-only quality prediction, parameter-based models as the E-model enable the assessment of the whole conversational quality. In table 2.5 an overview on more ITU standards for instrumental QoE assessment of speech and audio is given for the sake of completeness. Furthermore, the hybrid-model approach is added which is also the objective of this thesis.

Category	**Media-Layer** Model	**Parametric Packet-Layer Model**	**Parametric Planning Model**	**Hybrid Model**
Input information	Media signal	Packet header information	Design parameters	Combination of any
Primary application	Quality benchmarking	In-service monitoring	Network planning	In-service monitoring
Estimated quality	Listening MOS	Listening MOS	Conversational MOS	Combination of any
ITU standards	ITU-R BS.1387 (PEAQ) ITU-T P.862 (PESQ)	ITU-T P.564 (Speech)	ITU-T G.107 (E-model)	IPTV project ITU-T J.bitvqm
Output metric	DMOS/MOS	MOS	R-factor	Open
Procedure	Full reference	Non-reference	Reduced reference	Any

Table 2.5: Instrumental QoE assessment methods as to [86].

The Media-Layer Models PESQ, WB-PESQ, PEAQ and PEMO-Q

Diverse application-oriented psycho-acoustical models for objective estimation of respective listening tests outputs are available. The most recent ITU standards were used for the present thesis: PESQ, dedicated to speech quality evaluation (ITU-T recommendation P.862 [35]), and PEAQ, dedicated to audio quality evaluation (ITU-R recommendation BS.1387-1 [27]). Both methods were proposed in the 1990's combining ideas from several previous methods [18] [4]. Furthermore, the not standardized but powerful new development PEMO-Q, proposed in 2003 from [22], was used. For the PESQ algorithm, the wideband extension was employed, standardized in ITU-T Recommendation P.862.2 [40] in 2005. In the following, these instrumental QoE assessment tools shall be presented.

First, the general intrusive modeling approach shall be surveyed. In fig. 2.3 an overview of the most important system components is given. In principle, all of the mentioned models use psycho-acoustically motivated *perceptual models*, often referred to as *ear models* [46] or *auditory models* [23], which shall incorporate the perceptual behavior of the human ear with respect to specific sound signal characteristics, i.e. speech or audio as music. The structure of this model is similar to the structure of transform-based audio coding schemes [27].

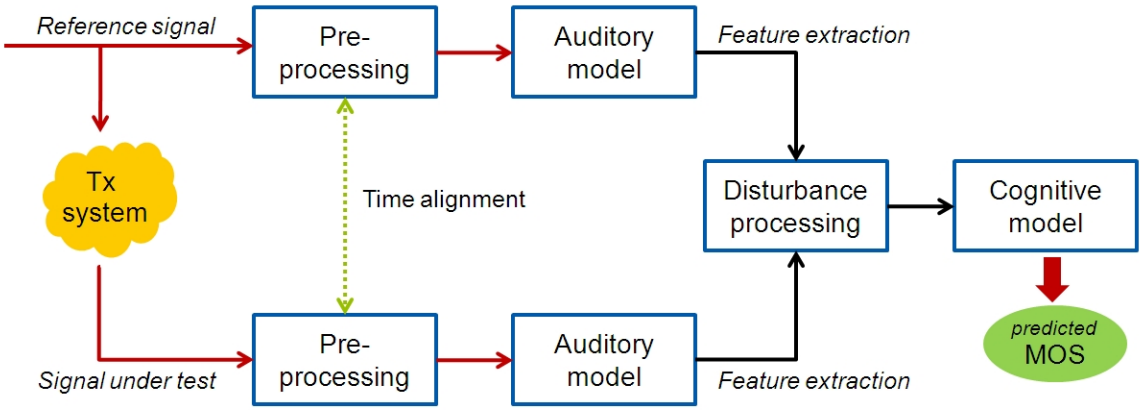


Figure 2.3: Principle stages of processing for signal-based objective audio quality assessment.

In general, a perceptual model is fed with the original and the distorted signal whose degradation with respect to QoE shall be determined, while often a *preprocessing* is incorporated before performing the common transformation in a time-frequency representation of the signals for the perceptual evaluation. This transformation is justified because also the real human ear performs a time-frequency transformation [4].

Furthermore, a *cognitive model*, evaluating the perceptual model outputs describing the disturbance, is used [69]. Here, usually a linear combination of the model output values is performed as in PESQ, while the more advanced approach of using an artificial neural network was exploited for PEAQ [5] and PEMO-Q [22]. The resulting scalar output is the QoE rating metric of the method which is an estimate of the perceived QoS evaluated with a corresponding subjective test.

Perceptual Evaluation of Speech Quality (PESQ)

This method was accepted by the ITU-T in 2001. It is an objective speech quality model for telephone applications proposed in [67], releasing an QoE metric comparable with the mean output of ITU-T recommendation P.800 subjective tests performed with the ACR method (see above). The computational model is able to predict the subjective speech quality rating in a wide range of conditions, that may

include coding distortions, errors (also packet loss), noise, filtering, delay and variable delay [4].

PESQ assumes listening through a narrowband telephone handset. Therefore it includes in the preprocessing an *Intermediate Reference System (IRS)* filtering with the characteristics of an narrowband limited telephone handset. Furthermore, the signals are time-aligned accounting for a possible delay and delay variation between the input signals. Also a level alignment is performed in the preprocessing [68].

The time-frequency transform for the perceptual model is performed with a windowed short-term *Fast Fourier Transform (FFT)*. The model output parameters are the computed *average disturbance value* and the *average asymmetrical disturbance value* where the later one was incorporated because of the so called asymmetry effect caused by non-linear coding distortions in the frequency domain. From the both measures the final MOS-like *PESQ score* MOS_{PESQ} is computed by linear combination (the cognitive model), which was optimized with respect to subjective results. The range of the PESQ score resulted to $[-0.5, 4.5]$ [4].

Wideband Perceptual Evaluation of Speech Quality (WB-PESQ)

In 2005, the PESQ algorithm was extended to wideband evaluation. Thereby it was found, that by simply removing the IRS filtering in the preprocessing and using a samplerate of $F_s = 16$ kHz, the resulting *wideband PESQ (WB-PESQ)* is able to give reasonable results. In general, the IRS filter is not really substituted by an allpass filter but by a 100 Hz highpass filter [40]. Nevertheless, for mainly normalized signals with bandlimiting, only removing the IRS filter is assumed to be sufficient in this work, because the dedicated professional audio signals normally are lowcut processed and normalized. The auditory as well as the cognitive model remain as they are. In [66], it was evaluated, that this algorithm is also able to evaluate the quality for mono audio signals. Therefore it gets increasing interesting for the present work.

In general, the PESQ algorithm has the tendency to underpredict the performance when the perceived quality is high, and to overpredict the performance when the perceived quality is low [35]. Therefore, a final mapping of the PESQ score to respective MOS LQO values was proposed to allow a direct comparison of the objective results with MOS LQS values produced from subjective experiments [40]. Firstly a mapping for the narrowband implementation was proposed and recently also for the WB-PESQ score.

Thereby a MOS scale with a range of $[1, 4.5]$ is assumed because this is the normal range found in an ACR experiment [35]. The mapping from PESQ score to MOS LQO is performed as

$$y = 0.999 + \frac{4}{1 + e^{ax+b}} \tag{2.2}$$

with $a = -1.3669$ and $b = 3.8224$ for the WB-PESQ from ITU-T recommendation P.862.2 [40], mentioned in [12]. For the ordinary narrowband PESQ the values result to $a = -1.4945$ and $b = 4.6607$ [35]. The wideband mapping is depicted in fig. 2.4.

Perceptual Evaluation of Audio Quality (PEAQ)

The PEAQ algorithm is currently the only available standardized computational method for the purpose of audio quality assessment. It includes a high-quality audio model for small impairment rating, comparable with BS.1116 listening tests [5]. The focus of the rating scheme are coding distortions on fullband audio signals and possible impairments due to packet losses were not considered during development. Therefore, the PEAQ algorithm is not able to give reliable results for packet loss impairments which was proven by comparison with subjective tests in [61].

In the PEAQ algorithm preprocessing step, only a level adaption and a simple correlation based time-alignment is performed, while it is stated in the specification that an accurate synchronization is essential

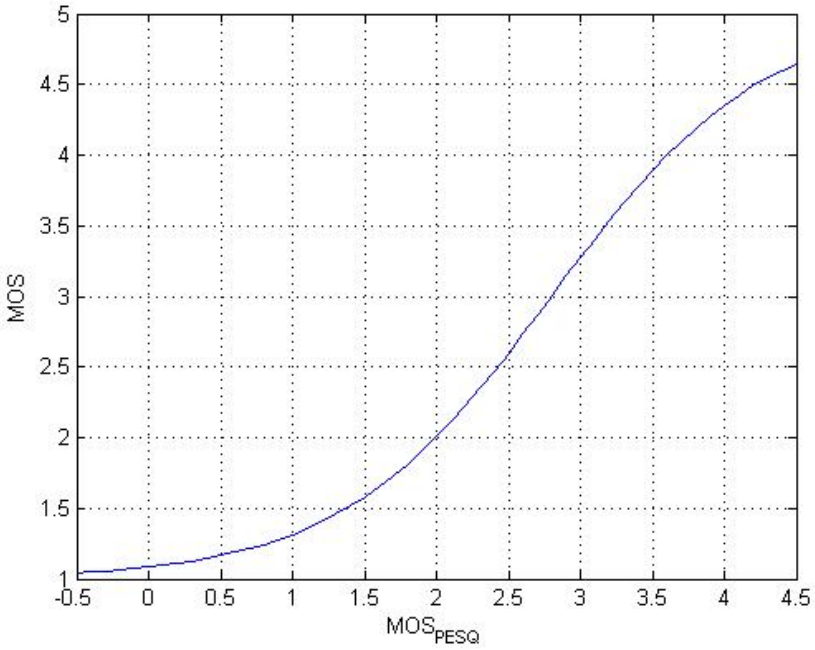


Figure 2.4: PESQ score to MOS mapping.

but no implementation is defined there [27]. For the perceptual model two approaches are possible: the *basic version* (with an FFT-based ear model) and the *advanced version* (with a filterbank based ear model). Hereby the later one is more accurate, but it is four times more demanding on computational power [5].

The different models compute a different number of *Model Output Values (MOVs)*, corresponding to different psycho-acoustical characteristics of the human auditory perception [97], i.e. changes in modulation, frequency-response curve modification, loudness difference and masking [46]. These MOVs are fed into an artificial neural network (the cognitive model) which was trained to map the MOVs to the scalar *Distortion Index (DI)*. This measure is linearly related to the perceived audio quality in terms of an Objective Difference Grade (ODG) with a range of $[-5, 0]$ corresponding to the SDG in the subjective domain (see above). This is the final output of these objective rating system [27]. Thereby, values near to zero (> -0.5) can normally be classified as transparent (i.e. perceptually undistorted) because of measurement uncertainties [47].

Quality Assessment with PEMO-Q

The PEMO-Q algorithm was proposed in [22]. Thereby the name comes from the acronym for the perception model, “PEMO”, while “Q” in PEMO-Q was added to indicate quality assessment [23]. It is able to perform a evaluation of speech signals as well as audio signals and was developed to address the rating of a wide range of distortions, while for packet loss impairments it is not trained so far as confirmed the inventor responding a question of the author of the present thesis.

As usual, in the preprocessing step a delay compensation and a level alignment is conducted. For the perceptual model, a special auditory filterbank-based transform is exploited. Thereby, the input signals are transformed into an internal representation which can be described as threedimensional activity patterns varying in time, frequency and modulation-frequency. Based on this model, two *Perceptual Quality Measures (PSM)* are calculated using also an artificial neural network as cognitive model. From the more refined one, an ODG output is derived by a linear transform [22].

2.2.3 The E-Model for Conversation Quality Rating

The *E-model* is currently the ITU-T recommendation for a mathematical model to perform speech transmission rating and network planning for bidirectional telephone services (ITU-T recommendation G.107 [41]). It gives an QoE estimate based on instrumentally measurable characteristics of the system [64] and therefore it relates the quality perception to physical parameters (more in [55]). Its main advantage is the ability to predict the overall *conversational quality* due to the incorporation of both equipment impairments on listening quality as well as impairments due to talker echo or transmission delay.

The main characteristics of the E-model are that it is parameter-based and non-intrusive. It operates with reduced-reference (RR), because it uses metrics which are derived from the incorporated signals (i.e. SNR). The E-model is empirical by nature and was developed on a large amount of auditory tests [64]. This helps to ensure that users will be satisfied with the end-to-end transmission performance of an observed system.

The E-model was original developed by the *European Telecommunications Standards Institute (ETSI)* at the end of the nineties [44] for the qualitative description of telephone networks with 3.1 kHz audio bandwidth (narrowband) handset telephony. In 2000, the E-model was approved as an ITU-T recommendation. Since 2002, it was extended for additional characteristics of IP based networks (mainly packet loss) [64] and recently for wideband transmission with up to 8 kHz audio bandwidth [56]. It was proposed in [65] to extend the model further for arbitrary network conditions.

The E-model is usable for estimating the transmission quality during network planning and set-up optimization [55], but also for monitoring in operation [9] as well as QoS optimization and control [84]. But until now it is only suitable for speech communication up to wideband transmission and therefore not applicable for fullband quality rating. Nevertheless, it can serve to build a fullband QoE rating scheme.

The R-Factor and the Impairment Factors Concept

The E-model outputs a scalar rating grade of transmission quality, the *transmission rating factor R*, on a range of $[0, 100]$, where 100 as maximal value indicates perfect perceived QoS [44]. The R-factor is obtained by summing up different impairment factors. Herewith different scalar input parameters (i.e. signal-to-noise ratio, packet loss or transmission delay) are grouped into different classes of impairments. The assumption behind this approach is an *impairment additivity* which is reasonable because of a concept from the science of psychology stated by Alnat in the seventies [1] so [44]:

"Psychological factors on the psychological scale are additive."

This is the fundamental principle of the impairment factor concept. For example, the impacts of a coding algorithm and impacts due to packet loss can be separated and summed up.

The R-factor is composed as

$$R = R_0 - I_s - I_d - I_{e,eff} + A\,, \tag{2.3}$$

where the basic transmission rating factor R_0 reflects the signal-to-noise ratio, the *simultaneous impairment factor* I_s accounts for degradations simultaneous to the transmitted speech signal (i.e. signal-correlated noise) and the *delay impairment factor* I_d includes the delay impacts in a bidirectional transmission. The *advantage factor A* in the E-model stands for some "advantage in access" [44], for example

if a wireless system is used and the mobility possibly compensates some quality degradation.

$I_{e,eff}$ is the *effective equipment impairment factor* which is a function of the equipment impairment factor $I_e = I_{e,eff}(\rho = 0)$ (i.e. of degradations due to lossy coding) and the packet loss ratio ρ, accounting for dependent loss with the inclusion of the $BurstR$ metric,

$$I_{e,eff} = f(I_e, \rho, BurstR) \; . \tag{2.4}$$

The E-model for narrowband transmission rating has $R_0 = 93.2$, which is the result for a standard narrowband ITU-T G.711 telephone connection with a "clean channel" where all parameters are set to its default values. For wideband transmission, the R-factor range is extended to $[0, 129]$, based on subjective evaluations, therefore the maximum value is increased from $R_{max} = 100$ to $R_{max} = 129$ and by definition, the basic transmission rating factor $R_{0,WB} = 129$. A clean wideband communication is therefore $\approx 36\%$ "better" than a clean narrowband channel [93]. This was adapted by the ITU-T in the E-model (ITU-T rec. G.107, Appendix II [41]).

MOS Values to R-Factor Scale Conversion

The R-factor from the E-model can finally be mapped to an $MOS\ CQE$ (estimated conversational Quality MOS). In the ITU-T recommendation G.107 [41] the conversion is defined as

$$MOS = \begin{cases} 1 & \text{for } R \leq 6.5 \; , \\ 1 + 0.035\, R + R\,(R-60)(100-R)7 \cdot 10^{-6} & \text{for } 6.5 < R < 100 \; , \\ 4.5 & \text{for } R \geq 100 \; . \end{cases} \tag{2.5}$$

The inverse mapping from $MOS\ CQE$ to the R-factor can be done by the complicated Candono's Formula as in [19] (since 2005 mentioned in G.107, Appendix I [41]). In [84], a simplified 3rd-order polynomial fitting is proposed as

$$R = 3.026\ MOS^3 - 25.314\ MOS^2 + 87.06\ MOS - 57.336 \; . \tag{2.6}$$

Wideband Case

The wideband R-factor extension is included in the MOS-to-R-factor computation by stretching the result. This can be done with a linear or non-linear model, proposed in [56] and [64],

$$R_{\mathrm{WB}} = \frac{R_{0,\mathrm{WB}}}{R_{0,\mathrm{NB}}}\, R_{\mathrm{NB}} = 1.29\, R_{\mathrm{NB}} \; , \tag{2.7}$$

$$R_{\mathrm{WB}} = a\ (e^{R_{\mathrm{NB}}/b} - 1) \; , \tag{2.8}$$

with $a = 169.38$ and $b = 176.32$ from subjective evaluations [56] [64].

If narrowband results for I_e exist (i.e. from the "provisional planning values" list in the ITU-T recommendation G.113 [42]), the respective wideband value $I_{e,\mathrm{WB}}$ can be obtained by simple shifting,

$$I_{e,\mathrm{WB}} = (R_{0,\mathrm{WB}} - R_{0,\mathrm{NB}}) + I_{e,\mathrm{NB}} = 35.8 + I_{e,\mathrm{NB}} \; . \tag{2.9}$$

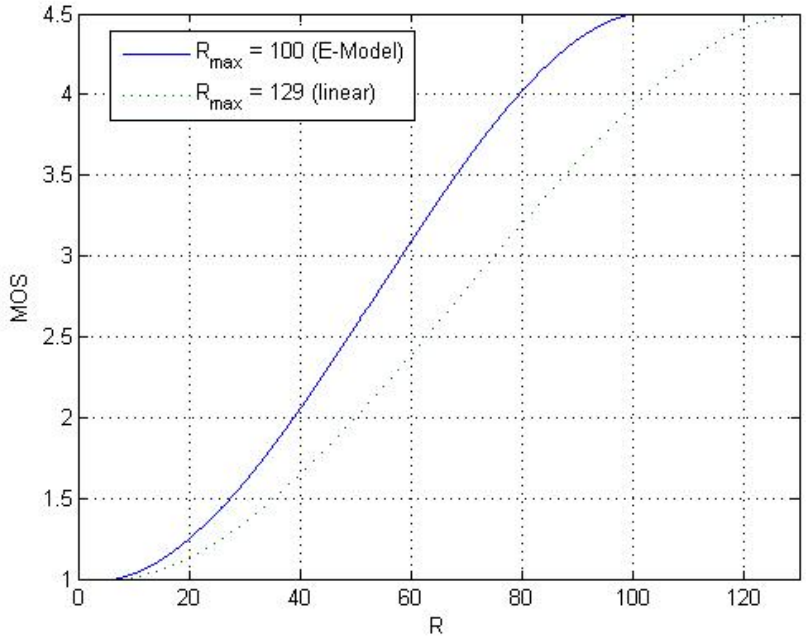


Figure 2.5: *MOS CQE* versus transmission rating factor *R* for NB and WB case.

Equipment Impairment Factor Accounting for Degradations Due to Packet Loss

In the extended E-model [41], the effective equipment impairment factor $I_{e,eff}$ is formulated as

$$I_{e,eff} = I_e + (I_{e,0} - I_e)\frac{\rho}{\frac{\rho}{BurstR} + Bpl} \,, \tag{2.10}$$

with ρ is the packet loss ratio in percent and Bpl the packet-loss robustness factor, introduced in [63]. Some Bpl provisional planning values are also found in the ITU-T recommendation G.113 [42]. I_e is the initial equipment impairment factor for $\rho = 0$ and $I_{e,0}$ was chosen as $I_{e,0,\mathrm{NB}} = 95$ for narrowband transmission respective $I_{e,0,\mathrm{WB}} = 129$ for wideband transmission rating.

The E-model accounts for independent ("random") packet loss since its revision in 2002 [63]. Following a proposal from [65], the formula was later extended for *dependent* packet loss based on a 2-state Markov model, including the special "burst ratio" $BurstR$ [54], a measure "expressing the tendency of a particular loss distribution for consecutive packet loss independently of the overall loss rate" [65].

Another approach is presented in [9] and used in [84], where the E-model compatibility was no explicit demand. Here a non-linear least squares curve fitting for a logarithmic model is proposed,

$$I_{e,eff} = a\ ln(1 + b\rho) + c \,, \tag{2.11}$$

while it is so far only used for the description of independent loss impacts.

Derivation of the Equipment Impairment Factor I_e

While the provisional planning values in ITU-T recommendation G.113 [42] are only specified for a limited set of narrowband and wideband codecs (i.e. ITU-T G.711 and G.722), sometimes new parameters

for new coding algorithms respective codec configurations have to be derived. This can be done based on subjective evaluations, following the methodology for the derivation of equipment impairment factors from subjective listening-only tests, defined in the ITU-T recommendation P.833 [36]. Because for the subjective approach a high effort is needed, also a methodology for the derivation of equipment impairment factors from instrumental models is proposed in ITU-T Recommendation P.834 for narrowband codecs using PESQ [37] and in P.834.1 for wideband codecs using WB-PESQ [43], while the latter one was recently approved and previously proposed in [56] and [64]. Furthermore, it was used in [84].

If equipment impairment factors I_e for discrete points of a loss process with altering packet loss ratio (i.e. $\rho = \{0, 2, 5, 10, 15\}$ %) exist, obtained from one of the mentioned methods, $I_{e,eff}$ is obtained by non-linear least squares curve fitting. The only fitting variable in the E-model is then Bpl.

Considering Bandwidth Impairment

For the future, the incorporation of an impairment factor for linear distortion of narrowband and wideband speech transmission was proposed [93], based on findings in [64]. The so called *bandwidth impairment factor* I_{bw} shall be incorporated as an additional impairment factor for linear distortions due to band limitations by the separation of coding algorithm impairment (i.e. non-linear distortions) and a linear band limiting impairment, both until now contained in the equipment impairment factor I_e. That is now decomposed as

$$I_{e,\mathrm{WB}} = I_{bw} + I_{res} \,, \tag{2.12}$$

where I_{res} is the residual portion of I_e and reflects the coding algorithm only impairment.

The reason behind this approach is, that with the possibility of wideband consumer telephony transmission, a narrowband codec is perceived as distorted in comparison with a wideband codec. Subjective listening tests showed that wideband has an $1.3 - 1.5$ points MOS advantage over narrowband [64]. While only narrowband telephone channels were available, narrowband was not perceived as distorted [64] because codecs with different audio bandwidths were not used.

To obtain I_{bw}, the resulting MOS values from subjective auditory tests were mapped on the R-factor scale. Then a formula for I_{bw} was found by curve fitting in [64] as

$$I_{bw} = 3.5 \cdot 10^{-2} \, |s| - 6.7 \cdot 10^{-3} s - 7.4 \, z_{bw} + 129.2 \tag{2.13}$$

with

$$s = f_c - 9.9 \, (z_{bw} + 101.8) \,, \tag{2.14}$$

where f_c is the center frequency in Hz computed by the geometric mean of the lower frequency f_{low} and the upper frequency f_{up},

$$f_c = \sqrt{f_{low} \cdot f_{up}} \,, \tag{2.15}$$

and z_{bw} is the transmission bandwidth in Bark [97], computed also from the limiting frequencies f_{low} and f_{up},

$$z_{bw} = z(f_{up}) - z(f_{low}) \,. \tag{2.16}$$

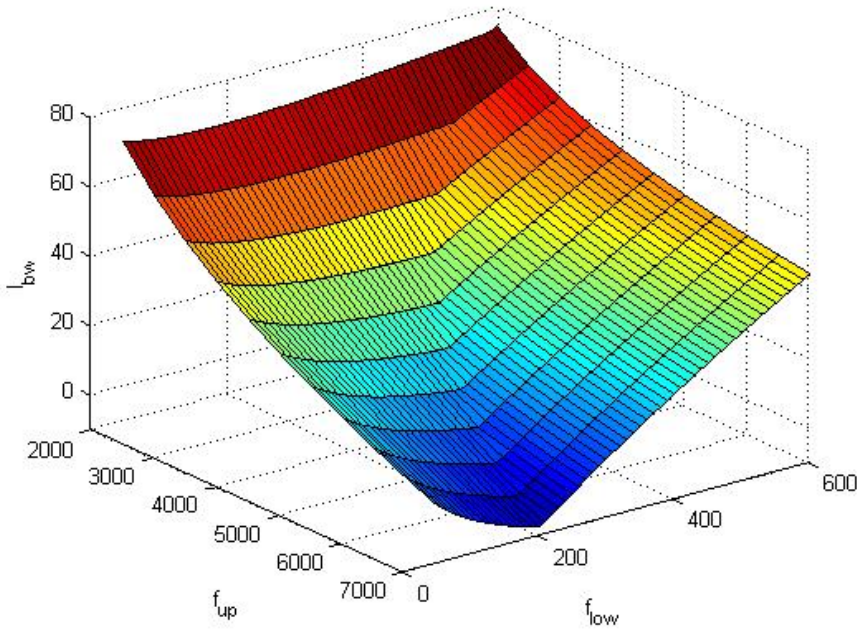


Figure 2.6: Bandwidth impairment factor for wideband transmission with respect to different lower and upper cutoff frequencies in Hz based on the formula from [64].

The *Bark scale* is a psycho-acoustical scale for the perceived pitch of a tone ("tonalness"), named after Heinrich Barkhausen (1881-1956). The scale is defined from 0 to 25 Bark and corresponds to the audio bandwidth perceivable by the human ear in frequency (20-20000 Hz). Doubling the Bark value corresponds to a doubling in perceived tone pitch. It is related to the Mel scale as

$$1 \text{ Bark} = 100 \text{ Mel} . \tag{2.17}$$

Frequency can be converted to the Bark scale with

$$z = 13 \arctan(0.00076 f) + 3.5 \arctan((f/7500)^2) . \tag{2.18}$$

For practical reasons, z_{bw} can be approximated by the *Equivalent Rectangular Bandwidth (EBR)* [64], which is estimated from the amplitude spectrum.

This new approach is very promising for a further extension of the model from wideband to fullband, which was exploited in this work. In fig. 2.7, the results presented in [93] are documented, where for different NB and WB speech codecs the obtained wideband equipment impairment factors, the bandwidth impairment factors and the residual portion for the coding algorithm only impairments are presented.

Delay Impairment Factor

The delay impairment factor I_d accounts for impairments due to transmission delay but also due to talker and listener echo. For these impairments, different possibilities for the modeling of I_d as function of the one-way delay d_{e2e} (short d) were proposed:

- The current E-model recommendation in ITU-T G.107 [41], $I_{d,1}$,

[2] The *Symmetric Disturbance* d_{SYM} is the spectral distance according to the processing in PESQ [35].

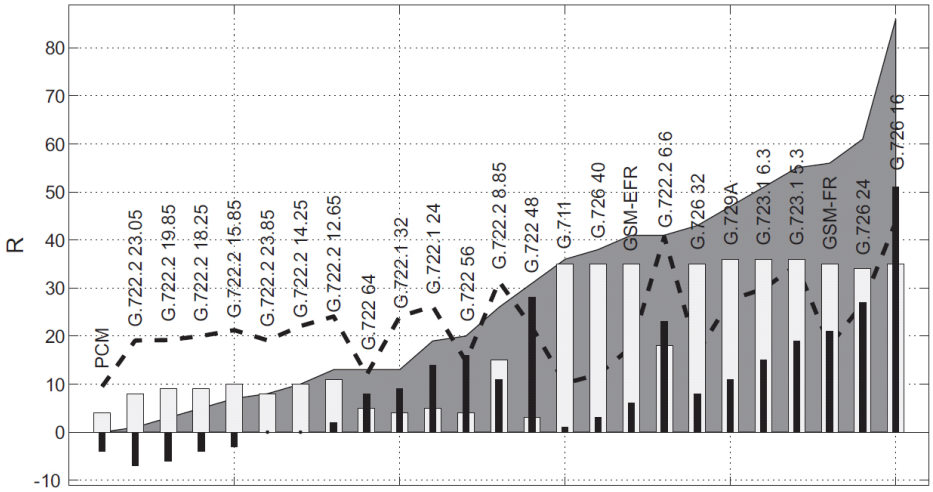


Figure 2.7: $I_{e,WB}$ (dark gray), I_{bw} (light gray), I_{res} (black) and d_{SYM} [2](re-scaled, dashed) for different NB and WB speech codecs, diagram taken from [93].

- A 6th order polynomial fitted curve, $I_{d,2}$ (accurate for delay $d < 600$ ms) from [84],
- The AT&T simplified model, $I_{d,3}$ (accurate for delay $d < 400$ ms) from [9],
- A simplified E-model recommendation in ITU-T G.114 [39], $I_{d,4}$.

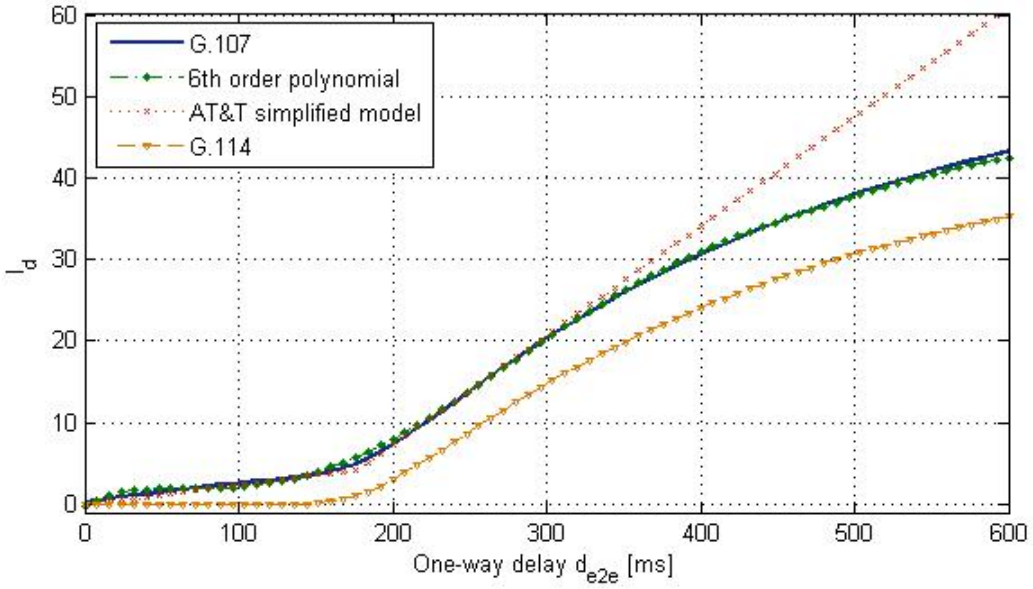


Figure 2.8: Different approaches for I_d versus one-way delay d_{e2e}.

$I_{d,1}$ can be calculated by a series of equations given in G.107 [41] and is composed as

$$I_{d,1} = I_{dd} + I_{dte} + I_{dle} \,, \tag{2.19}$$

where I_{dd} expresses the impairment caused by absolute delay higher than 100ms, I_{dte} is the *talker echo impairment* and I_{dle} is the *listener echo impairment* [64]. All parameters are calculated via some dedicated equations from the default settings (corresponding to a clean connection) in dependency of the delay

while R_0 is obviously set to $R_{0,max} = 93.2$ (for narrowband) and a perfect echo cancellation is assumed like in [84],

$$I_{dd} = 25\left[(1+X^6)^{1/6} - 3\,(1+(X/3)^6)^{1/6} + 2\right] \tag{2.20}$$

with

$$X = \frac{\log_{10}(d/100)}{\log_{10}(2)}\,. \tag{2.21}$$

The formulas for calculating I_{dte} and I_{dle} can be found in Appendix E of [64].

The other models are less accurate with respect to telephony research, while $I_{d,2}$ is well aligned to the E-model curve for delay less than 600ms which can assumed to be sufficient for the majority of VoIP calls [84]. $I_{d,2}$ and $I_{d,3}$ are composed as

$$I_{d,2} = a_6 d^6 + a_5 d^5 + a_4 d^4 + a_3 d^3 + a_2 d^2 + a_1 d^1 + a_0\,, \tag{2.22}$$
$$I_{d,3} = 0.024d + 0.11(d - 177.3)\cdot H(d - 177.3)\,, \tag{2.23}$$

for $I_{d,2}$ with

$$\mathbf{a} = [\,a_0,\ a_1,\ \ldots\ a_6\,] \tag{2.24}$$
$$= [-0.1698, 0.103, -0.001802, 1.344\cdot 10^{-5}, -3.903\cdot 10^{-8}, 5.062\cdot 10^{-11}, -2.468\cdot 10^{-14}]\,, \tag{2.25}$$

and for $I_{d,3}$ with the step function

$$H(x) = \begin{cases} 0, & \text{if } x < 0\,, \\ 1, & \text{if } x \geq 0\,. \end{cases} \tag{2.26}$$

In $I_{d,4}$ the E-model formula is used but with ignoring the initial echo impairment factors I_{dte} and I_{dle} and setting $I_{d,4} = I_{dd}$.

2.3 Real-time Audio Content Delivery over IP

After the discussion of QoS, real-time audio content delivery over IP, a subset of multimedia content delivery and the basis for audio communication services, shall be introduced. It is often shortly denoted as simply *Audio over IP (AoIP)*. Now, first the notion of real-time is defined and a classification of services in terms of audio bandwidth is given, while also the different delay requirements are addressed. Then, the necessary IP technology for communication is surveyed before depicting the main system components for real-time audio transmission over IP.

2.3.1 Notion of Real-time

Todays multimedia content delivery over IP and especially communication services over IP is in principle enabled by the ensuring of a *real-time processing*. In the following this important requirement is addressed following the definition given in [79] and [80]. A *real-time process* in digital systems is in general be defined in standardization bodies as the DIN (German National Institute of Standardization) and the ANSI (American National Standards Institute) as

"A real-time process is a process which delivers the results of the processing in a given time-span."

This indicates, that the presentation of data must be done within a certain time period T, because a continuous playout is required for real-time applications. This time period need not to be small in general, it depends on the specific requirements of a service and the related demands of the respective user. While the time instance is an absolute term, especially for the specific requirements of professional audio transmission, the processing and transmission must be guaranteed by the system components inside the defined bounds.

Furthermore, error-free processing and transmission is no more the only demand for correct data presentation in the case of real-time systems, but also the time when the processing and transmission is finished have to be considered here because errors of real-time systems, i.e. distorted playout, do not appear only because of hardware or software failures but also if the system can not give the desired result after a previously defined time instance.

As smaller the desired certain time period T for data presentation, the problem complexity for the system design increases. Therewith, the analysis and cross-layer optimization of the QoS parameters with relation to time on the different QoS layers is of particular importance for finding a solution. The problem complexity increases even more if the system has determinants to consider which can not be manipulated. In this thesis, the problem of assuring real-time transmission for audio contribution over IP is concerned from an end-to-end perspective, where the network is assumed to be given. In the case of a relatively high network delay, the application is forced to ensure that the time requirements nevertheless are met which could also be impossible and a trade-off would have to be accepted.

2.3.2 Classification in Terms of Audio Bandwidth

In the following, audio content delivery over IP services shall be presented and classified. The main domains with respect to different requirements and QoS levels are

- audio on demand Internet streaming,
- Internet radio broadcasting,
- professional AoIP streaming,
- Internet telephony (i.e. for interactive gaming),
- IP-based telephony: *Voice over IP (VoIP)*,
- teleconferencing using Internet telephony or VoIP,
- audio communication over IP (i.e. for telepresence systems),
- professional audio contribution (ACIP).

The first three applications are often TCP-based (see below) or use error correction schemes as *Forward Error Correction (FEC)* for providing a reliable transmission. Nevertheless, this introduces additional delay but this is not of major importance here because no interactivity is required. The other ones are communication services which especially need acceptable delay bounds for the desired interactivity. Thereby, Internet telephony provides the lowest QoS because of the unpredictable nature of the Internet while the demands on QoS and especially delay are increasing for consumer telephony upto professional audio contribution (ACIP). Internet telephony and VoIP can be separated because of their used networks. Internet telephony uses the best-effort Internet while for VoIP normally managed networks providing a

specific QoS are exploited. Furthermore, VoIP relies on international telecommunication standards. The particular differences between VoIP and ACIP are discussed in chapter 1.3 in further detail.

Besides the important influence of the delay on QoS and the user satisfaction, the different applications provide different audio quality from ordinary telephony speech quality up to high fidelity audio. While this quality is strongly influenced by the quality of the used coding algorithm, the main factor in general determining the audio quality in terms of acoustics and human perception is the *audio bandwidth B* [97]. In table 2.6 an overview on different audio bandwidths used in digital audio transmission and especially for IP-based communication services is presented while also some dedicated coding algorithms are mentioned as examples.

	Audio Bandwidth [Hz]	**Sample Rate** [kHz]	**Example Codec**
Narrowband (NB)	300-3400	8	ITU-T G.711 [31]
Wideband (WB)	50-7000	16	ITU-T G.722 [32]
Super-wideband (SB)	70-12000	24	Skype SILK [92]
Ultra-wideband (UB)	40-15000	32	Eapt-X [52]
Fullband (FB)	20-22000	48	MPEG [77]

Table 2.6: Possible audio transmission bandwidths in audio communication.

Therewith, *narrowband (NB)* transmission is the traditional configuration in telephony. Today, also a *wideband (WB)* and *super-wideband (SB)* transmission is enabled with VoIP and Internet telephony. While WB was introduced by telephone service providers in the last years, the later one SB was recently introduced for the commercial Internet telephony service Skype and proposed to standardize it in the IETF [92]. The *ultra-wideband (UB)* and *fullband (FB)* transmission is usually used for audio transmission (also referred to as *broadband* transmission).

The classification here is based on definitions for speech communication [64] but extended by the author for audio communication services. The intermediate fullband audio bandwidth with sample frequency $F_s = 44.1$ kHz (also known as "CD-quality") is neglected here because it is usually not used for communication purposes while its usage for consumer audio streaming is quite possible beyond telecommunications and professional audio standards.

2.4 EBU N/ACIP Specification

The present work was designed to account always for international standards and compatibility. This is especially important here, because for professional audio contribution over IP often remote places are connected and the used end-devices must provide the functionality needed for reliable transmission. Therefore the EBU N/ACIP specification was comprehensively examined.

The European Broadcasting Union (EBU) project group for Audio Contribution over IP (N/ACIP) under the auspices of the Network Management Committee (NMC) has presented the EBU N/ACIP specification in 2007 as EBU TECH 3326 [57]. Since the inaugural meeting in February 2006, they had discussed the issue based on an initiative of German vendors and broadcasters [81]. Before, solutions from different manufacturers for audio contribution over IP end-devices have been incompatible with each other.

The main objectives of the N/ACIP specification are

1. ensure *interoperability* between ACIP devices and
2. provide *guidlines* on the utilization of audio over IP systems to broadcasters.

The interoperability is a necessary requirement for the functional capability of professional audio contribution over IP. Therefore the specification is based on standardized methods for audio communication over IP which are recommended i.e. for VoIP or similar professional digital audio systems as well. The N/ACIP working group elaborated the areas to account for in the interoperability framework as

- *IP transport protocols*, including "port definition and packet loss recovery" [57],
- *audio coding algorithms* to be implemented,
- *audio frame encapsulation* definitions, and
- *signaling* challenges.

The different technology principles and components were desired to be selected appropriately with the goal of delivering a performance "equal to traditional telephony lines" [81]. Therefore, the N/ACIP document describes a minimal set of *musts*, but also a set of *recommended* and *optional* additions. In the following, the different specifications are surveyed. Thereby, only the mandatory ones are mentioned in particular, recommended and optional implementations are only noted if necessary in the context. The whole specification can be found in the EBU TECH 3326 [57][3].

2.4.1 IP Transport

For audio contribution over IP, certainly the Internet protocol (IP) *must* be available in version 4 (RFC 791) at the network layer, while also version 6 (RFC 2460) and IP multicast (RFC 1112) is desired. For the media transport RTP (RFC 3550/3551) *must* be used over UDP (RFC 768) providing also the UDP checksum. Furthermore, RTP over TCP is also considered. Retransmission for active recovery (RFC 4588) may be used, while therefore also a support of the extended RTP profile (RFC 4585) would be necessary. The implementation of RTCP and respective sender and receiver reports (SR/RR) is only recommended.

Forward Error Correction (FEC) is not integrated, because "it is currently a work in progress" [57]. A possible option considered is the RTP payload format for generic forward error correction (RFC 5109). If FEC would be used, it *must* be provided as a separate stream.

Application-specific port numbers are surveyed in fig. 2.7, already including also the SIP signaling port. Port numbers are address identifies of the type of service signaled in the UDP header. Here it is mandatory that the sent and received streams are always on the same port.

2.4.2 Mandatory Audio Codecs

The N/ACIP project group selected four well-known audio coding formats as mandatory set for compatible audio contribution over IP to provide a minimum common base [58]. Thereby the most important selection criteria were patents costs and implementation complexity. Therefore, traditional standards for digital audio contribution were chosen with the drawback that within the development of this coding algorithms, the IP streaming application was not regarded especially [77]. Consequently the mandatory coding algorithms are in principle not optimized for IP streaming and do not include i.e. scalability or effective Packet Loss Concealment (PLC). The four mandatory audio coding algorithms are

[3] The EBU TECH 3326 is available at: http://tech.ebu.ch.

Transport Protocol	Port Number
RTP over UDP	5004
RTP over TCP	5004
RTCP	5005
TCP	5004
FEC over RTP	5006
SIP	5060

Table 2.7: Standard application-specific port numbers specified in N/ACIP.

- ITU-T G.711,
- ITU-T G.722,
- ISO MPEG-1/2 Layer 2,
- Linear PCM.

Parameter	ITU-T G.711	ITU-T G.722	ISO MPEG-1/2 L2	16 bit PCM	20/24 bit PCM
Bitrate [kbit/s]	64	64	32-384	512-1536	640-1152
No. of channels	1	1	1-2	1-2	1-2
Audio samplerate [kHz]	8	16	16-48	32/48	32/48
RTP clock [kHz]	8	8	90	32/48	32/48
MIME subtype name	PCMA/PCMU	G722	MPA	L16	L20/L24
RTP payload type	8/0	9	14 / dynamic	dynamic	dynamic
Inter-packet time [ms]	20	20	24	4	4

Table 2.8: Standard parameters specified in N/ACIP for using audio coding algorithms with RTP transport.

In table 2.8 an overview on the main media profile parameters important for RTP/IP transport of these audio coding formats is given. Thereby the RTP encapsulation *must* be done according to RFC 3551 [74].

In addition to the basis mandatory set of coding algorithms, the EBU also specified the parameters for recommended and optional coding algorithms which are free to implement or not for the manufacturers. It is important, that for a specific coding algorithm an RTP audio frame encapsulation is defined in an IETF RFC for ensuring interoperability[4].

The recommended formats are ISO MPEG-1/2 Layer 3 and MPEG-4 AAC-LC (*low complexity*) respective AAC-LD (*low delay*). Optional are further mentioned the proprietary Enhanced APT-X (Eapt-X) algorithm[5], MPEG-4 HE-AACv2 (*high efficiency*), Dolby AC-3[6] and the recent communication algorithms AMR-WB+ (3GPP TS 26.290) and AMR-WB (ITU-T G.722.2). The expected most suitable ISO MPEG AAC-ELD (*enhanced low delay*) algorithm [24], developed dedicated for audio communication over IP is not mentioned so far.

[4] A comprehensive overview on RTP payload types can be found on Wikipedia, available: http://en.wikipedia.org/wiki/RTP_audio_video_profile.
[5] So far, there is no RTP payload format specified by the IETF even if there is a draft provided by the Eapt-X vendor.
[6] More information on Dolby AC-3 is available at http://www.atsc.org/standards/a_52b.pdf.

2.4.3 Signaling Issues

For signaling, SIP (RFC 3261) and SDP (RFC 4566) are mandatory as well as the session announcement protocol (SAP, RFC 2974). The later shall be used for unidirectional multicast links. The session description *must* include the following fields:

- *v* (the protocol version),
- *o* (the originator and session identifiers),
- *s* (the session name, single space if not used),
- *c* (the connection data),
- *t* (the session times, '0 0' if not used),
- *m* (the media description).

The SDP session description i.e. for the description of a session with 16 bit linear PCM coding ($F_s = 48$ kHz, stereo) would be (taken from the specifications in [57]):

```
v=0
o=alice 2890844526 2890844526 IN IP4 host.anywhere.com
s= (single space)
c=IN IP4 host.anywhere.com
t=0 0
m=audio 5004 RTP/AVP 98
a=rtpmap:98 L16/48000/2
```

For the session management with SIP, three issues are mentioned in particular which are important for professional audio contribution. These are

- the codec negotiation,
- independent encoder/decoder settings,
- reconfiguration.

For the codec negotiation between to end-devices, the model from RFC 3264 is recommended (see chapter **??**). Independent encoder/decoder settings are desirable, because often a bidirectional connection with narrowband return is exploited for audio contribution (see chapter 1.1). By default, the same coding algorithms are negotiated during the call initiation with SIP/SDP. For independent coding algorithms at sender and receiver, the establishment of two unidirectional SIP sessions are mandatory. Sometimes also a reconfiguration of an already established connection is needed, i.e. for changing the coding scheme. Therefore, it is mentioned also as recommended for an implementation.

3 Objective QoE Evaluation for AoIP

For the construction of a non-intrusive parameter-based QoE rating model for professional audio contribution over IP, a basis of QoE rating results with respect to model parameters must exist. Therefore normally subjective tests are performed. In this work, the instrumental WB-PESQ algorithm as well as the PEAQ algorithm should be used for quantifying user perceived audio quality for different impairments due to the usage of professional AoIP. The approach is a trade-off because WB-PESQ is originally dedicated to speech quality assessment but the dedicated audio quality prediction algorithm PEAQ can not rate stronger impairments as packet loss in a correct manner while PESQ was developed to consider this problem.

The final goal of the evaluation is a discrete set of QoE metrics related to quantitative parameters of the audio stream and particular network properties. Furthermore, the results with respect to packet loss are non-linear curve fitted for a better comparison of the results. Finally, only WB-PESQ results were produced in this work while necessary results of the PEAQ algorithm were taken from already existent evaluations [47].

3.1 Analyzed Impacts on Perceived Audio Quality

The quality impairment analysis in this chapter is focused on the evaluation of listening-only quality, resulting in method specific MOS LQO metrics. Conversation quality evaluation for ACIP can only be done with subjective conversation tests [20] which would be to much additional effort here. Therefore, the analysis was restricted to the listening-only quality assessment.

The analyzed quality degradations can be categorized in two parts:

1. The *coding algorithm impairments* and
2. possible *network impairments* due to packet loss.

The coding algorithm impairments due to coding artifacts depend mainly on the general algorithm type performance, possible audio bandwidth restrictions as well as the coding bitrate. A possible quantization distortion due to the sampling resolution Q and the consequences for the perceived audio quality due to different channel configurations ch were not regarded in particular in this thesis. Therefore, the perceived audio quality based on coding impairments, assessed with some MOS_{coding} metric can be described in general as

$$MOS_{\text{coding}} = f(c_{type}, r_c, F_s, Q, B, ch) = f(\mathbf{c}) \tag{3.1}$$

with the coding algorithm type c_{type}, the coding bitrate r_c, the audio sampling frequency F_s, the quantization resolution Q, the audio bandwidth B and the channel configuration ch. All these parameters can be collected in the vector of coding parameters for each coding algorithm, $\mathbf{c} = [c_{type}, r_c, F_s, Q, B, ch]^T$. For the coding impairment QoE analysis, the packet loss ratio is forced to be $\rho = 0$.

The impairments on the perceived audio quality through network packet loss are determined by the amount of packet lost in a period of time, the burstiness of the loss process as well as the packet size of the audio stream L because as longer the packet size more consecutive audio information gets lost. This impairments can also be described by a QoE metric, now with respect to loss as

$$MOS_{\text{loss}} = f(\mathbf{P}_{\text{loss}}, \mathbf{c}, L) \tag{3.2}$$

while the packet loss process is characterized by its state transition matrix $\mathbf{P}_{\text{loss}}$ of the 2-state Markov model which includes the description of loss amount and correlation (see chapter **??**). The analysis with respect to different packet loss processes has the main focus in this chapter and its QoE evaluation was being done most detailed.

3.2 Experiments Preparation

In the following, the preparations necessary for the objective QoE assessment procedure are addressed. First, the principal dependencies of the previous mentioned QoS parameters with respect to the perceived audio quality are depicted, before some notes on the QoE assessment method are stated. Then the chosen audio coding algorithms to evaluate and their selected configurations are presented. Before continuing with the setup description, the set of audio testfiles for the objective evaluation as well as the chosen packet loss process configurations for the packet loss impairment analysis are introduced.

3.2.1 Impairment to Parameter Relation Analysis

Now an brief overview on the principal impairment trends shall be given, starting with the coding algorithm impairments.

The following equations show the principal dependence of the perceived audio quality assessed by a MOS on the principal coding parameters of a specific coding algorithm where only one parameter is varied,

$$MOS_{\text{coding}}(r_c) \propto r_c \tag{3.3}$$
$$MOS_{\text{coding}}(B) \propto B \tag{3.4}$$
$$MOS_{\text{coding}}(Q) \propto Q \tag{3.5}$$

For the loss perception, the principal expected dependency for a specific coding algorithm evaluated for a specific loss process configuration with a packet loss ratio $\rho > 0$ can be described as

$$MOS_{\text{loss}}(L_{IP}) = \propto 1/L_{IP} \tag{3.6}$$

This shows another dimension of the packet size influence on the perceived quality. Already for the impairments due to latency, the packetization delay which depends on the packet size has to be considered as was shown in chapter **??**. Therefore, a larger packet size has a negative impact both on latency as well as on the quality if loss is considered. This should be accounted for the later development of a QoS optimization algorithm.

On the other hand, a larger packet size reduces the overall bandwidth requirement because the IP overhead is less. This have to be considered if the available bandwidth reduces and a trade-off between delay and quality have to be done [10]. Also, a larger packet size has a lower packet loss rate (amount of losses per time) in comparison with smaller packet sizes if the available bandwidth is used effectively.

3.2.2 QoE Assessment Method

Unfortunately, no appropriate intrusive objective quality assessment method exists for the estimation of packet loss impairments on the perceived quality of an audio signal. The ITU-R standardized PEAQ

algorithm was only developed for evaluating the potentials of audio coding algorithms which introduce small impairments. PEMO-Q as a further development matches better with subjective results for a higher amount of impairments, but also do not account for loss impacts. In contrast, the ITU-T PESQ algorithm was only developed for objectively assessing the quality of narrowband speech signals but the packet loss impairment was considered during the development.

The wideband extension of PESQ (WB-PESQ) can rate speech signals up to an audio bandwidth of 7 kHz, which is unfortunately still too low for fullband audio signal evaluation. Furthermore, it was not extensively tested on perceptual transform coding algorithms (i.e. MPEG) or with non-speech signals. Nevertheless, some evaluations in [66] showed, that it can also be used for the quality assessment of wideband mono audio signals.

Therefore, a trade-off approach was chosen for the objective evaluation in the present work. While the PESQ algorithm seems to be the most reasonable one for packet loss impairment rating, which was also confirmed by the inventor of PEMO-Q in a statement delivered to the author, PESQ was chosen primarily but reasonable using the WB-PESQ extension. To overcome the wideband limitation of WB-PESQ, it was decided to incorporate also PEAQ results if higher bandwidth quality rating is addressed. Here only the coding distortion were evaluated, which means that only evaluations of signals without packet loss impairments ($\rho = 0$) were used.

While in [21] it is stated with reference to [62] that “composite objective measures are obtained by combining basic objective measures to form a new measure”, this approach should also in general be possible here. Therefore, for the combination of WB-PESQ and PEAQ results a linear combination approach was chosen which is offered in chapter 4.

For the real WB-PESQ evaluation, the Matlab code from [51] were used. Unfortunately, this implementation is only the ordinary PESQ algorithm which had to be modified for its usage as WB-PESQ. This was possible because there are not so much differences between the two approaches. The available implementation supports in general a quality evaluation of mono audio signals up to a sampling frequency of $F_s = 16$ kHz which is primarily necessary for a wideband extension. Therefore, only the narrowband IRS prefilter had to be excluded which changes the input characteristic to an allpass as desired for the evaluation in this work. To get also a rating for stereo signals, in this case an average of the QoE results for the left and right channel can be performed as simplest assumable approach. Furthermore, it has to be considered that most of the evaluated signals have a sampling frequency $F_s = 48$ kHz. Therefore, they had to be downsampled to $F_s = 16$ kHz which was done with the standard Matlab implementation for this task [88].

3.2.3 Audio Coding Algorithms and Configurations

Based on the set of available audio coding algorithms, a set of algorithms theoretically suitable for audio contribution over IP was selected. Table 4.3 summarizes the chosen algorithms and their configurations which are similar to the algorithms studied throughout the traffic analysis in chapter **??**. For the QoE analysis with respect to the packet size, appropriate values for these were selected which enable a comparison with respect to the impact of the packet size on a lossy link for a given coding algorithm. Because of the necessary high effort for evaluating an algorithm configuration for all chosen packet loss process characterizations (see chapter 3.2.5 below), only extreme values for the packet size were chosen (see table 4.3).

[1] The ITU-T G.711 coding algorithm was only analyzed with respect to coding impairments.

Coding Algorithm	Channels	Bitrate [kbit/s]	Audio Samplerate [kHz]	Packet Size [Byte]	Δt_p [ms]
MPEG L2 1	2	384	48	1196	24
MPEG L2 2	2	256	48	812	24
Eapt-X 1a	2	384	48	1100	21.25
Eapt-X 1b	2	384	48	332	5.25
Eapt-X 2a	2	256	32	828	24
Eapt-X 2b	2	256	32	336	8
Eapt-X	1	64	16	208	16
ITU-T G.722	1	64	16	200	20
ITU-T G.711 [1]	1	64	8	200	20

Table 3.1: Set of coding algorithms selected for the QoE evaluation and their basic configurations.

Unfortunately, many well known coding standards as MPEG1/2 have a fixed frame size, coupled with the coding rate and the sampling frequency. This dependency was already depicted in chapter **??** in eq. (**??**) and leads to an fixed packet size per coding configuration which could only be varied if more than one frame would be transmitted in one packet which i.e. for the MPEG1/2 coding algorithms is not recommended because the delay would increase in an not tolerable manner.

Finally, only the Eapt-X algorithm was available as algorithm with selectable packet size. These were chosen for different configurations comparable to the MPEG ones (see table 4.3) to later also have comparable QoE results with respect to different coding types with similar configuration. The variation in packet size for coding algorithms as the Eapt-X is made possible due to their relatively small frame length which is explicitly designed to be groupable in one packet. Generally no packet loss concealment was enabled in this thesis. There was always only a silence insertion chosen for missing audio data due to packet loss.

3.2.4 Input Audio Material

In ITU-T recommendations P.834.1 [43] a set of 4 speech files (2 female and 2 male voices) is used for the evaluation, available in different languages (French, Japaneses, American English). For the wideband case in this work, not only speech signals but also audio signals should be used for the evaluation of codec quality. Inspired by proposals from [61] (using 4 audio signals) and [91] (using 5 audio signals) different appropriate audio test signals were chosen, which reflect the dedicated application of professional audio contribution for broadcasting.

Generally, the set of signals shall be small for not extending the effort unnecessary but large enough to represent the most cases of possible signal characteristic. For audio signals, there can be distinguished roughly between music, instruments and vocals. Further, the type of music (i.e. classic or pop music, with voice or without) or type of instrument (i.e. more harmonic or more transient character) as well as vocal gender is are important characteristics.

Thereby also the signal durations has to be taken into account. In comparable QoE evaluations, durations of 5-20 seconds found to be appropriate. In [91], audio signals of 5-10 seconds duration were taken, in [65] 20 seconds of speech signals (speech evaluation) were taken and in [61] 1 minute of different audio and speech signals were used but cuted into 3 parts of 20 seconds.

Signal Description	Filename	Duration [s]
Jazz music with speedy drums, piano and walking bass	add_count.wav	20
German spoken male voice	ref_smg.wav	16
Opera with classical vocal singing	ref_sop.wav	42
Female vocal singing, (Suzanne Vega: “Tom’s Diner”)	ref_veg.wav	21
String orchestra, very harmonic	add_str.wav	28
Percussive castanets rhythm	ref_cas.wav	15

Table 3.2: Audio material for the QoE rating and $I_{e,FB}$ derivation.

Finally, a set of 6 stereo audio signals were chosen for the quality evaluation in this work, listed in table 3.2. The signals were taken from the set of PEAQ audio test signals of the the famous quality assessment tool Opera by Opticom [14] and are 16bit linear PCM coded signals with 48 kHz sampling rate. Most of them are used in different standard testsignal databases for quality evaluation from organizations as MPEG or ITU [27].

The signal durations were taken as they are. They are slightly longer as in the mentioned previous works but therefore also better suited for the later packet loss evaluation (see chapter 3.4.3).

3.2.5 Loss Model Parameters

The selection of the different loss process conditions to model, whose QoE impairment was later analyzed, were chosen based on the proposals in [65] (for voice over IP) combined with the proposals in [91] (for audio over IP) but modified for the dedicated application. Finally, a better resolution for small loss rates was found to be necessary for professional audio over IP analysis while the evaluation with respect to very high loss ratios ($\rho > 10$ %) was found to be neglectable because of the surely already to high impairments on audio quality at these loss ratios for the studied application. For professional audio demands, these impairments are in priniple not acceptable.

For modeling the network loss process, the 2-state Markov model described in chapter **??** was used with its parameters which fully describe a short-term loss process allowing for loss correlation. Based on the previous analysis of appropriate loss process conditions, finally the following parameter sets for the two parameters of the Markov model were chosen,

$$\text{Set for } \rho : \mathscr{P} = \{\ 0,\ 0.2,\ 0.5,\ 1.5,\ 3,\ 5,\ 10\ \}\ [\%] \tag{3.7}$$

$$\text{Set for } p_c : \mathscr{Q} = \{\ 0,\ 0.15,\ 0.3,\ 0.5\ \} \tag{3.8}$$

where ρ is the packet loss ratio and p_c the conditional loss probability. This selection leads to four different network packet loss processes considering burstiness independent of the current amount of packet loss.

In table 3.3 the resulting characteristics of these four configurations are given. From the different conditional loss probabilities p_c, the respective mean loss period (MLP) from eq. (**??**) and the main probabilities for the occurrence of a loss burst of length k, $p_{L,k}$ from eq. (**??**), are depicted. Therefore, the single loss probability $p_{L,1}$ and the different burst probabilities for two consecutive losses ($p_{L,2}$), three consecutive losses ($p_{L,3}$) and four consecutive losses ($p_{L,4}$) are given, while higher burst length are neglected here. The resulting mean loss periods are also comparable to the respective parameter for the

Conditional Loss Probability p_c	Mean Loss Period (MLP)	$p_{L,1}$	$p_{L,2}$	$p_{L,3}$	$p_{L,4}$
0	1.0	1.0	0.0	0.0	0.0
0.15	1.18	0.85	0.12	0.02	0.003
0.3	1.43	0.7	0.21	0.06	0.02
0.5	2.0	0.5	0.25	0.13	0.06

Table 3.3: Loss process conditions for different conditional loss probabilities p_c for arbitrary packet loss rate.

evaluation chosen in [61].

While the mean loss period only depends on the conditional loss probability p_c, the mean loss distances (MLD) from eq. (**??**) depend also on the current packet loss ratio. The resulting mean loss distances (MLD) for the selected model parameters are listed in table 3.4.

Cond. Loss Probability p_c	$\rho = 0.2$ %	$\rho = 0.5$ %	$\rho = 1.5$ %	$\rho = 3$ %	$\rho = 5$ %	$\rho = 10$ %
0	499.0	199.0	65.7	32.3	19.0	9.0
0.15	587.1	234.1	77.3	38.0	22.4	10.6
0.3	712.9	284.3	93.8	46.2	27.1	12.9
0.5	998.0	398.0	131.3	64.7	38.0	18.0

Table 3.4: Mean loss distances (MLD) in packets for the different loss process conditions.

It has to be mentioned that finally the promising approach from [70] for speech coding algorithm evaluation was not included for the loss model, because there a slightly different approach was chosen which was not found to be suggestive in the environment of this thesis. The loss model in [70] is not based directly on parameters describing loss ratios and loss correlation of an stochastic model. Instead, they fixed deterministic mean loss distances and related them to different burst sizes from where the different loss ratios could be obtained.

Therefore, they investigated the impairments directly with respect to different deterministic burst lengths and distances while therewith the relation to real loss traces observed on real packet networks was depreciate which was not desired in the present thesis while the impairments most closely to impairments to reality should be evaluated. Nevertheless, the more deterministic loss model in [70] is better suited for a rigorous performance analysis of the packet loss impairments on different coding algorithms in a way loosing the direct relation to their appearance in reality.

3.3 Experimental Test Setup

In the following, the experimental test setup and its components including the packet loss generation is described. The setup principle with its main components is depicted in fig. 3.1. There it can be seen that for the objective audio quality assessment as usual an undistorted reference audio signal from the testfile database (see chapter 3.2.4) is made available as well as the audio signal degraded by packet loss introduced by the IP network emulator. The impaired signal is obtained after a transmission of the audio testfile over an audio over IP infrastructure with first the audio encoder and RTP/UDP/IP packetizer followed by the IP network emulator which forces packets to be lost based on the desired loss process to

Coding Algorithm	MPEG L2	MPEG L2	Eapt-X	Eapt-X	Eapt-X	Eapt-X	Eapt-X	ITU-T G.722
Bitrate [kbit/s]	384	256	384	384	256	256	64	64
Samplerate [kHz]	48	48	48	48	32	32	16	16
Δt_p [ms]	24	24	21.25	5.25	24	8	16	20
Packet rate r_p [pps]	41.7	41.7	46.5	190.5	41.7	125	62.5	50
$r_\rho(\rho = 0.2)$ [pps]	0.08	0.08	0.09	0.38	0.08	0.25	0.13	0.10
$r_\rho(\rho = 0.5)$ [pps]	0.21	0.21	0.23	0.95	0.21	0.63	0.31	0.25
$r_\rho(\rho = 1.5)$ [pps]	0.63	0.63	0.70	2.86	0.63	1.88	0.94	0.75
$r_\rho(\rho = 3)$ [pps]	1.25	1.25	1.40	5.71	1.25	3.75	1.88	1.50
$r_\rho(\rho = 5)$ [pps]	2.08	2.08	2.33	9.52	2.08	6.25	3.13	2.50
$r_\rho(\rho = 10)$ [pps]	4.17	4.17	4.65	19.05	4.17	12.50	6.25	5.00

Table 3.5: Packet loss rates r_ρ (packets lost per second) for the evaluated coding algorithms and configurations.

emulate.

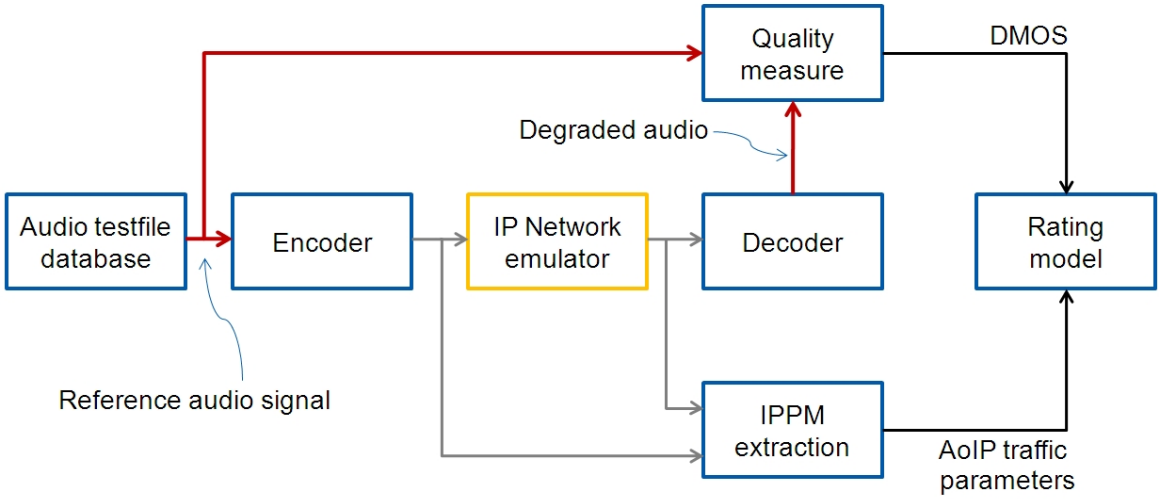


Figure 3.1: QoE evaluation environment and QoS/QoE metrics extraction.

The IP audio stream is then decoded for its handover in normal PCM audio format to the quality assessment module. Furthermore, the IP stream is observed by network analyzers ahead and after the IP network emulator to extract important IP performance metrics (IPPM) which can be delivered to a rating model as well as the determined QoE metric from the quality assessment tool. Throughout the measurement procedure, the IP performance measurement is also used for a monitoring of the connection to ensure that at the network layer only the impairments introduced by the IP network emulator are present.

3.3.1 System Components

For the audio coding and decoding as well as the IP stream generation from the digital audio signal, dedicated professional hardware audio codecs were used. Three different devices were found to be appropriate for the desired evaluations and each were used for the coding of different algorithms. The

three devices are the same as used for the traffic analysis (see chapter **??**).

The Fast Ethernet connection of the at a time used coding device were plugged on a Cisco Systems Catalyst 2950 series intelligent Fast Ethernet (100 Mbit/s) switch as well as one Fast Ethernet connection from the IP network emulator and one of the network analyzer. The second Ethernet connection of the IP network emulator was plugged on a second Cisco switch of the same type as well as the decoding device and the other network analyzer. Furthermore, a personal computer for the remote control of the audio coding devices were added which was only active when no evaluation was done. For the network protocol analysis again Wireshark was used as in chapter **??**.

The audio playout and recording was done simultaneously in AES/EBU digital audio with the numerical computing environment Matlab [88] using ASIO drivers. The later QoE analysis were also done with Matlab while a wideband-modified Matlab implementation of the PESQ algorithm from [51] was used (see chapter 2.2.2). Always bidirectional SIP calls where established if possible[2] while only the audio signal of one direction were recorded for later QoE analysis.

The Matlab audio processing as well as both used Wireshark instances ran on the personal computer Dell Precision R5400 with the professional digital audio card Digigram PCX881HR on 48 kHz sampling rate (performance details in chapter **??**). Owing to the lab conditions, the latency and jitter monitoring could be performed by only using Wireshark and therefore a complicated synchronization of the network devices could be avoided (concept introduced in chapter **??**).

All involved digital audio devices were synchronized with the help of an RME ADI-8DD digital audio format converter which was only used for the AES/EBU synchronization. For the analog audio monitoring a pair of the professional loudspeakers Genelec 1029A were used.

As IP network emulator for the packet loss generation the ZTI NetDisturb Enhanced Edition [96] was used running on a Spectra personal computer (Intel Pentium 3 (983 MHz), 512 MB RAM, OS: WinXP Prof. SP3) equipped with two Intel PRO/100S fast Ethernet network adapters. NetDisturb acts as a bridge between the two Fast Ethernet segments introduced above and operates bi-directional packet transfer. Generally, it can generate diverse impairments over IP networks, such as latency, delay, jitter, bandwidth limitation, loss, duplication and modification of the packets [96]. NetDisturb was used because this tool has the capability to include own packet loss processes while no application with comparable functions were found. Normally, only a limited set of basic statistical distributions is selectable in an emulation environment which is not sufficient for the desired analysis . Furthermore, NetDisturb is used by the Deutsche Telekom Laboratories for IP evaluations as mentioned in [64]. The packet loss generation is treated in further detail in the following.

3.3.2 Packet Loss Generation

The different modeled packet loss processes which should impair the IP audio streams in a realistic way are based on a 2-state Markov model introduced in chapter **??**. But Markov models are normally not available for selection in network emulators, also the chosen NetDisturb tool only provides basic statistical distributions for the packet loss processes but it provides the opportunity to include own deterministic packet loss traces. Thereby a chain of 1's (indicating packet is lost) and 0's (indicating packet is served) stored in a normal textfile is loaded and applied on the desired stream. The concept of NetDisturb allows the definition of different IP flows and for each a dedicated impairment can be defined after the selection

[2] At the time of the evaluations, the firmware of the APT Oslo were not able to handle SIP connections.

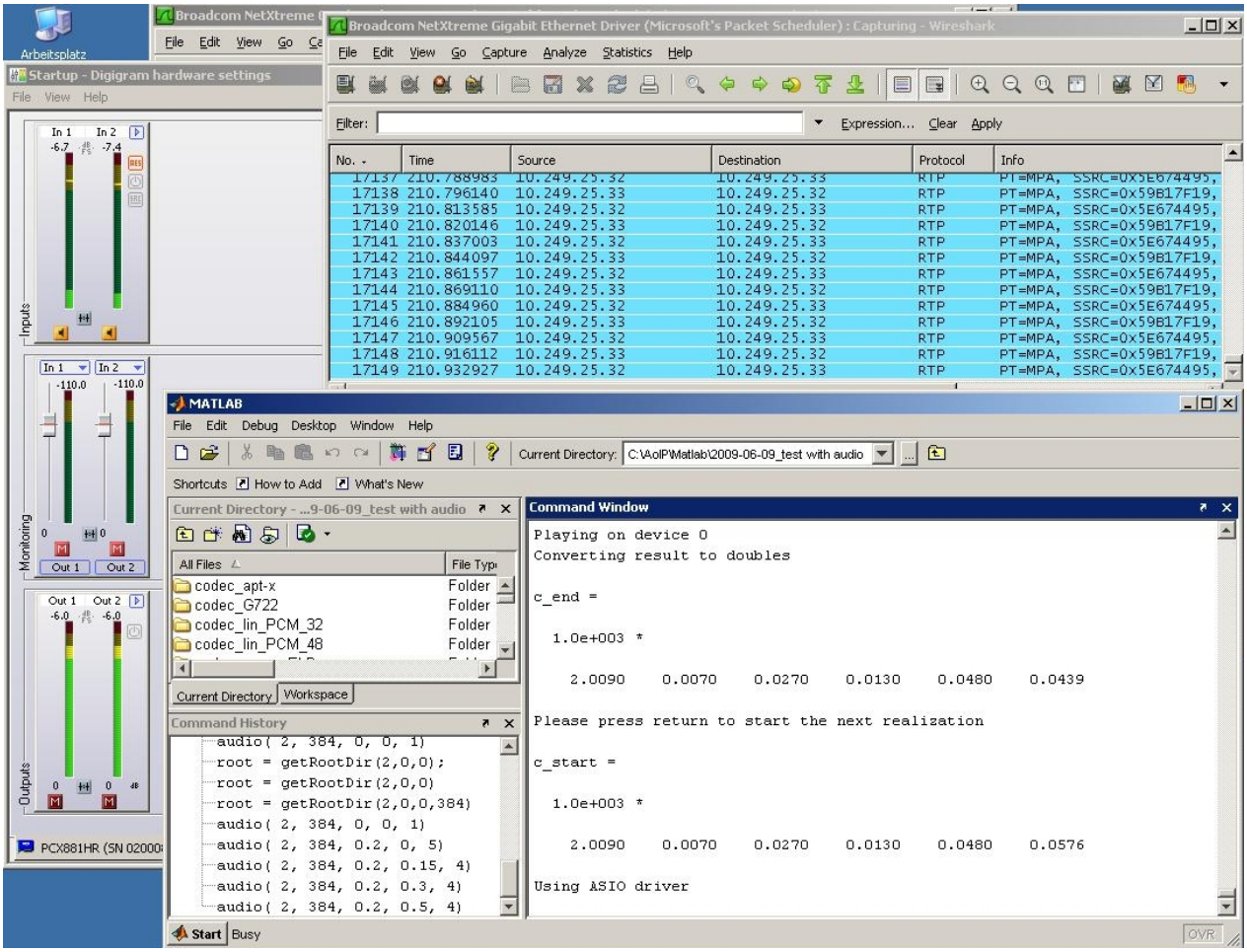


Figure 3.2: Screen shot of the data capturing process.

of an appropriate packet filter based on header information.

This was also useful here because therewith the RTP audio stream could be filtered and the packet loss impairments based on the Markov model where only applied on this media flow and possible other IP packets in the network as control messages did not disturb the desired statistical characteristics of the packet loss with respect to the audio stream.

The textfiles including the packet loss process realizations were generated using Matlab. The 2-state Markov model were implemented in this numerical computing environment and the necessary 24 different loss traces resulting from all possible combinations of the parameter sets for the packet loss ratio ρ and the conditional loss probability p_c were exported in the necessary format in textfiles. These textfiles were again loaded in Matlab for verification and their statistics based on the estimates defined in chapter **??** were computed. Then, realizations with high deviations in the estimates to the desired process characteristics (i.e. the theoretical probability values) were excluded and new realizations were generated until one with the acceptable deviations was obtained.

The loss indicator chains were chosen to have a length of 10000 events which was found to be appropriate to avoid a periodicity impact destroying the desired statistical characteristics. Even more, because the verification step necessarily was included. The selected length converted in time results normally in the range of 3-4 minutes audio stream duration for usual packet rates (see the traffic analysis results in

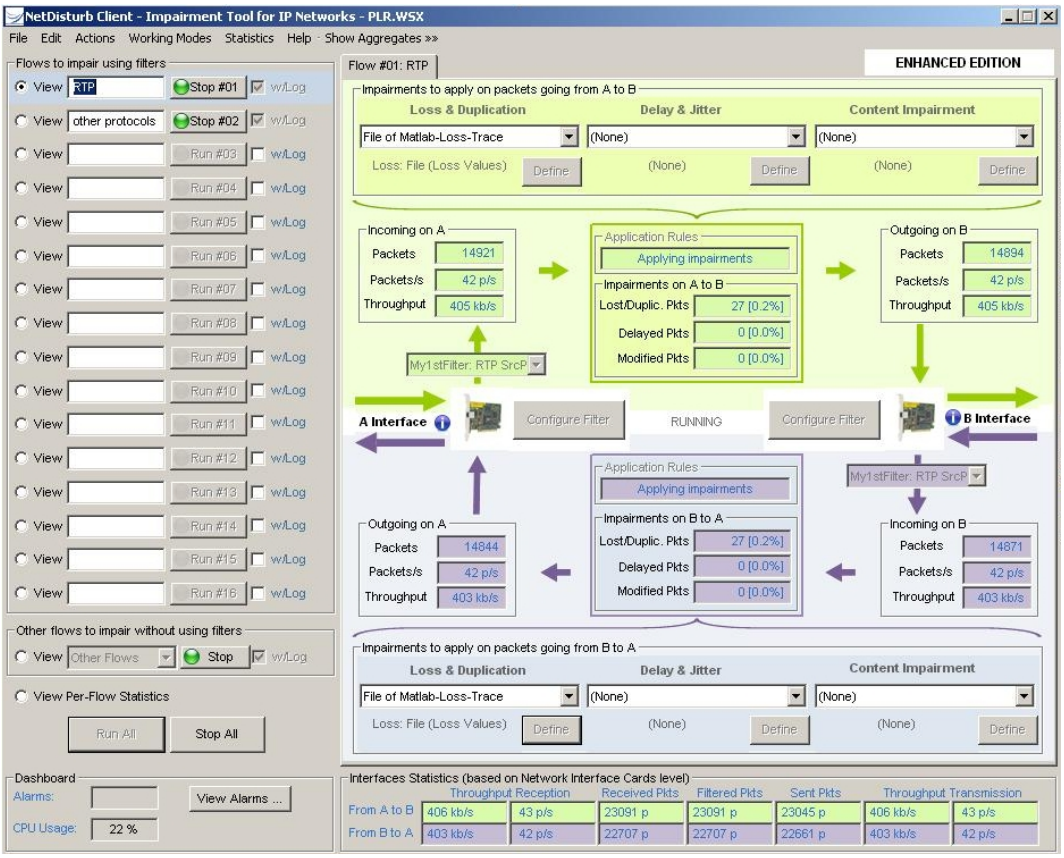


Figure 3.3: Screen shot of the network emulator running.

chapter **??**).

Considering the worst case for ACIP applications, a minimum of 4 ms audio frame duration respective inter-packet time results if linear PCM audio coding is chosen for the transmission over IP [15]. This leads to a maximum packet rate of 250 packets per second and a relative time duration of the loss indicator chain of 40 seconds which could be problematic. Nevertheless, finally the 4ms packetized linear PCM audio coding was not evaluated with respect to packet loss impairments on the perceived quality of service.

3.4 Data Acquisition and Analysis

Finally, the audio quality assessment of the coded test signals impaired with different packet loss processes was accomplished using WB-PESQ for the MOS LQO rating. For each selected coding algorithm type and configuration each loss model realization was employed.

First, the used codec device as well as the IP network emulator NetDisturb had to be configured. Then, the 6 different testfiles were played consecutively on each experimental setup configuration while the system output were simultaneously recorded, repeating the procedure 4 times to enhance the reliability of the outcomes. Then, the resulting WB-PESQ scores were averaged over the testfiles[3] as well as the trials, as usual, i.e. in [43] and [14]. If the testsignal had a channel configuration of stereo ($ch = 2$), also the average of the left and right channel were performed. Furthermore, tcpdump-traces captured with

[3] The performance of the WB-PESQ algorithm for $\rho = 0$ with respect to the different testfiles for each examined coding algorithm can be observed in Appendix A.

Wireshark for each connection initiated for every experiment anew were stored.

The concatenation of the audio testfile chain has a duration of approximately 2:30 minutes. While the generated packet loss process realizations where designed to cover normally 3-4 minutes audio stream durations (see above), the periodicity problem can be neglected here. The number of 4 trials for each of the 25 network emulator configurations (24 loss processes and $\rho = 0$) was chosen as compromise between time effort for the assessment and necessary trials for an accurate assessment of the quality metric. In [65], similar loss processes were used but each test was performed 8 times instead for even more significant results.

For the worst case with packet loss ratio $\rho = 0.2$ % and conditional loss probability $p_c = 0.5$ a mean loss distance (MLD) of 998 packets is obtained (see table 3.4). This results in a time-related mean loss distance of approximately 18 seconds. While the mean duration of the testfiles is approximately 24 seconds, obtaining meaningful results is assumed from playing them 4 times with random audio starting times with respect to the loss process realization. The RTP stream was always initiated approximately 5 seconds before the start of the audio files in Matlab.

Unfortunately, it was not possible to include accurate timestamps for the audio playout and recording, therefore a comparison or assignment of specific packet losses in the packet trace to resulting audio failures in the recorded signal could not be accomplished. Nevertheless, finally this was also of less importance here because the correlation of loss and audio failure is also indicated by the WB-PESQ score and especially packet loss concealment methods were not examined in this work. For packet loss concealment methods it is especially interesting how much perceivable audio failures result from a specific loss ratio, i.e. which amount of network packet losses are concealed by a algorithm.

3.4.1 Procedure for Data Mining

The set of averaged WB-PESQ scores obtained for the clean and impaired testfiles were first stored with respect to each packet loss process realization for which the selected audio material currently was evaluated. Then, they had to be resorted for the goal of constructing packet loss ratio dependent functions of the WB-PESQ scores as

$$MOS_{\mathrm{PESQ}}(\mathbf{P}_{\mathrm{loss}}, \mathbf{c}, L) = f(\rho) \tag{3.9}$$

while also the different burstiness evaluations had to be considered. Therefore the relation in eq. 3.9 was assembled for each of the chosen four different mean loss periods (MLP) (derived from the conditional loss probability p_c as to eq. (**??**)).

Now, first the coding algorithm only impairments on the listening quality shall be discussed, which were assessed by the sorted WB-PESQ scores results with respect to $\rho = 0$.

3.4.2 Coding Algorithm Impacts

The mean ratings of the WB-PESQ algorithm with respect to the different selected coding algorithms and configurations are depicted in fig. 3.4. The differentiations for some coding algorithms with respect to the packet size respective the inter-packet time were excluded here because without the existence of packet losses as ensured here, they have no impact on the quality rating.

The WB-PESQ scores for the wideband and narrowband coding algorithms (“EAPTX 64k”, “G.722”, “G.711” in fig. 3.4) was taken as meaningful compared with subjective evaluations [64] and subjective

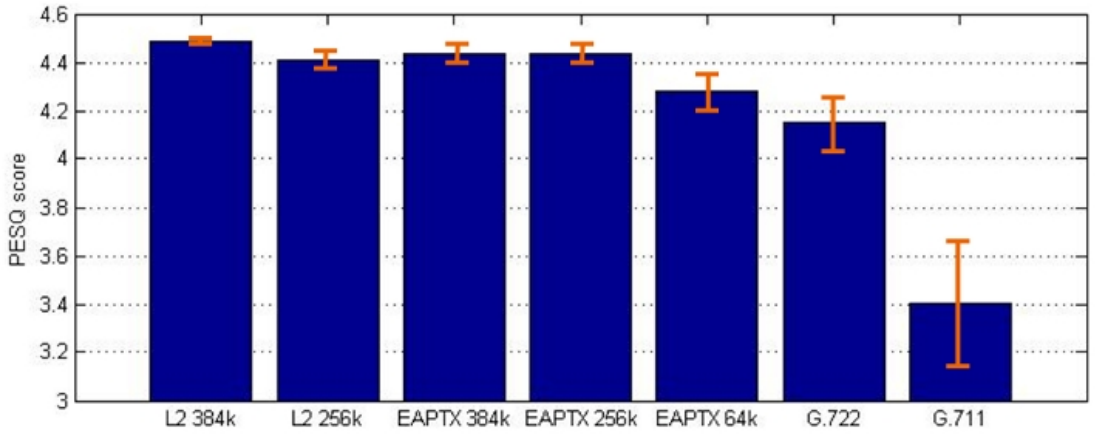


Figure 3.4: Obtained WB-PESQ scores for different AoIP coding algorithms (with error-free network, $\rho = 0$) and their 95 % confidence intervals (orange).

impressions by the author. As expected, the assessed WB-PESQ scores for high quality codecs on fullband ($F_s = 48$ kHz) and ultra-wideband ($F_s = 32$ kHz) configuration were not as reasonable as desired and correlate bad with subjective impressions from the author even if the principal relations are not totally wrong with one exception: the WB-PESQ rating of the fullband Eapt-X algorithm with 384 kbit/s coding rate ($F_s = 48$ kHz, "EAPTX 384k") and the rating for the ultra-wideband Eapt-X algorithm with 256 kbit/s coding rate ($F_s = 32$ kHz, "EAPTX 256k") are very similar which is inappropriate with respect to subjective impressions and the respective ODG output of the PEAQ algorithm where a clearer difference between the ratings of the two mentioned Eapt-X coding algorithms can be seen.

An explanation of the observation that the WB-PESQ algorithm can not deal with the perceptual differences of Eapt-X with 48 kHz samplerate and Eapt-X with 32 kHz samplerate is, that WB-PESQ is only evaluating at a samplerate of 16 kHz, therefore it can not determine impairments on frequencies above 8 kHz (see chapter 2.2.2. Besides the missing coding algorithm impact rating for higher frequencies, the relative bandwidth impairment of using ultra-wideband instead of fullband comprehensively can not be assessed by WB-PESQ.

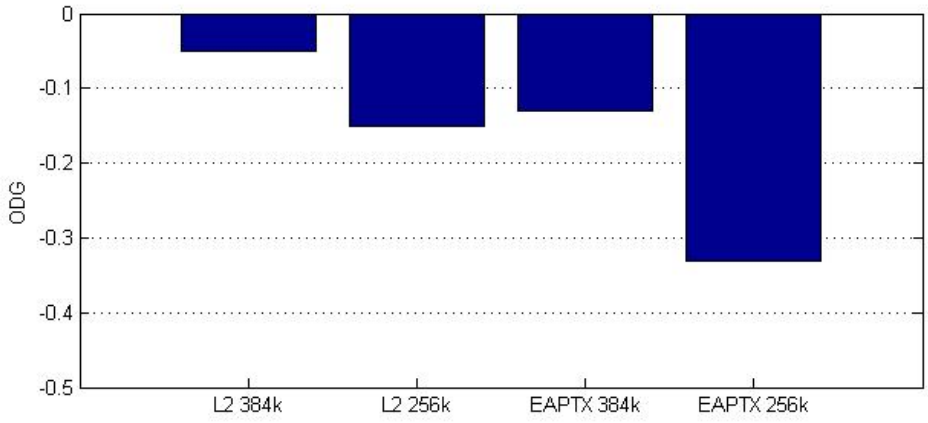


Figure 3.5: PEAQ Objective Difference Grades (ODG) for different AoIP coding algorithms, the standard deviation is almost constant at $\sigma = 0.26$ on average (values taken from [47]).

But the WB-PESQ scores with respect to the mentioned Eapt-X algorithms can be used for another interesting insight. Here, the specific results indicate, that both coding algorithm configurations have similar impairments on the lower frequencies which is reasonable because the reduction in coding rate is obtained by lowering the samplerate with the same relation ($\frac{348}{256} = \frac{48}{32} = \frac{3}{2}$). Therefore apparently the algorithm is not changing the coding efficiency with respect to different subbands when it reduces the bitrate from 384 kbit/s to 256 kbit/s but it simply excludes the information of the higher frequency components and reduces therefore the necessary capacity for transmitting the remaining information which enables the bitrate reduction. This reasonable explanation states also that the WB-PESQ with respect to its limited capabilities gives relative accurate results.

In fig. 3.5 also some Objective Difference Grade (ODG) results of the PEAQ algorithm are shown, taken from [47]. Here, only the respective results for the coding algorithm configurations fullband ($F_s = 48$ kHz) and ultra-wideband ($F_s = 32$ kHz) with respect to the audio bandwidth are included. This is due to the observation, that the PEAQ algorithm can not give reliable results for narrowband and wideband coding algorithm configurations, as expected, because it was designed to rate only relatively small impairments on the audio quality due to coding artifacts.

In general, the used PEAQ results are known to correlate well with subjective impressions and can be seen as reliable results. The observable relations on the ODG scale show meaningful clearer differences in the rating of the different codecs included here. The overall analysis of the WB-PESQ and PEAQ results with respect to a lossless transmission ($\rho = 0$) support the assumption, that a combination of the two approaches could be used for an instrumental based QoE prediction rating framework for audio quality assessment. This is discussed and finally exploited in the following chapter 4.

The performance of the WB-PESQ algorithm for $\rho = 0$ with respect to the different testfiles for the examined coding algorithm can be additionally be observed in Appendix A. These results are interesting because it gaves an insight on the ability of a coding algorithm to process an audio signal independent of its specific characteristic, i.e. speech-like, more harmonic or transient. Therefore, ideally a coding algorithm has a perceptual degradation independent of the signal characteristics which is normally not the case because coding algorithms are normally optimized with respect to a dedicated signal characteristic (see chapter **??**).

3.4.3 Packet Loss Impacts

In the following, the results of the WB-PESQ processing for the different coding algorithm types and configurations with respect to an possible audio quality impairment due to IP packet loss are presented and analyzed. The results also include the general coding impairment because it was found to be perceptual more meaningful to compare the entire results and not to assume an additivity of the impairments in the PESQ score dimension which would enable a subtraction of the results with respect to lossless transmission. While this could be in general be possible, the comparability of the resulting curves would be not as good as with respect to the overall results.

First, different reasonable coding algorithm performance comparisons are discussed based on a nonlinear curve fitting of the obtained WB-PESQ scores for discrete packet loss ratios ρ. Later, for each coding algorithm a two-dimensional bar plot is presented, which gives a coding algorithm dependent overview on all performed perceptual ratings of the performance due to both the packet loss ratio ρ as well as the mean loss period (MLP), therewith incorporating the normally observed burstiness of packet losses on IP networks.

For the coding algorithm performance comparisons with respect to the packet loss ratio ρ (depicted in fig. 3.6, fig. 3.7 and fig. 3.8), the results for the different conditional loss probabilities p_c were first averaged over the whole set of these parameter for each packet loss ratio ρ from the respective set from the packet loss process model (both sets are defined in chapter 3.2.5). It is undertaken that therewith a more long-term description can be obtained for an simplified overall comparison of single curves. Therefore, an uniform distribution of the different loss processes were assumed. This results in an overall averaged mean loss period (MLP) $\mu_{MLP} = 1.4$ supposed here.

The averaged discrete results now only depending on the packet loss ratio ρ as to eq. 3.9 were then interpolated by non-linear least-squares curve fitting in Matlab. Finally, a 3-parameter model was chosen for the description, inspired by [91] where an 2-parameter model was used. The model for the continuous WB-PESQ scores with respect to the packet loss ratio ρ can be formulated as

$$MOS_{PESQ}(\rho) = a \cdot (b \cdot \rho)^c + MOS_{PESQ}(\rho = 0) \tag{3.10}$$

with the coding algorithm type and configuration dependent interpolation parameters a, b, and c while $0 < c < 1$. The parameters for the different schemes can be found in table 3.6, including the *Residual Sum of Squares (RSS)* (i.e. the squared 2-norm of the errors) as a descriptor for the goodness of fit of the interpolation.

Coding Algorithm	**MPEG L2 1**	**MPEG L2 2**	**Eapt-X 1a**	**Eapt-X 1b**	**Eapt-X 2a**	**Eapt-X 2b**	**Eapt-X**	**G.722**
Bitrate [kbit/s]	384	256	384	384	256	256	64	64
Samplerate [kHz]	48	48	48	48	32	32	16	16
Δt_p [ms]	24	24	21.25	5.25	24	8	16	20
Packet rate r_p [pps]	41.7	41.7	46.5	190.5	41.7	125	62.5	50
WB-PESQ scores	4.49	4.41	4.44	4.44	4.44	4.44	4.28	4.15
a	-0.9385	-0.8716	-1.3980	-1.4649	-1.4550	-1.0964	-1.0061	-0.9321
b	0.8983	0.7870	1.7684	1.2087	1.8773	1.1227	1.0074	0.9082
c	0.5128	0.5268	0.3468	0.2770	0.3257	0.4208	0.5033	0.4839
RSS	0.0133	0.0236	0.1544	0.1416	0.0761	0.1306	0.2313	0.0363

Table 3.6: Interpolation parameters a, b, c and the Residual Sum of Squares (RSS) of the curve fitting for the WB-PESQ scores with respect to the packet loss ratio ρ for the averaged mean loss period (MLP).

The following comparisons were selected based on similar properties of the chosen coding algorithm sets:

1. Both MPEG L2 configurations versus both Eapt-X configurations in fig. 3.6
   (all $r_c = 384$ kbit/s $F_s = 48$ kHz but different packet sizes),
2. Eapt-X versus G.722 (both $r_c = 64$ kbit/s $F_s = 16$ kHz) in fig. 3.7,
3. Results of all Eapt-X configurations with $F_s \geq 32$ kHz in fig. 3.8 .

These comparisons were enabled due to the sophisticated choice of coding algorithms characteristics for this evaluation. The detailed configurations can be checked in table 4.3.

Overall, all results seem in principle meaningful in comparison to subjective impressions, while the quality rating prediction performance of the WB-PESQ algorithm for the cases treated here should also

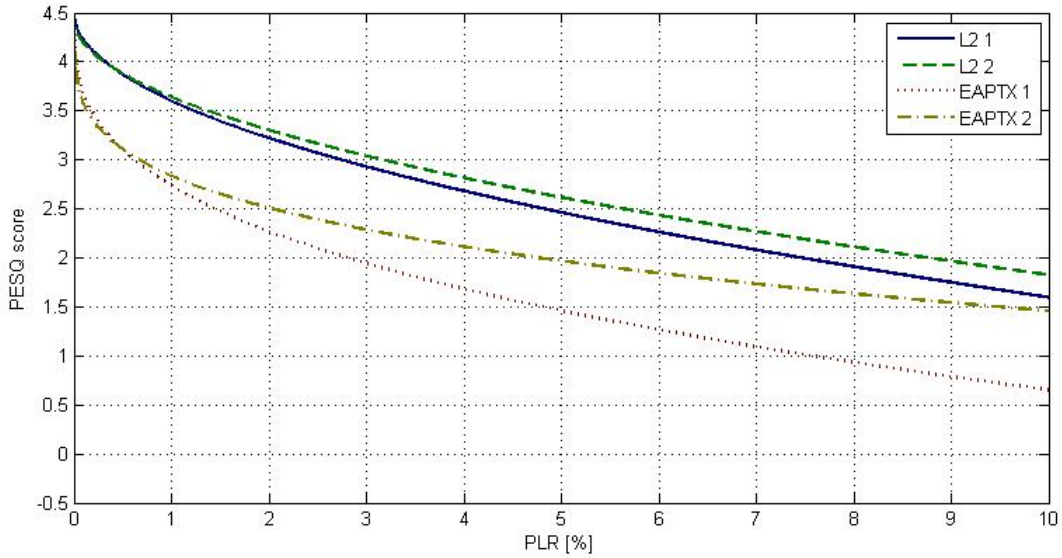


Figure 3.6: PESQ score with respect to packet loss ratio comparison between MPEG L2 and Eapt-X at 384 kbit/s coding rate (numbering indicating the different configurations).

be investigated by comparison with comprehensive subjective tests for reliably confirmation.

In fig. 3.6 it can be seen that in general the MPEG L2 algorithms is said to perform better under packet loss in comparison with the respective Eapt-X algorithms based on the obtained WB-PESQ scores. Thereby it has to be mentioned that the hardware coding device used for the Eapt-X evaluation had sirious problems already to cope with a small amount of packet loss. Single losses caused already clock synchronization problems between the internal and the RTP clock, therefore the loss gaps often were longer than the theoretical lost audio signal parts due to one packet loss. For higher loss ratios, the problem was less and the slopes of the decreasing quality rating curves for both MPEG and Eapt-X are similar for higher loss rates.

The comparison of the G.722 coding algorithm with the Eapt-X coding algorithm for a similar configuration can be seen in fig. 3.7. The results there indicate, that the G.722 performs slightly better than Eapt-X for higher loss rates while the Eapt-X is slightly better for very small loss amounts, also owed to its better rating for lossless transmission. It is suspected that the Eapt-X simply was not designed to cope also with higher loss rates uncommon in digital transmission over dedicated lines, the main application for Eapt-X coding in broadcasting.

The WB-PESQ scores for different Eapt-X configurations are compared in fig. 3.7. Here, for different bitrates and audio bandwidth the influence of the packet size respective the inter-packet time can clearly be seen as the configurations with the smaller packet sizes are for higher loss ratios nearly one point WB-PESQ score better rated than the configurations with longer packet size while apparently the different coding rates and bandwidths do not have such an impact here.

Finally, overall two-dimensional bar plots of the different discrete mean WB-PESQ scores with respect to the packet loss ratio ρ and the mean loss period MLP are presented for each of the 8 evaluated coding algorithm types and configurations in fig. 3.9 and fig. 3.10. The $\rho = 0$ case where excluded here for clarity of the diagrams. They show relatively reasonable ratings but with one exception for the Eapt-X algorithm ($r_c = 384$ kbit/s $F_s = 48$ kHz) with the relatively small packet size ($L_c = 256$) at $\rho = 0.5$ and $MLP = 1.18$ as well as $MLP = 1.43$. Here the clock synchronization problems mentioned above can also

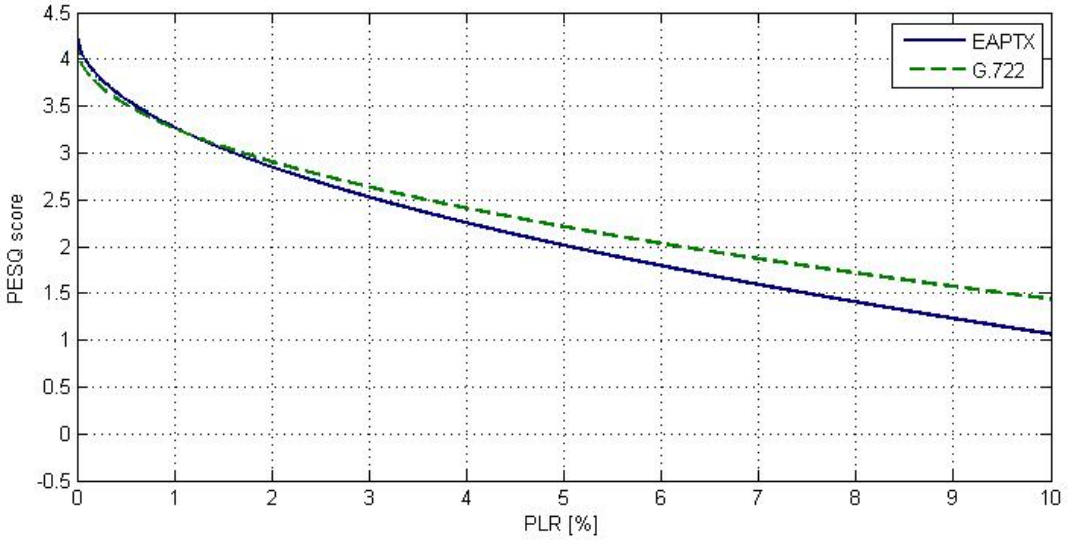


Figure 3.7: PESQ score with respect to packet loss rate comparison between Eapt-X and G.722.

be detected.

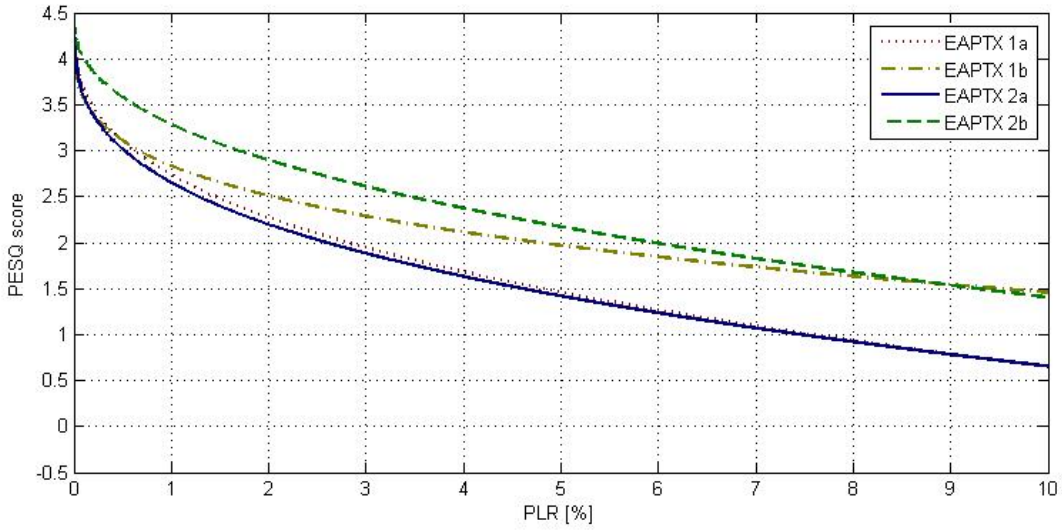


Figure 3.8: PESQ score with respect to packet loss rate comparison for Eapt-X (numbering indicating the different configurations, 1: $r_c = 384$ kbit/s, 2: $r_c = 256$ kbit/s).

4 Proposed QoE Rating Model

We chose the E-model as the basic framework for our broadband QoE rating model for ACIP because it is inherently extendable to fullband quality rating and allows the incorporation of both equipment related impairments such as coding characteristics and network packet loss as well as the impairment due to transmission delay. First, we simplified the general R-factor formula of the E-model focusing on coding and IP transmission impairments, resulting in

$$R_{FB} = R_{0,FB} - I_{e,eff,FB} - I_{d,\mathrm{ACIP}} \tag{4.1}$$

with the fullband basic transmission rating factor $R_{0,FB}$ and the delay impairment factor $I_{d,\mathrm{ACIP}}$, dedicated to audio contribution. The fullband equipment impairment factor $I_{e,eff,FB}$ can be further separated to

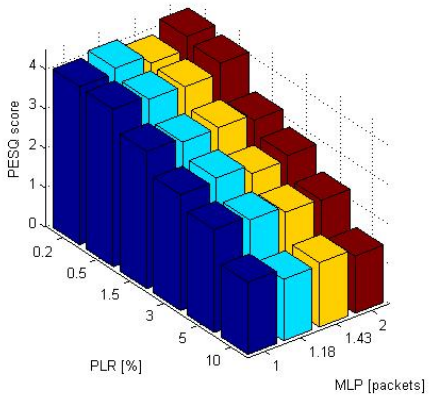
$$I_{e,eff,FB} = I_{bw,FB} + I_{res,FB} + I_{loss,FB} \ , \tag{4.2}$$

with the bandwidth impairment factor $I_{bw,FB}$ representing linear bandwidth distortions for the fullband case, while $I_{res,FB}$ represents non-linear coding distortions. The loss impairment factor $I_{loss,FB}$ includes a continuous QoE model for the loss-only perception, obtained by non-linear least-squares curve-fitting of discrete experiment outcomes with a logarithmic model dependend on the packet loss ratio ρ [84],

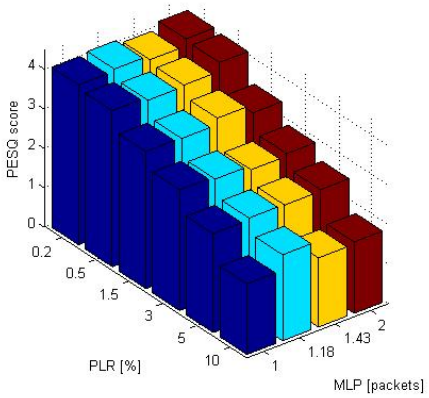
$$I_{loss,FB} = a \ ln((1+b)\rho) \ . \tag{4.3}$$

Hence, our ACIP QoE rating framework includes separate models for the delay, loss, coding and bandwidth impairments, which are the most important ones for audio contribution applications. We excluded the advantage factor for now and also the simultaneous impairment factor I_s of the E-model is neglected, because a perfect audio source is assumed to be present at the input.

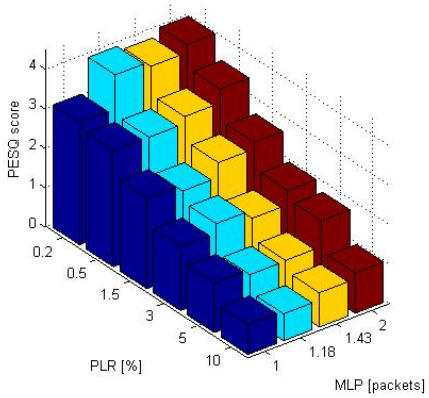
The challenge is now to find a reasonable fullband basic transmission rating factor $R_{0,FB}$ as well as the mathematical models for the delay impairment factor $I_{d,\mathrm{ACIP}} = f(d_{\mathrm{e2e}})$ and the fullband equipment impairment factor $I_{e,eff} = f(\mathbf{P}_{\mathrm{loss}}, \mathbf{c}, L)$.



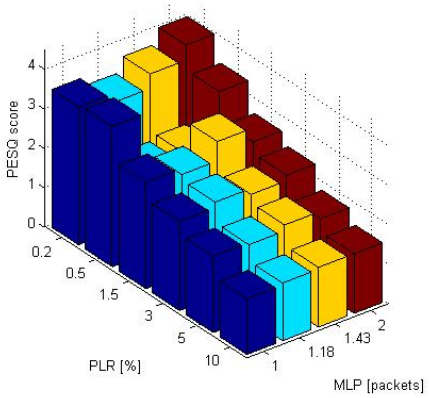
(a) MPEG L2 384 kbit/s $L_c = 1152$.



(b) MPEG L2 256 kbit/s $L_c = 384$.



(c) Eapt-X 384 kbit/s $L_c = 1024$.



(d) Eapt-X 384 kbit/s $L_c = 256$.

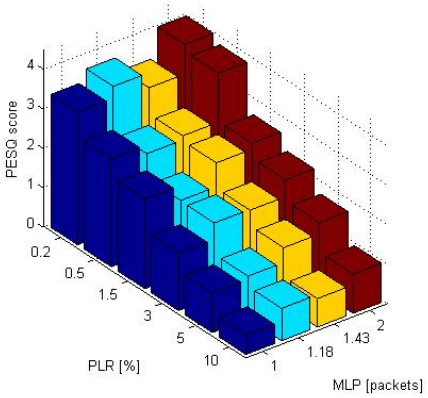
Figure 3.9: PESQ scores for loss impact analysis for different AoIP coding algorithms (1 of 2).

4.1 The Fullband *R*-Factor

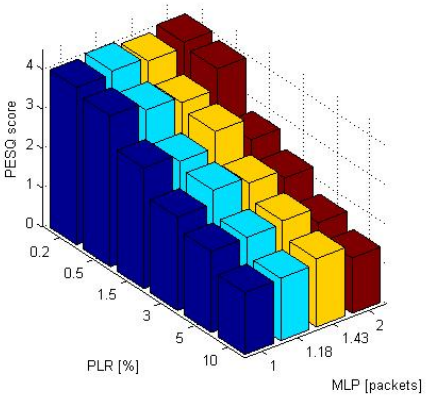
In the following, we first derive the necessary *R*-factor scale extension for fullband quality rating exploiting the bandwidth impairment model approach (see section 2.2.3). For this, we extrapolate the model for wideband speech communication to the full audio bandwidth in order to derive $R_{0,FB}$. This is justified by a comparison of the model outcomes for different audio bandwidths up to fullband with subjective results from the literature as from EBU research [2], objective measures from the PEMO-Q algorithm [23] and subjective impressions from our experts.

As anchor, the basis for the derivation of the wideband improvement was taken: in [64], it is stated that the wideband advantage over narrowband is in the range of 1.3-1.5 points *MOS*, based on the evaluations of the author from [64]. We compared this result with objective difference grade (*ODG*) results from the PEMO-Q algorithm [23] for bandwidth-only impairments, which is the best suited instrumental method for evaluating high impairments on audio quality [5], such as the bandwidth restriction by downsampling, i.e. from 48 to 16 kHz sampling frequency. For the PEMO-Q evaluations, a 16bit linear PCM testfile was evaluated with 48, 32, 16 and 8 kHz sampling rate which results in respective audio bandwidth restrictions with cutoff frequencies of approx. 22, 15, 7.5 and 3.5 kHz [16].

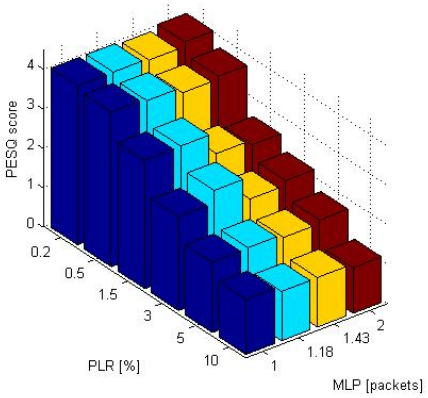
Further, results from EBU subjective listening tests [2], obtained with the MUSHRA method (Multi Stimulus Test with Hidden Reference and Anchors), were employed. These EBU QoE evaluations were performed for low bitrate codec evaluation. Here, one property of the MUSHRA method was especially



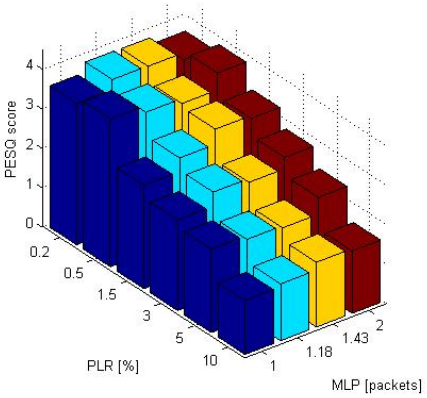
(a) Eapt-X 256 kbit/s $L_c = 768$.



(b) Eapt-X 256 kbit/s $L_c = 256$.



(c) Eapt-X 64 kbit/s $L_c = 128$.



(d) G.722 64 kbit/s $L_c = 160$.

Figure 3.10: PESQ scores for loss impact analysis for different AoIP coding algorithms (2 of 2).

interesting: MUSHRA uses a fullband hidden reference and hidden anchors which are low-pass filtered versions of the reference. In [2], two anchors are used, limited to 3.5 kHz and 7kHz, respectively. The overall quality gradings for the hidden reference and anchors are also given beside the codec results. It was assumed, that this results can be used for the comparison in our work, even if the Continuous Quality Scale ($CQS \in [0, 100]$) of the MUSHRA method is not directly relatable to the MOS scale.

The averaged results of the EBU listening tests are noted in table 4.1, compared with respective I_{bw} results and the ODG scores obtained with the PEMO-Q algorithm. In table 4.2, the relative advantages of wideband to narrowband and ultra-wideband respective fullband to wideband are depicted.

In general, all methods give results with the same trend. Wideband transmission has a high advantage over narrowband while for ultra-wideband compared to wideband a similar but slightly higher advantage can be read out of the table, but it is necessary to assume a non-linear dependence of small CQS results from the MUSHRA method with respective results of the other measures, which is reasonable. For the fullband advantage with respect to ultra-wideband, only a small improvement results for all methods. This is meaningful, even more because the subjective impression of consulted quality experts confirmed the result. This is evident because also from the theory the results can be approved. Findings from the psychoacoustics state that the sensibility of the human ear is getting lower for higher frequencies and is decreasing exponentially for audio frequencies above 10 kHz. The so called hearing area for audio signals has its limits for most subjects between 16 and 18 kHz and the limit decreases with age [97].

[1] Result from non-linear curve fitting between ODG values and MUSHRA CQS results.

Bandwidth	F_s [kHz]	I_{bw}	ODG	MUSHRA CQS
Narrowband	8	35.4	-2.15	27
Wideband	16	6.7	-1.19	52
Ultra-WB	32	-25.47	-0.10	approx. 95 [1]
Fullband	48	-27.89	0	100

Table 4.1: Different bandwidth impairment measures.

Rel. advantage in B	$\Delta MOS\ LQS$	ΔODG	ΔI_{bw}	ΔCQS
WB to NB	ca. 1.4	1	28.7	25
Ultra-WB to WB		1.1	32.2	≈ 43
FB to WB		1.2	34.6	48

Table 4.2: Relative bandwidth impairments.

Finally, we assume an applicability of the wideband bandwidth impairment model extrapolation based on the assumption of a linear dependence between the I_{bw} and the PEMO-Q ODG values (squared 2-norm of the residual: 0.22). This enables that we can use the difference in I_{bw} between the best possible wideband transmission (audio bandwidth 200-7000 Hz [64] and linear PCM coding) with $I_{bw} = 0$ and the result for undistorted fullband transmission (linear PCM with $F_s = 48$ kHz) for extending the wideband R-factor scale to fullband using $\Delta I_{bw} \approx 28$. This results in our proposal for a fullband basic transmission rating factor as

$$R_{0,FB} = R_{0,WB} + \Delta I_{bw} = 129 + 28 = 157\ . \tag{4.4}$$

4.2 A Delay Impairment Factor for ACIP

Now, the formulation of the delay impairment factor for ACIP is addressed. Usually, for formulating a delay impairment factor I_d for a communication service's use case, huge subjective conversational tests are necessary (such as compared in [20] for telephony) which even need more effort than subjective listening-only tests. Furthermore, the "perception of delay strongly depends on the conversational situation" [20], because conversation situations are strongly influenced by the degree of interaction between the participants [20]. Hence, a delay perception model must be geared to the use case it is dedicated for. While ACIP has various use cases with different expectations on the delay (resulting in different delay models), the focus in the following is on the applications requiring high interactivity, e.g. "live discussions". For a delay impairment factor dedicated to this ACIP application, a simple but promising approach was chosen to enable an incorporation of the delay impairment on the perceived QoS in the QoE prediction framework. Nevertheless, it can be assumed that it will be not as accurate as one based on conversational tests.

In the operational requirements for different audio contribution use cases, presented in table 1.1, a maximum one-way delay of 100 ms is specified for interviews or discussions. This was used as anchor for the delay model derivation. Different approaches for I_d are available, we finally decided to use the AT&T simplified model [84], because it is relatively simple, has a linear increase even for higher delays which are even less tolerated for interactive ACIP and most important: it has a precise turning point where the slope increases, easy realized using the step function. For the dedicated ACIP approach, the turning point of the AT&T simplified model was shifted to the left to introduce the stricter delay requirements. The

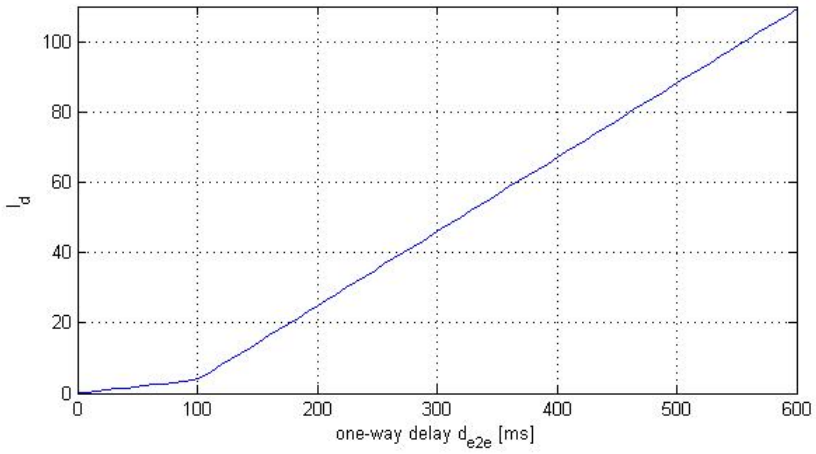


Figure 4.1: Model for $I_{d,\mathrm{ACIP}}$ versus one-way delay d_{e2e}.

result was furthermore rescaled from the ordinary narrowband design to the fullband framework. In fig. 4.1 the resulting characteristic can be observed. The underlying delay impairment factor formula is

$$I_d = 0.024d + 0.11(d_{\mathrm{e2e}} - d_0) \cdot H(d_{\mathrm{e2e}} - d_0)\,, \tag{4.5}$$

which is the original AT&T simplified model but with a turning point at $d_0 = 100$ ms instead of using $d_0 = 177.3$ ms. $H(x)$ is the step function. The ACIP delay impairment factor results after the linear conversation to fullband as $I_{d,\mathrm{ACIP}} = 1.57\ I_d$.

This solution could then further be used in the conversational quality estimation model while its suitability should be further evaluated with subjective tests comparison which was not possible in the context of the present work because neither such results were available nor the time for performing such tests remained.

4.3 Fullband Equipment Impairment Factors

Because in this work a new fullband quality rating model based on the E-model was derived, until now no provisional planning values as specified by the ITU-T for the narrowband and wideband cases exist (see chapter 2.2.3), also an methodology for the derivation of equipment impairment factors from instrumental models for the fullband case is missing so far. Fortunately, the principle approach in the ITU-T rec. P.834.1 [43] (for wideband) can be modified for the fullband case, because the whole approach here is based on the E-model concept. Therefore the challenge was to find an equipment impairment factor derivation methodology for the fullband case based on the assumption that the WB-PESQ algorithm can also be used for the fullband transmission quality rating if results from the PEAQ algorithm (ITU-R rec. BS.1387-1) [5] are considered as well (see below). The goal is to find a reasonable mapping $I_{e,eff,FB} \Longleftrightarrow MOS$. This is addressed in section 4.3.2.

4.3.1 Objective QoE Evaluations

Primarily, for the design of the desired equipment impairment model, a basis of QoE rating results related to quantitative model parameters must exist, such as from the audio stream and from particular network properties. Therefore, normally, subjective tests are performed. Instead we decided to focus on objective

evaluation because subjective auditory testing is expensive and slow, which makes it unsuitable for day-to-day quality evaluations which we required. However, we cross-checked the suitability of our model and results by individual expert testing.

We used the WB-PESQ as well as the PEAQ algorithm for objectively quantifying the user perceived audio quality for different equipment impairments due to the usage of VoIP technology. We combined the method's scores to a new QoE metric, based on a linear combination of the used ojective methods, presented in [17]. Thereby, we exploited the ability of the PEAQ algorithm to describe small impairments on fullband audio signals, while the WB-PESQ algorithm more ideally rates stronger impairments and accounts for possible impairments due to packet loss. Furthermore, some evaluations in [66] showed, that it can also be used for the evaluation of wideband mono audio signals. This trade-off approach was chosen because so far no unified fullband method for objective audio quality evaluation is available, which would regard also impairments possibly introduced by IP-based transmissions. Finally, we calculated WB-PESQ scores with respect to coding and loss impairments while we took necessary results of the PEAQ algorithm with respect to coding impairments from already existent evaluations[2]. For the loss evaluation, a loss process characterization based on a 2-state Markov process [72] has been derived, which allowed for taking bursty loss into account. We performed our experiments in a dedicated application-oriented experimental environment for the coding algorithm configurations specified in table 4.3. Our experimental methodology is illustrated in depth in [16], where we also present results with respect to coding and packet loss, furthermore considering different IP packet length. In fig. **??**, the PEAQ modified PESQ results of our QoE metric in comparison to the raw WB-PESQ scores are demonstrated.

4.3.2 Methodology for $I_{e,FB}$ Derivation

The focus in the following is on clean transmission (i.e. no packet loss), while the $I_{e,FB}$ derivation methodology was also used for computing impairment factor anchors for results with respect to packet loss ratios $\rho \neq 0$ and mean loss periods $\mu \geq 1$ to later interpolate the discrete results for different loss ratios to an comprehensive $I_{e,eff,FB}$ using a non-linear least-squares curve fitting strategy [16].

In principle, the procedure for the derivation of equipment impairment factors is as follows, similar to the narrowband and wideband case (see section 2.2.3): the derived *MOS LQO* values based on the WB-PESQ and PEAQ evaluations are mapped to a *MOS LQS* which is then transformed on the fullband *R*-factor scale. From the obtained R-factor estimates $\hat{R}_{FB}$ a raw equipment impairment factor K is calculated with $\hat{R}_{FB} = R_{0,FB} - K$. K is finally mapped to the desired $I_{e,FB}$ value with an adjusting function derived with reference conditions which ensures that the fullband results are consistent with the results of the

[2] These PEAQ results are recorded in a measurement report from M. Karle (Hessischer Rundfunk, 2006)

Coding Algorithm	Channels	r_c [kbit/s]	F_s [kHz]
MPEG Layer2 (1)	2	384	48
MPEG Layer2 (2)	2	256	48
Eapt-X (1)	2	384	48
Eapt-X (2)	2	256	32
Eapt-X (3)	1	64	16
ITU-T G.722 / G.711	1	64	16/8

Table 4.3: Selected ACIP coding algorithms.

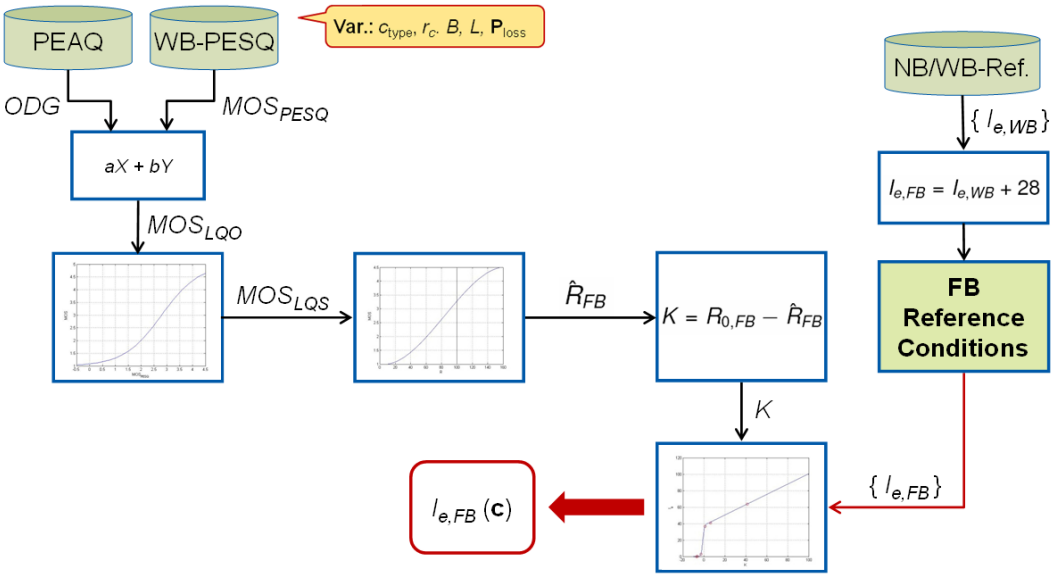


Figure 4.2: Methodology for the derivation of I_e, FB.

Codec name	Channels	F_s [kHz]	r_c [kbit/s]	$I_{e,FB}$
16 bit lin. PCM	2	48	1536	0
MPEG L2	2	48	384	0.2
Eapt-X	2	32	256	6.5
Eapt-X	1	16	64	36.7
ITU-T G.722	1	16	64	41
ITU-T G.711	1	8	64	63.8

Table 4.4: $I_{e,FB}$ reference conditions.

E-model for narrowband and wideband transmission. The overall procedure is described in more detail in [16] but depicted in fig. 4.2. The procedure can be in short form formulated as

$$MOS_{LQO} \Longrightarrow MOS_{LQS} \Longrightarrow \hat{R}_{FB} \Longrightarrow K \Longrightarrow I_{e,FB} \ . \tag{4.6}$$

In the following, we describe the mapping of the *MOS LQS* to the R_{FB} as well as the transformation of the K value to a stable $I_{e,FB}$.

For the mapping of the *MOS LQS* to the R_{FB}, the calculated *MOS* estimations based on the objective evaluations (see the previous section 4.3.1) are first transformed to the non-extended R_{NB}-scale (range $[0, 100]$), using the relationship of *MOS* and *R*-factor from the narrowband E-model, which still reflects the narrowband use of the *MOS* scale assumed by the original E-model [43]. For obtaining results which reflect the superior quality of fullband transmission to the narrowband and wideband cases, the *R*-factor scale is linearly extended to the fullband case with $R_{FB} = 1.57\, R_{NB}$.

In the last step, the obtained K value has to be transformed to a stable $I_{e,FB}$ value. Hence, we had to design a transformation function. This K to $I_{e,FB}$ mapping was found by using reference conditions, which firstly had to be defined because until now, no reference conditions for $I_{e,FB}$ are available. Hence, we derived 6 reference conditions: we derived 3 of them throughout the methodology design process for audio coding algorithms to ensure a high correlation of the results with the underlying QoE results of the WB-PESQ and PEAQ algorithm as well as subjective impressions. The other 3 are the undistorted linear fullband case, the narrowband speech codec ITU-T G.711 and the wideband codec ITU-T G.722.

For all these coding algorithms we computed raw K values following the steps of the methodology described before. Then, the reference conditions for ITU-T G.711 and ITU-T G.722 were obtained by shifting the wideband values from the provisional planning values in ITU-T rec. G.113 [42] to fullband with $I_{e,FB} = I_{e,WB} + 28$ following the wideband principle, while for the undistorted linear fullband case we defined $I_{e,FB} = 0$. In table 4.4 the 6 reference conditions are listed with their basic configuration and finally defined fullband equipment impairment factors $I_{e,FB}$. The interrelationship of raw K and defined $I_{e,FB}$ values were finally be used to derive the normalization function by section-wise linear and non-linear curve fitting.

4.4 Bandwidth Impairment Factor

We now also incorporate the concept for a seperate bandwidth impairment factor (see section 2.2.3) in our fullband QoE model. For this, We simply shifted I_{bw} for wideband to the fullband R-factor scale,

$$I_{bw,FB} = I_{bw,WB} + \Delta I_{bw} = I_{bw,WB} + 28 \ , \tag{4.7}$$

using the wideband to fullband rating conversion presented in section 4.3.2. Now, the best performance without perceived bandwidth impairments is shifted to full audio bandwidth with $I_{bw,FB} = 0$, while the previous best rating for the bandwidth 200-7000 Hz which resulted in $I_{bw,WB} = 0$, is now rated with $I_{bw,FB} = -28$.

Similar to the in [93] recently presented splited results for different modern speech codecs, the obtained final fullband equipment impairment factors for ACIP algorithms are plotted against their resulting bandwidth and coding-only impairment in fig. 4.3. The reasonable results consolidate the proposed solutions and give a nice picture of the different impairment relations to the total equipment impairment value.

4.5 Simplified Model for MOSc

We were able to use our objective results in order to build simplified two-dimensional parametric models of the *conversational MOS* (*MOSc*) [84]. We obtained the *MOSc* results by transformation of the objective QoE evaluations $MOS_{\text{obj}}(\mathbf{P}_{\text{loss}})$ for a specific coding algorithm and configuration $\mathbf{c}$ as well as packet size L to a fullband equipment impairment factor $I_{e,eff,FB}$ in the R-factor domain enabled by the QoE rating model framework for ACIP described in section **??**. The resulting R_{FB} from eq. (4.1) can then be transformed to an estimate of the $MOSc$, depending continuously on the overall packet loss ratio ρ and the end-to-end delay d_{e2e}, which are the most important QoS parameters besides the available bandwidth a_{bw}. The latter one is not considered directly in the QoE model, but it limits the maximum coding bitrate which can be used, $r_c = f(a_{bw})$.

The principal dependencies of a continuous $MOSc$ model can be described as

$$MOSc = f(\rho, d_{\text{e2e}})|_{\mu=\mu_k, \mathbf{c}=\mathbf{c}_i, L=L_j(\mathbf{c}_i)} \ . \tag{4.8}$$

Here the codec configuration $\mathbf{c}_i$ is taken from the set of coding algorithm configurations $\mathscr{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_i, \ldots \mathbf{c}_I\}$ of size I, while the packet size L_j is taken from the respective set of possible packet length $\mathscr{L} = \{L_1, L_2, \ldots, L_j, \ldots, L_J\}$ of size J. Moreover, for the different discrete mean loss periods μ_k ($k = 1 \ldots 4$) from the set $\mathscr{Q}$ of size $K = 4$ defined in eq. (**??**), dedicated parameter models can be accessed. The $MOSc$ surfaces can finally be least-squares fitted using a general polynomial model [84].

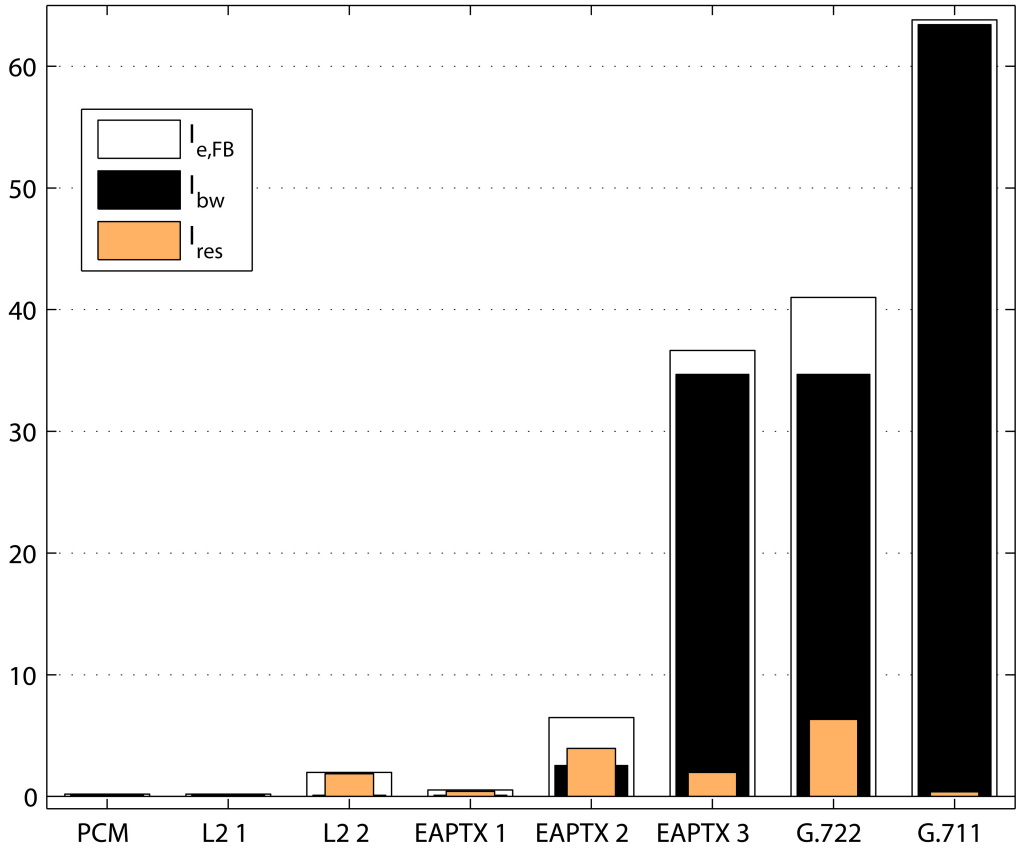


Figure 4.3: $I_{e,FB}$ (white), $I_{bw,FB}$ (black) and I_{res} (beige) for the examined coding algorithms.

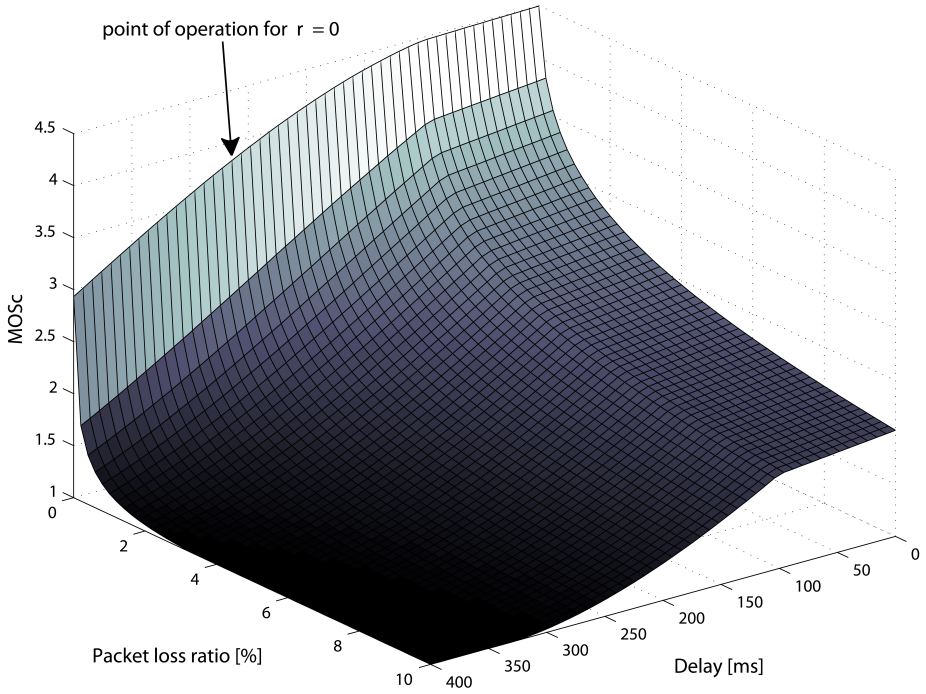


Figure 4.4: $MOSc$ surface for MPEG Layer 2 (1).

Fig. 4.4 and fig. 4.5 depict exemplary $MOSc$ surfaces for $\mu = 1$ of dedicated ACIP algorithms. Fig. 4.4 shows the resulting model for the MPEG Layer 2 coding algorithm with a 384 kbit/s coding bitrate and $F_s = 48$ kHz, while fig. 4.5 shows the one for the Eapt-X coding algorithm with 64 kbit/s and $F_s = 16$ kHz. In both figures a QoE result is marked, which corresponds to a typical non-optimized ACIP operation, assuming a network delay of 70 ms and a playout buffer delay of 50 ms, which represent minimal VoIP contribution values to the overall delay [64]. For the packet loss $\rho = 0$ is assumed. Even if the MPEG algorithm in fig. 4.4 has the desired audio quality, the high coding delay reduces the conversational quality rapidly. If loss were present, we could not recommend its usage with IP communication, even if it was specified by the EBU.

The $MOSc(\rho, d_{\mathrm{e2e}})$ models can be used for estimating the conversational quality in ACIP systems based on non-intrusive passive measurements of the network QoS parameters. This enables a perceptually-driven QoS optimization, e.g. by choosing a perceptually optimal rate control at the sender-side or a perceptually optimal playout buffer size at the receiver-side. For example, the optimum playout buffer size is a trade-off between buffer delay d_b and late loss ρ_b, which is the possible loss due to buffer constraints additional to the network packet loss ρ_n, $\rho = \rho_n + \rho_b$. Sometimes accepting a late loss (ideally concealed thereafter) may be perceptually more meaningful than a higher playout buffer size to cope a wider range of network delay variation Δd_n (jitter), because this would increase the overall end-to-end delay which is desired to be as small as possible to ensure interactivity. In fig. 4.5 such a trade-off is depicted (thick arrow on the surface). Allowing for late losses, the conversational MOS value is increased while the delay is decreased because the playout buffer delay d_b is lowered. Hence, a QoS optimization is achieved.

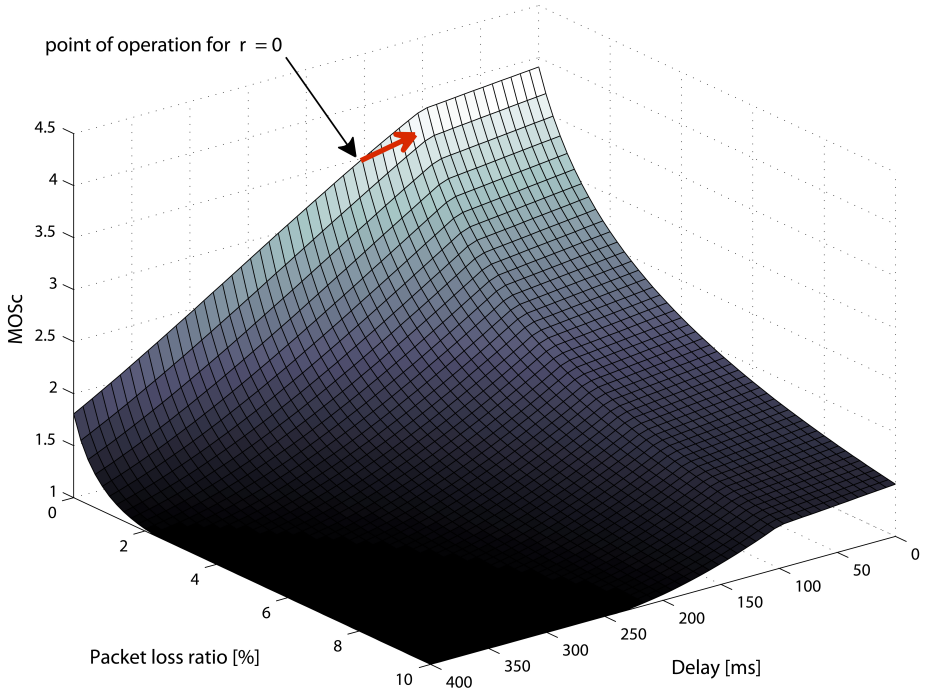


Figure 4.5: *MOSc* surface for Eapt-X (3).

5 Conclusions and Further Work

In this paper, we proposed a non-intrusive parametric QoE assessment model for ACIP quality rating, providing separated impairment factors for the delay, loss, coding and bandwidth impairment. For this, we derived a broadband extension of the wideband R-factor of the extended E-model to enable the conversational quality prediction for ACIP and presented a delay impairment factor for ACIP as well as a derivation methodology for fullband equipment impairment factors from instrumental models, based on our novel listening-only quality evaluation approach for broadband. We also proposed a necessary set of reference conditions for ACIP. Moreover, we derived *MOSc* surfaces and gave an example for their application in perceptually-driven QoS optimization.

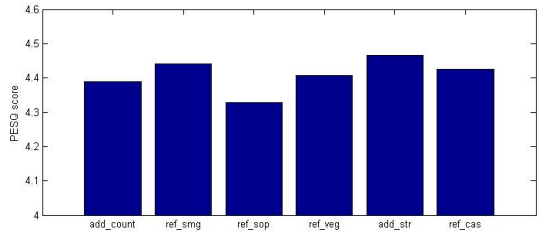
Our findings correlate mostly well with the opinion of our experts, while the validation of the framework with sophisticated subjective testing is desirable. Alternatively, the accuracy of our approach may be investigated in the future with the P.OLQA model (Objective Listening Quality Assessment), which is still in standardization process in the ITU-T. The ITU-T plans to replace PESQ by P.OLQA, which should be able to assess speech quality up to full audio bandwidth signals [69].

For a comprehensive quality rating, the framework should be further extended. Until now, only the packet loss and delay impairments were parametrized, while the model can be enhanced if also parametric models for the QoE dependency on coding bitrate, packet size and loss burstiness are incorporated. Moreover, the advantage of using stereo instead of mono transmission was not considered so far. Also the E-model advantage factor A could be incorporated for describing an access advantage, for example in an extreme case where someone is exclusively at a catastrophe location and has mobile communication access via mobile phone, so that he can give some live comments for a broadcasting station, then A will be high because the audio quality is then not the main issue, it prevails the advantage of actuality combined with the access possibility. Therefore actuality and exclusiveness of transmitted information may be considered in this measure. Furthermore, a greater set of audio coding algorithms potentially

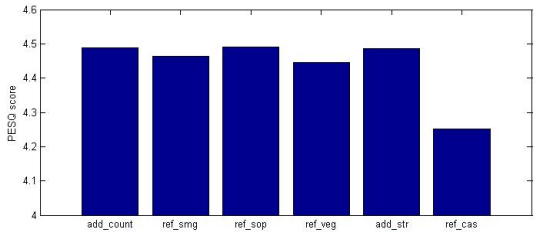
useful for professional audio communication should be evaluated (e.g. AMR-WB+, AAC ELD). Thereby, also their packet loss concealment methods should be examined with respect to the QoE for quantifying their advantage and capabilities.

A Appendix: WB-PESQ Score Results per Testfile

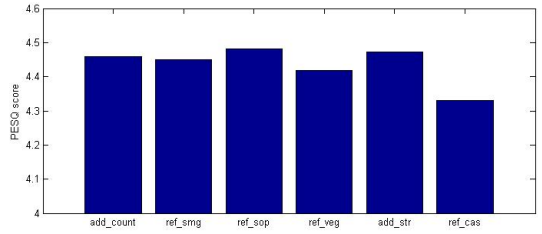
Fig. A.1 shows the WB-PESQ score results per testfile (see table 3.2) for the different audio coding algorithms. The results for MPEG L2 (384 kbit/s 48 kHz) were excluded while it is perceived nearly transparent and the mean WB-PESQ score was 4.48 without significant deviations with respect to the different testfiles.



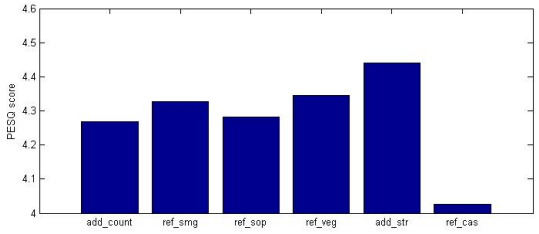
(a) MPEG L2 256 kbit/s 48 kHz.



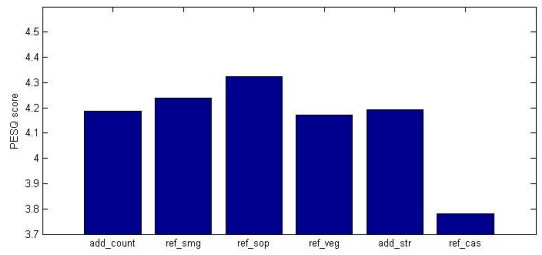
(b) Eapt-X 384 kbit/s 48 kHz.



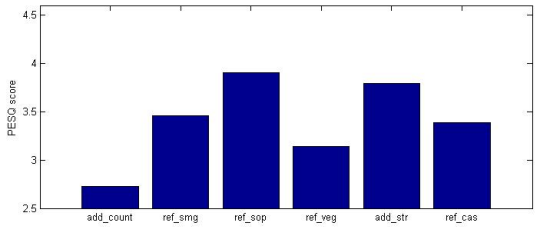
(c) Eapt-X 256 kbit/s 32 kHz.



(d) Eapt-X 64 kbit/s 16 kHz.



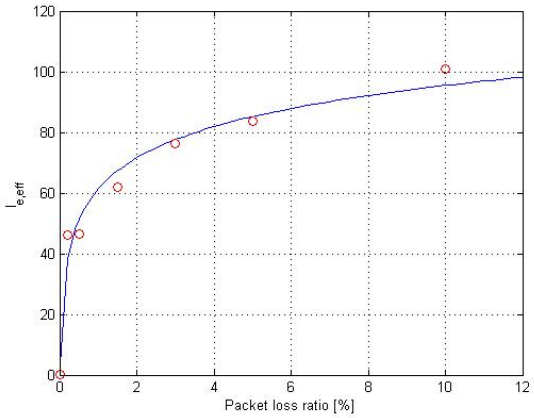
(e) G.722 64 kbit/s 16 kHz.



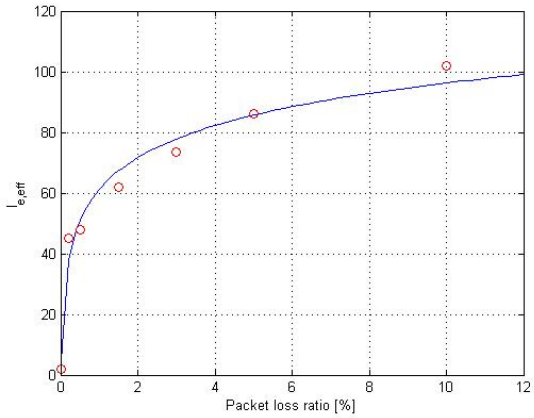
(f) G.711 64 kbit/s 8 kHz.

Figure A.1: WB-PESQ score results per testfile for different audio coding algorithms (with error-free network, $\rho = 0$).

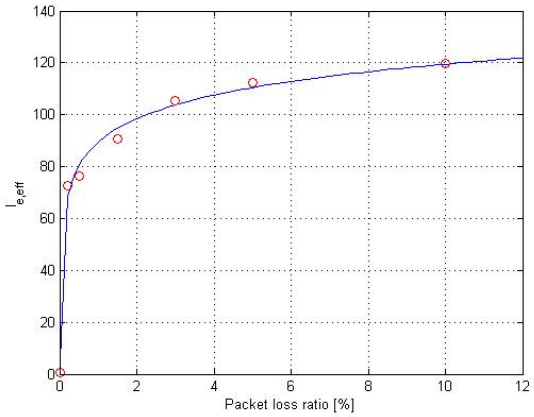
B Appendix: $I_{e,eff,FB}$ Characteristics



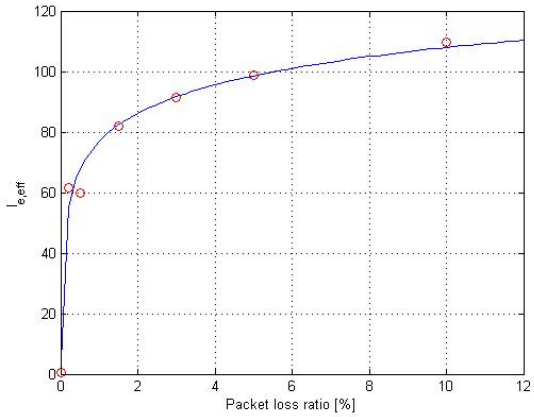
(a) MPEG L2 384 kbit/s $L_c = 1152$.



(b) MPEG L2 256 kbit/s $L_c = 384$.

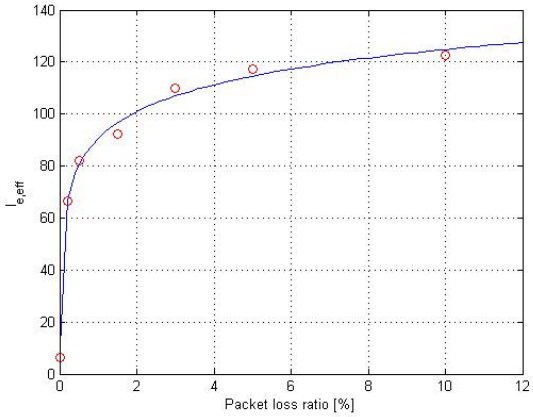


(c) Eapt-X 384 kbit/s $L_c = 1024$.

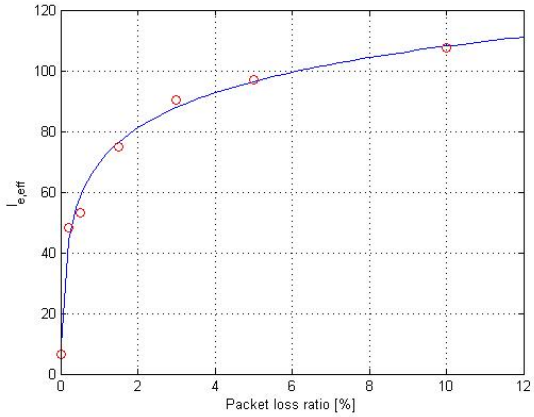


(d) Eapt-X 384 kbit/s $L_c = 256$.

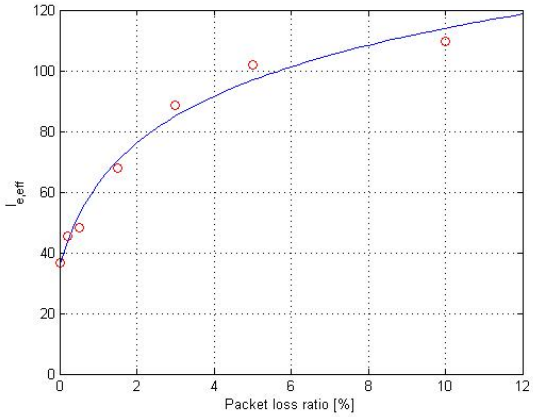
Figure B.1: Resulting $I_{e,eff,FB}$ characteristics in dependence of independent packet loss ($MLP = 1$) for different AoIP coding algorithms (1 of 2).



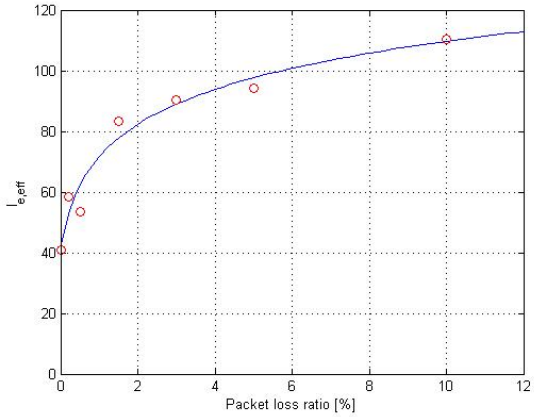
(a) Eapt-X 256 kbit/s $L_c = 768$.



(b) Eapt-X 256 kbit/s $L_c = 256$.



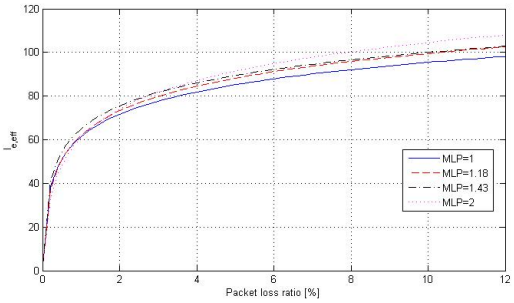
(c) Eapt-X 64 kbit/s $L_c = 128$.



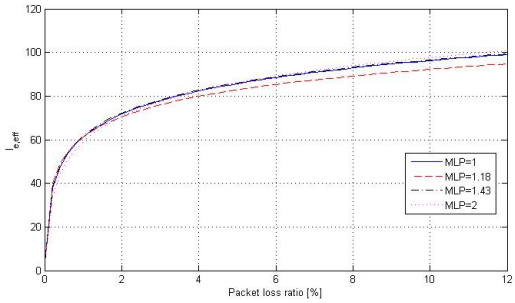
(d) G.722 64 kbit/s $L_c = 160$.

Figure B.2: Resulting $I_{e,eff,FB}$ characteristics in dependence of independent packet loss ($MLP = 1$) for different AoIP coding algorithms (2 of 2).

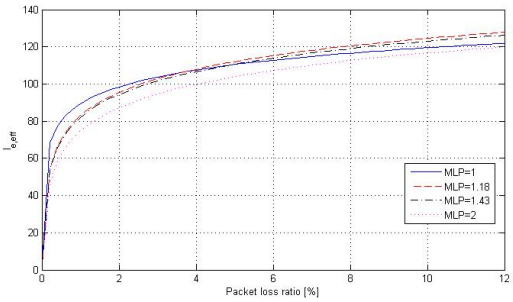
C Appendix: $MOSc$ Surfaces



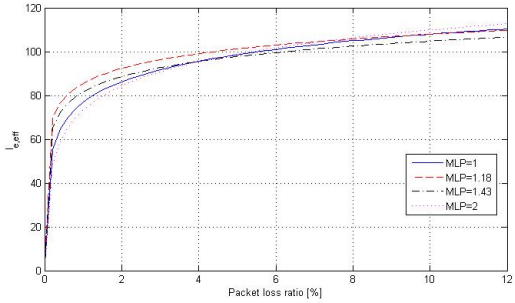
(a) MPEG L2 384 kbit/s $L_c = 1152$.



(b) MPEG L2 256 kbit/s $L_c = 384$.



(c) Eapt-X 384 kbit/s $L_c = 1024$.



(d) Eapt-X 384 kbit/s $L_c = 256$.

Figure B.3: Resulting $I_{e,eff,FB}$ characteristics in dependence of different depenend packet loss processes for different AoIP coding algorithms (1 of 2).

Bibliography

[1] J. Alnatt. Subjective Rating and Apparent Magnitude. In *International Journal of Man-Machine Studies* Johannesson [44], pages 801–816.

[2] EBU Project Group B/AIM. EBU Subjective Listening Tests on Low-Bitrate Audio Codecs. Tech 3296, EBU, Jun. 2003.

[3] S. Bech and N. Zacharov. *Perceptual Audio Evaluation: Theory, Method and Application*. John Wiley Wiley & Sons, Chichester, UK, 2006.

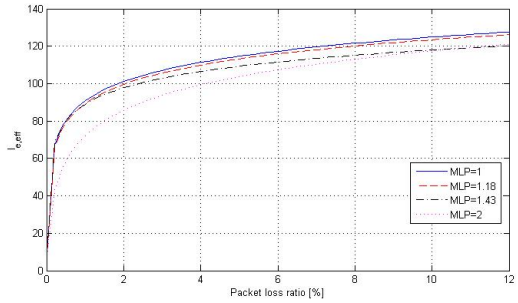
[4] J. Beerends, A. Hekstra, A. Rix, and M. Hollier. Perceptual Evaluation of Speech Quality (PESQ), the New ITU Standard for End-to-End Speech Quality Assessment. Part II: Psychoacoustic Model. *JAES*, 50(10):765–778, Oct. 2002.

[5] D. Campbell, E. Jones, and M. Glavin. Audio Quality Assessment Techniques - A Review, and Recent Developments. *Signal Processing*, 89(8):1489–1500, 2009.

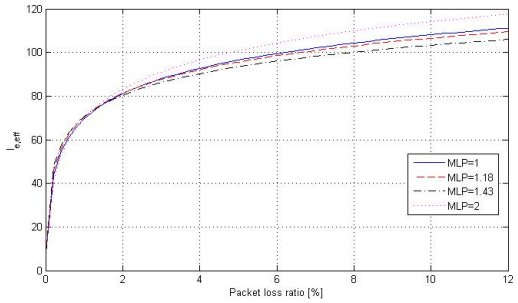
[6] K. Campbell. Deploying Large Scale Audio over IP Networks. In *AES Convention paper 7652, presented at the 126th AES Convention*, May 2009. preprint.

[7] P. Casas, P. Belzarena, and S. Vaton. End-2-End Evaluation of IP Multimedia Services, a User Perceived Quality of Service Approach. In *18th ITC Specialist Seminar Quality of Experience 2008*, Karlskrona, Sweden, May 2008.

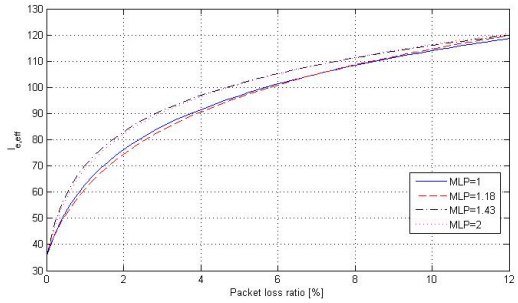
[8] C.J. Chambers. The Development of ATM Network Technology for Live Production Infrastructure. Bbc research white paper whp 074, British Broadcasting Corporation (BBC), Sep. 2003.



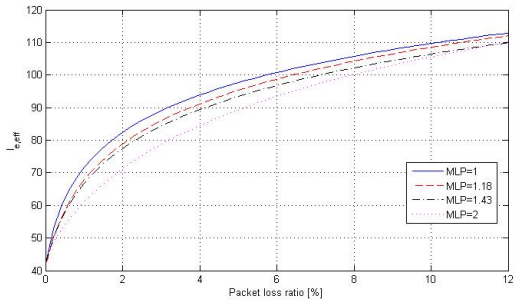
(a) Eapt-X 256 kbit/s $L_c = 768$.



(b) Eapt-X 256 kbit/s $L_c = 256$.



(c) Eapt-X 64 kbit/s $L_c = 128$.



(d) G.722 64 kbit/s $L_c = 160$.

Figure B.4: Resulting $I_{e,eff,FB}$ characteristics in dependence of different depenend packet loss processes for different AoIP coding algorithms (2 of 2).

[9] R. G. Cole and J. H. Rosenbluth. Voice over IP Performance Monitoring. *ACM SIGCOMM Computer Communication Review*, 31(2):9–24, 2001.

[10] S. Daniels. Can the Public Internet be used for Broadcast? In *AES Convention paper 7324, presented at the 124nd AES convention*, May 2008.

[11] S. Daniels. 20 Things you should know before Migrating your Audio Links to IP. In *AES Convention paper 7651, presented at the 126th AES Convention*, May 2009. preprint.

[12] A.A. de Lima, F.P. Freeland, R.A. de Jesus, B.C. Bispo, L.W.P. Biscainho, S.L. Netto, A. Said, A. Kalker, R. Schafer, B. Lee, and M. Jam. On the Quality Assessment of Sound Signals. In *IEEE International Symposium on Circuits and Systems, ISCAS '08*, pages 416–419, May 2008.

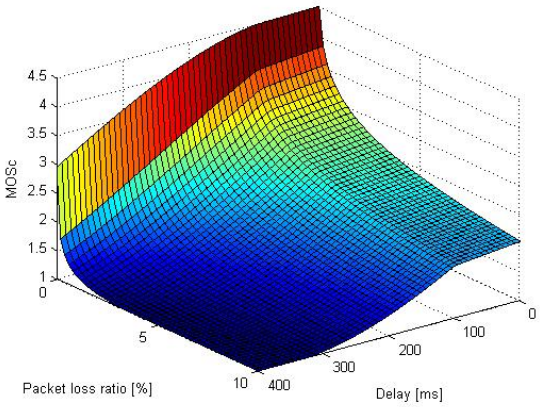
[13] B. Feiten, A. Raake, M.-N. Garcia, U. Wüstenhagen, and J. Kroll. Subjective Quality Evaluation of Audio Streaming Applications on Absolute and Paired Rating Scales. In *AES Convention paper 7787, presented at the 126th AES Convention*, May 2009. preprint.

[14] OPTICOM GmbH. OPERA (TM) Audio Quality Analysis Tool. available: http://www.opticom.de/products/audio-quality-testing.html, 1999.

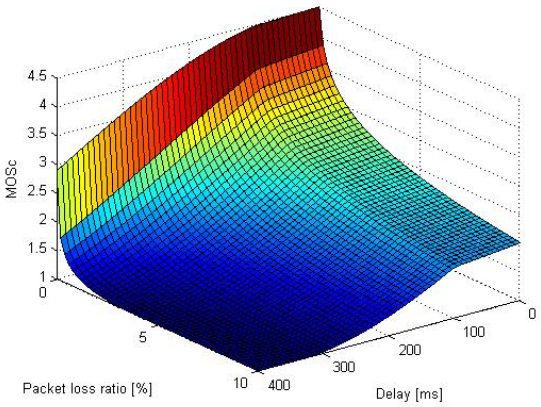
[15] J.-N. Gouyet. Towards Reliable IP Networks for Broadcast Applications. Network technology seminar 2009, EBU Network Technology Management Committee (NMC), Geneva, Jun. 2009.

[16] Maxim Graubner. QoE Assessment and a Perceptually-Driven QoS Optimization Model for Audio Contribution over IP. diploma thesis, TU Darmstadt, Germany, Sep. 2009.

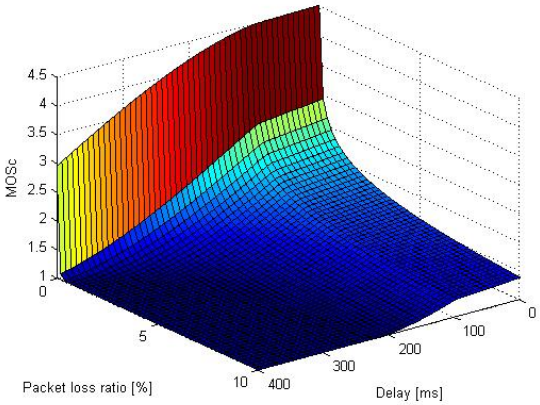
[17] Maxim Graubner, Parag Mogre, Ralf Steinmetz, and Thorsten Lorenzen. A New QoE Model and Evaluation Method for Broadcast Audio Contribution over IP. submited to: *NOSSDAV*, Jun. 2010.



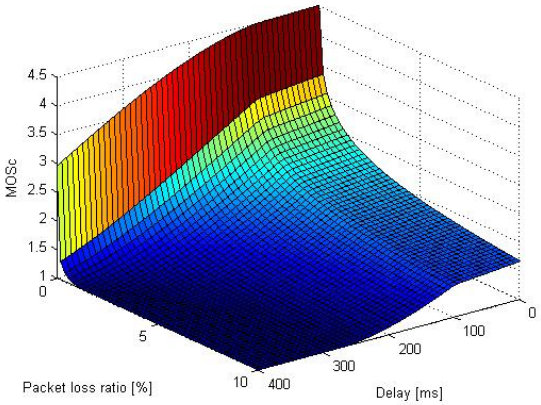
(a) MPEG L2 384 kbit/s $L_c = 1152$.



(b) MPEG L2 256 kbit/s $L_c = 384$.



(c) Eapt-X 384 kbit/s $L_c = 1024$.



(d) Eapt-X 384 kbit/s $L_c = 256$.

Figure C.1: Derived $MOSc$ surfaces in dependence of end-to-end delay and packet loss ratio with $MLP = 1$ for different audio coding algorithms (1 of 2).

[18] C. Herrero. Subjective and Objective Assessment of Sound Quality: Solutions and Applications. In *International Conference on Acoustics and Musical Research, CIARM '05*, pages 1–20, 2005.

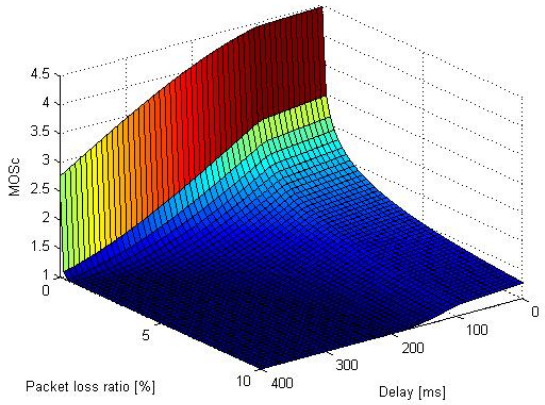
[19] C. Hoene, H. Karl, and A. Wolisz. A Perceptual Quality Model for Adaptive VoIP Applications. In *International Symposium on Performance Evaluation of Computer and Telecommunication Systems, SPECTS '04*, Jul. 2004.

[20] J. Holub, M. Kastner, and O. Tomiska. Delay Effect on Conversational Quality in Telecommunication Networks: Do We Mind? In *Wireless Telecommunications Symposium, WTS '07*, pages 1–4. IEEE, Apr. 2007.

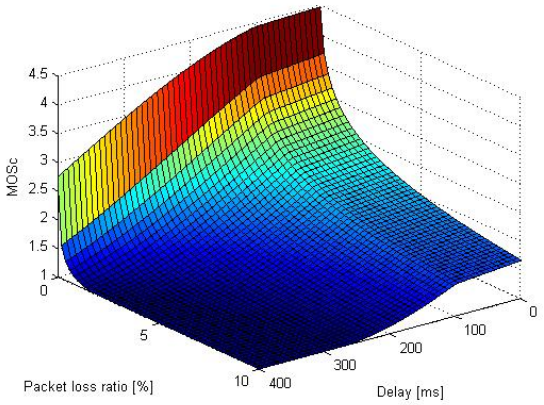
[21] Y. Hu and P.C. Loizou. Evaluation of Objective Quality Measures for Speech Enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):229–238, 2008.

[22] R. Huber. *Objective Assessment of Audio Quality Using an Auditory Processing Model*. PhD thesis, Universität Oldenburg, 2003.

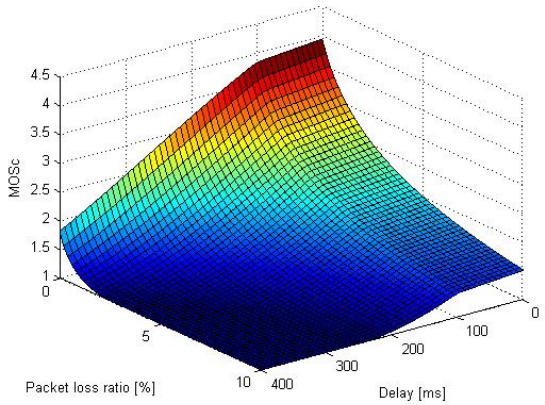
[23] R. Huber and B. Kollmeier. PEMO-Q - A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception. *IEEE Transactions on Audio, Speech, and Language Processing*, 14 (6):1902–1911, 2006.



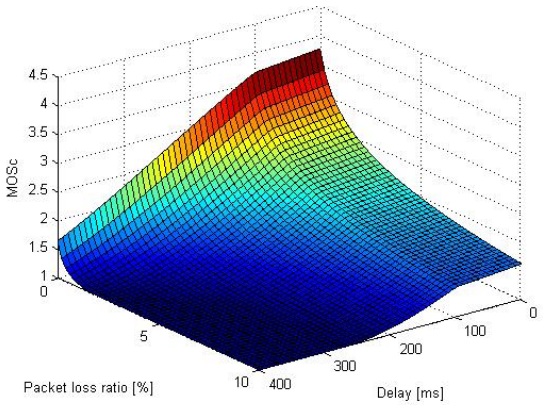
(a) Eapt-X 256 kbit/s $L_c = 768$.



(b) Eapt-X 256 kbit/s $L_c = 256$.



(c) Eapt-X 64 kbit/s $L_c = 128$.



(d) G.722 64 kbit/s $L_c = 160$.

Figure C.2: Derived *MOSc* surfaces in dependence of end-to-end delay and packet loss ratio with $MLP = 1$ for different audio coding algorithms (2 of 2).

[24] J. Issing, N. Färber, and M. Lutzky. Adaptive Playout for VoIP based on the Enhanced Low Delay AAC Audio Codec. In *AES Convention paper 7395, presented at the 124th AES Convention*, May 2008.

[25] ITU-R. Subjective Assessment of Sound Quality. ITU-R recommendation BS.562-3, 1990.

[26] ITU-R. Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems. ITU-R recommendation BS.1116-1, Oct. 1997.

[27] ITU-R. Method for Objective Measurements of Perceived Audio Quality. ITU-R recommendation BS.1387-1, Nov. 2001.

[28] ITU-R. Methodology for the Subjective Assessment of the Quality of Television Pictures. ITU-R recommendation BT.500, 2002.

[29] ITU-R. Method for the Subjective Assessment of Intermediate Quality Levels of Coding Systems. ITU-R recommendation BS.1534-1, Jan. 2003.

[30] ITU-R. General Methods for the Subjective Assessment of Sound Quality. ITU-R Recommendation BS.1284-1, 2003.

[31] ITU-T. Pulse Code Modulation (PCM) of Voice Frequencies. ITU-T recommendation G.711, Nov. 1988. Series G: transmission systems and media, digital systems and networks.

[32] ITU-T. 7 kHz Audio-Coding within 64 kbit/s. ITU-T recommendation G.722, Nov. 1988. Series G: transmission systems and media, digital systems and networks.

[33] ITU-T. Methods for Subjective Determination of Transmission Quality. ITU-T Recommendation P.800, Aug. 1996. Series P: Telephone transmission quality, telephone installations, local line networks.

[34] ITU-T. Definition of Categories of Speech Transmission Quality. ITU-T Recommendation G.109, Sep. 1999. Series G: transmission systems and media, digital systems and networks.

[35] ITU-T. Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs. ITU-T Recommendation P.862, 2001.

[36] ITU-T. Methodology for Derivation of Equipment Impairment Factors from Subjective Listening-only tests. ITU-T Recommendation P.833, Feb. 2001. Series P: Telephone transmission quality, telephone installations, local line networks.

[37] ITU-T. Methodology for the Derivation of Equipment Impairment Factors from Instrumental Models. ITU-T Recommendation P.834, Jul. 2002. Series P: Telephone transmission quality, telephone installations, local line networks.

[38] ITU-T. Mean Opinion Score (MOS) Terminology. ITU-T Recommendation P.800.1, Apr. 2003. Series P: Telephone transmission quality, telephone installations, local line networks.

[39] ITU-T. One-way Transmission Time. ITU-T Recommendation G.114, May 2003. Series G: transmission systems and media, digital systems and networks.

[40] ITU-T. Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs. ITU-T Recommendation P.862.2, Nov. 2005. Series P: Telephone transmission quality, telephone installations, local line networks.

[41] ITU-T. The E-model, a Computational Model for Use in Transmission Planning. ITU-T Recommendation G.107, Mar. 2005. Series G: transmission systems and media, digital systems and networks.

[42] ITU-T. Transmission Impairments due to Speech Processing. ITU-T Recommendation G.113, Nov. 2007. Series G: transmission systems and media, digital systems and networks.

[43] ITU-T. Extension of the Methodology for the Derivation of Equipment Impairment Factors from Instrumental Models for Wideband Speech Codecs. ITU-T Recommendation P.834.1 (prepublished), Apr. 2009. Series P: Telephone transmission quality, telephone installations, local line networks.

[44] N.O. Johannesson. The ETSI Computation Model: a Tool for Transmission Planning of Telephone Networks. *IEEE Communications Magazine*, 35(1):70–79, Jan. 1997.

[45] J. Joskowicz, J.C. López Ardao, and M.A. Gonzáles Ortega. A Mathematical Model for Evaluating the Perceptual Quality of Video. 2009.

[46] P. Kabal. An Examination and Interpretation of ITU-R BS.1387: Perceptual Evaluation of Audio Quality. Tech. report, McGill University, Montreal, Canada, 2002.

[47] M. Karle. PEAQ Evaluations for Different Audio Coding Algorithms. measurement reports, Hessischer Rundfunk IT Systemservice Hörfunk, 2006.

[48] J. Korhonen, Y. Wang, and D. Isherwood. Toward Bandwidth-Efficient and Error-Robust Audio Streaming over Lossy Packet Networks. *Multimedia Systems*, 10(5):402–412, Aug. 2005.

[49] Swedish Radio L. Jonsson and EBU Technical Department M. Coinchon. Streaming Audio Contributions over IP - a New EBU Standard. Ebu technical review Ű 2008 q1, EBU N/ACIP, 2008.

[50] U. Löbbert, G. Amatucci, W. Richter, and N. Roen. Contribution-Workflows Use Cases Man-to-machine Interface. Draft version 1.0, EBU N/ACIP, 2006.

[51] P. C. Loizou. *Speech Enhancement : Theory and Practice*. CRC Press, 2007.

[52] Audio Processing Technology Ltd. The apt-X Suite of Audio Compression Algorithms. White Paper Introduction to apt-XŹ, Feb. 2003.

[53] D. McDysan. *Qos and Traffic Management in IP and ATM Networks*. Osborne/McGraw-Hill, USA, Dec. 1999.

[54] J. McGowan. Burst Ratio: A Measure of Bursty Loss on Packet-based Networks, Aug 2005.

[55] S. Möller. *Assessment and Prediction of Speech Quality in Telecommunications*. Kluwer Academic Publ., Oct. 2000.

[56] S. Möller, A. Raake, N. Kitawaki, A. Takahashi, and M. Waltermann. Impairment Factor Framework for Wide-Band Speech Codecs. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6): 1969–1976, Nov. 2006.

[57] N/ACIP. Audio Contribution over IP - Requirements for Interoperability. Ebu Ű tech 3326rev, EBU, Geneva, 2008. Technical Specification.

[58] N/ACIP. Audio Contribution over IP Tutorial. Ebu Ű tech 3329, EBU, Stockholm, Apr. 2008. tutorial.

[59] D. Pan. A Tutorial on MPEG Audio Compression. *IEEE Multimedia*, 2(2):60–74, 1995.

[60] C. Perkins. *RTP: Audio and Video for the Internet*. Addison-Wesley, 4 edition, 2006.

[61] AQUAVIT project. Assessment of Quality for Audio-Visual Signals over Internet and UMTS - Main Findings. Eurescom project p905 report, European Institute for Research and Strategic Studies in Telecommunications (EURESCOM), Apr. 2001. EDIN 0145-0910.

[62] S. Quackenbush, T. Barnwell, and M. Clements. *Objective Measures of Speech Quality*. Volume 16 of Hu and Loizou [21], 1988.

[63] A. Raake. Predicting Speech Quality under Random Packet Loss: Individual Impairment and Additivity with other Network Impairments. *Acta Acustica united with Acustica*, 90:1061–1083, 2004.

[64] A. Raake. *Speech Quality of VoIP: Assessment and Prediction*. Wiley, Chichester, UK, 2006.

[65] A. Raake. Short- and Long-Term Packet Loss Behavior: Towards Speech Quality Prediction for Arbitrary Loss Distributions. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6): 1957–1968, Nov. 2006.

[66] A. Rix. Perceptual Wideband Speech and Audio Quality Measurement. In *ETSI STQ/ITU Workshop on Wideband Speech Quality in Terminals and Networks: Assessment and Prediction*, Mainz, Germany, Jul. 2004.

[67] A. Rix, M. Hollier, J. Beerends, and A. Hekstra. PESQ-The New ITU Standard for End-to-End Speech Quality Assessment. In *AES Convention paper 5260, presented at the 109th AES Convention*, Sep. 2000.

[68] A. Rix, M. Hollier, A. Hekstra, and J. Beerends. Perceptual Evaluation of Speech Quality (PESQ), the New ITU Standard for End-to-end Speech Quality Assessment. Part I: Time Alignment. *JAES*, 50(10):755–764, Oct. 2002.

[69] A.W. Rix, J.G. Beerends, K. Doh-Suk, P. Kroon, and O. Ghitza. Objective Assessment of Speech and Audio Quality - Technology and Applications. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):1890–1901, Nov. 2006.

[70] L. Roychoudhuri. *Proactive Rate and Error Control for Packet Multimedia Transmissions based on Loss Prediction*. PhD thesis, DePaul University, Chicago, USA, Jan. 2008.

[71] L. Roychoudhuri and E. S. Al-Shaer. *Management of Multimedia Networks and Services*, volume 3271 of *Lecture Notes in Computer Science*, chapter Real-Time Analysis of Delay Variation for Packet Loss Prediction, pages 213–227. Springer Berlin / Heidelberg, 2004 Real-Time Analysis of Delay Variation for Packet Loss Prediction.pdf 2004.

[72] H. Sanneck and G. Carle. A framework Model for Packet Loss Metrics based on Loss Runlengths. In *SPIE/ACM SIGMM Multimedia Computing and Networking Conference*, pages 177–187, 2000.

[73] J.B. Schmitt. *Heterogeneous Network Quality of Service Systems*. Kluwer-Academic-Publ., 1st edition, 2001.

[74] H. Schulzrinne and S. Casner. RTP Profile for Audio and Video Conferences with Minimal Control. Standards Track RFC 3551, Jul. 2003. IETF Network Working Group.

[75] S. Shenker, C. Partridge, and R. Guerin. Specification of Guaranteed Quality of Service. Informational RFC 2212, Sep. 1997. IETF Network Working Group.

[76] M. Siller and J. Woods. Improving Quality of Experience for Multimedia Services by QoS Arbitration on a QoE Framework. In *Packet Video Workshop*, pages 28–29, Apr. 2003.

[77] A. Spanias, T. Painter, and V. Atti. *Audio Signal Processing and Coding*. Wiley-Interscience, 2007.

[78] V. Srivastava and M. Motani. Cross-Layer Design: a Survey and the Road Ahead. *IEEE Communications Magazine*, 43(12):112–119, Dec. 2005.

[79] R. Steinmetz. *Multimedia-Technologie: Grundlagen, Komponenten und Systeme*. Springer-Verlag, Berlin, 3rd edition, 2000.

[80] R. Steinmetz and K. Nahrstedt. *Multimedia Systems*. Springer-Verlag, Berlin, 1st edition, 2004.

[81] P. A. Stevens and M. Zemack. Standardising Audio Contribution over IP Communications. Bbc research white paper whp 170, British Broadcasting Corporation (BBC), Oct. 2008.

[82] G. Stoll and F. Kozamernik. EBU Listening Tests on Internet Audio Codecs. Ebu technical review, EBU Project Group B/AIM, Jun. 2000.

[83] L. Sun and E. Ifeachor. New Methods for Voice Quality Evaluation for IP Networks. In *18th International Teletraffic Congress, ITCI 18*, Berlin, Germany, Sep. 2003.

[84] L. Sun and E.C. Ifeachor. Voice Quality Prediction Models and their Application in VoIP Networks. *IEEE Transactions on Multimedia*, 8(4):809–820, Aug. 2006.

[85] A. Takahashi, H. Yoshino, and N. Kitawaki. Perceptual QoS Assessment Technologies for VoIP. *IEEE Communications Magazine*, 42(7):28–34, Jul. 2004.

[86] A. Takahashi, D. Hands, and V. Barriac. Standardization Activities in the ITU for a QoE Assessment of IPTV. *IEEE Communications Magazine*, 46(2):78–84, Feb. 2008.

[87] P. Tesch. Switching to the Packet - an Approach for Changing to Audio Transport over IP. Master's thesis, University of Liverpool, Nov. 2007.

[88] Inc. The MathWorks. MATLAB. available at: http://www.mathworks.de.

[89] E.P.J. Tozer, editor. *Broadcast Engineer's Reference Book*. Elsevier Oxford, 2004.

[90] U. Trick and F. Weber. *SIP, TCP/IP und Telekommunikationsnetze*. Oldenbourg, 3rd edition, May 2007.

[91] J.E. Voldhaug, E. Hellerud, and P. Svensson. Evaluation of Packet Loss Distortion in Audio Signals. In *AES Convention paper 6855, presented at the 120th AES Convention*, May 2006.

[92] K. Vos, S. Jensen, and K. Soerensen. SILK Speech Codec. IETF Standards Track Internet-Draft, Jul. 2009. URL https://developer.skype.com/silk. Skype Technologies S.A.

[93] M. Waltermann and A. Raake. Towards a New E-Model Impairment Factor for Linear Distortion of Narrowband and Wideband Speech Transmission. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '08*, pages 4817–4820, Apr. 2008.

[94] XiPeng Xiao. *Technical, Commercial and Regulatory Challenges of QoS: An Internet Service Model Perspective*. Morgan Kaufmann, 2008.

[95] U. Zölzer. *Digital Audio Signal Processing*. Wiley & Sons, Chichester, 2nd edition, Aug. 2008.

[96] ZTI. NetDisturb Enhanced Edition version 4.7. available at: http://www.zti-telecom.com/EN/NetDisturb.html.

[97] E. Zwicker and H. Fastl. *Psychoacoustics: Facts and Models*. In Waltermann and Raake [93], Apr. 1999.