Qual User Exp (2017) 2:9
https://doi.org/10.1007/s41233-017-0012-7

REVIEW ARTICLE

Review of recent standardization activities in speech quality of experience

Sebastian Möller[1] · **Friedemann Köster**[1]

Received: 8 June 2017 / Published online: 4 September 2017

Abstract Speech communication services have been amongst the first telecommunication services to be used by a wide public, and the quality experienced by their users has been an object of concern since then. Methods on how to evaluate quality using test participants or using technical measurements and algorithms have been standardized mostly in Study Group 12 of the International Telecommunication Union (ITU-T SG12) and the Technical Committee Speech and multimedia Transmission Quality (STQ) of the European Telecommunications Standards Institute, ETSI. This paper reviews new and updated ITU-T Recommendations and ETSI documents which have emerged within the last 12 years, and puts them into the general framework of available standards for this type of service. It also discusses current work items of ITU-T SG12 to illustrate directions of thoughts and future Recommendations to be addressed within the next study period.

Keywords Quality of experience (QoE) · Speech communication service · Standardization · Subjective evaluation · Quality prediction · International Telecommunication Union (TU) · European Telecommunications Standards Institute (ETSI)

✉ Sebastian Möller
sebastian.moeller@tu-berlin.de

Friedemann Köster
friedemann.koester@tu-berlin.de

[1] Quality and Usability Lab, TU Berlin, Ernst-Reuter-Platz 7, 10587 Berlin, Germany

Introduction

A paradigm shift has been reached during the past three decades. Whereas until the 1980s, telecommunication service providers mostly tried to optimize the performance of individual technical characteristics, a more wholistic view has gained ground since then. What is considered more important than the optimization of individual technical characteristics (such as attenuation, noise levels, echo compensation and delay, non-linear distortions, etc.) is the optimization of the quality experienced by the end user–taking into account the totality of technical service characteristics, and translating them into an experience of a prototypical user. This paradigm shift is reflected by the transition from the term Quality of Service (QoS), i.e. the "[t]otality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service" [1], to the term Quality of Experience (QoE).

The necessity to measure and optimize quality resulted in a framework of recommended—or standardized—methods related to performance, QoS and QoE. The definition of the related concepts themselves, in particular QoS and QoE, has led to considerable activities in the international standardization bodies. The body which carries the terms QoS and QoE explicitly under its mandate is Study Group 12 of the Telecommunication Sector of the International Telecommunication Union, ITU-T SG12. This body has recently updated its definition of QoE in Amendment 5 to P.10/G.100 [2] as follows: "Quality of experience (QoE) is the degree of delight or annoyance of the user of an application or service". [3]. This definition replaces the former 2007 definition at the same place: "The overall acceptability of an application or service, as per-

ceived subjectively by the end-user".[1] The new definition results from discussions with experts from the European Network on Quality of Experience in Multimedia Systems and Services,[2] see [3], and with participants of the Dagstuhl seminar series[3] where a similar definition was developed.

The definitory underpinning was an important progress reached during the past years, but it was by far not the only one. The very nature of QoE, namely to be "the degree of delight or annoyance of a user", requires to put the user and their experiences into the center of investigation if one wants to measure and optimize QoE. This makes subjective methods, i.e. methods which rely on human test participants as perceiving, judging and coding organs, indispensable. Such methods are usually the starting point when a new aspect of a service, or a new type of service, is addressed. Furthermore, service providers are usually not only interested in finding out whether their service is experienced positively, they also would like to know which service elements—in terms of technical characteristics and parameters—make it generating positive, or not-so-positive, experiences in their users. Thus, they would like to obtain in a second step links between subjective experiences and technical parameters, i.e. between QoE and QoS, in order to optimize their services. This optimization was previously mostly performed in a one-to-one manner, i.e. the impact of one characteristic or parameter on perceived QoE was measured, leaving the other technical characteristics (and parameters) at predefined, default settings.

With the increasing complexity of services and underlying systems, as well as with the distribution of responsibilities between different players serving one particular service (e.g. in case of over-the-top services, leased lines, etc.), this one-to-one mapping was no longer meaningful. Instead, service providers needed to have a picture of the *joint* effects of a number of system characteristics on QoE. This was reached by developing prediction models[4] estimating QoE on the basis of signals, parameters, or protocol information. Unfortunately, the development of such instrumental models sometimes led to a loss of information on *which* technical characteristic caused suboptimal QoE, as only estimations of integral QoE of the entire service were provided. This led to the necessity to develop more *diagnostic* models, as we will see in the following.

In this paper, we would like to give a review of standards for the subjective and instrumental assessment of QoE of speech services. The focus will be on speech communication services, as these are the most common speech services used nowadays, but we will also include services which make use of text-to-speech synthesis, or of spoken dialogue systems including speech recognition and interpretation, dialogue management, response generation, and speech output (such as voice portals). The corresponding standards or recommended methods are commonly to be found in the P- and partially also in the G-Series of Recommendations of the Telecommunication Standardization Sector of the International Telecommunication Union, ITU-T, more precisely in the ITU-T P.8X, P.8XX, P.13XX and G.1XX series of Recommendations. Some useful information is also contained in the Standards, Guides, Technical Specifications and Technical Reports issued by the European Telecommunications Standards Institute, ETSI, mostly prepared by its Technical Committee Speech and multimedia Transmission Quality (STQ), as well as in the P.14XX series of ITU-T Recommendations; we will make reference to these documents where appropriate. We deliberately left out standards which refer to methods for pure technical performance measurement, such as the determination of loudness ratings in ITU-T Rec.s P.76-79, the use of objective measurement apparatus and test signals (ITU-T Rec. P.5X and P.5XX series), etc. We also left out recommendations that are rather directed to audio-visual services (ITU-T Rec. P.9XX series), although the borderline between speech-only and audio-visual services involving speech is not always sharp (especially in the P.13XX series of Recommendations). Historically, there is a clear separation between speech services and audio services (such as broadcasting), as the latter were expected to provide a wide audio bandwidth, leading to much higher quality and fidelity of the audio signals. Arguably, this borderline is about to fall, but in standardization, the territories are still separated, with the Radiocommunication Sector of the ITU (ITU-R) dealing with the latter, and ITU-T dealing with the former. Thus, we will also leave audio broadcasting services as a topic for another review.

The paper is structured as follows: in the following section we will review the Recommendations which were

[1] The definition also includes two notes: "NOTE 1—Quality of Experience includes the complete end-to-end system effects (client, terminal, network, services infrastructure, etc.). NOTE 2—Overall acceptability may be influenced by user expectations and context".

[2] COST Action IC 1003 "Qualinet", see http://www.qualinet.eu.

[3] Dagstuhl Seminars 09192 "From Quality of Service to Quality of Experience" (2009), 12181 "Quality of Experience: From User Perception to Instrumental Metrics" (2012), and 15022 "Quality of Experience: From Assessment to Application" (2015), see http://www.dagstuhl.de.

[4] These models are sometimes called "objective models" in order to distinguish them from subjective methods. This dichotomy does however not indicate that the "objective model" would be independent of subjective influence—in fact all "objective models" have been optimized to best estimate the results of subjective experiments. Thus, in the following we rather use the term "instrumental model" instead of "objective model", as the input to the models are instrumental measurements of signals or parameters, rather than subjective opinions.

available in the year 2005, which we consider to be the state-of-the-art for our paper. We will then discuss the considerable advances which have been reached since then, separately for subjective evaluation methods (Sect. "Subjective evaluation methods") and for the instrumental quality prediction methods (Sect. "Instrumental quality prediction methods"). Finally, we will address new emerging paradigms which so far have not resulted in new recommended standards, but which are expected to do so in the near future. We conclude with a summarizing discussion and topics of future work in last section.

State-of-the-art

Rather than going for historical preciseness and completeness, we will describe the state-of-the-art by reviewing a number of Recommendations which were (more or less) frequently used around the year 2005, and which focus on the subjective and/or instrumental assessment of speech quality. Some of these Recommendations have a long-standing tradition (such as Rec. P.800, formerly P.80 and P.74) and have frequently been updated throughout the years, others have been one-shot Recommendations which have not seen many changes. We briefly review the relevant content of each Recommendation, by ordering them in their logical order, and in groups of Recommendations dealing with a similar topic. The precise content of each Recommendation can be found in the referenced documents, and all of them are available free-of-charge under http://www.itu.int.

The following documents contain *general information on subjective test procedures*:

- *ITU-T Handbook on Telephonometry* [4]: Whereas this is not a formal ITU-T Recommendation, and its focus is on telephonometric measurements rather than on QoE, the handbook contains a wealth of information on how to carry out subjective evaluations of speech communication services in a passive (listening-only) or interactive (conversational) way. This includes a discussion of the test procedure and planning, the test rooms, the instructions given to test participants, the test scenarios, questionnaires and ratings scales, as well as a short section on the analysis and interpretation of the results. As instrumental models were not yet commonly available when the handbook was written (in parts in the 1970-80s), these are not handled in the book.
- *ITU-T Rec. P.800: methods for subjective determination of transmission quality* [5]: This Recommendation, formerly numbered P.80 and P.74, is the central point of all Recommendations dealing with subjective speech quality evaluation in ITU-T. Interestingly, it has not been updated since 1996. It contains a short general overview of listening-only and conversational tests (including references to field-test principles used at that time) in its main body, and then provides more detailed information in (normative) annexes. For conversation opinion tests, it describes test room and noise conditions, test participants and instructions, the standard Absolute Category Rating (ACR) scale, and the Difficulty Scale, leading to the percentage of listeners experiencing difficulty in the conversation. On the listening-only side, it describes ACR tests with speech material recording and playback, test procedure, classical rating scales such as the listening-quality scale, the listening-effort scale, and the loudness preference scale, and gives some hints to the statistical analysis. It also describes the Quantal-Response Detectability Test which is not frequently used, mainly to detect the audibility and annoyance of impairments. Regarding comparative listening-only tests, it describes Degradation Category Rating (DCR) tests (paired-comparison against a high quality reference) and Comparison Category Rating (CCR) tests (paired-comparison without a high-quality reference). It also describes a method for assessing speech quality with the help of a reference degradation, by comparing the speech sample under investigation with speech samples which have been degraded with a scalable impairment, such as signal-correlated noise produced with the help of a Modulated-Noise Reference Unit, MNRU [6].
- *ITU-T Rec. P.800.1: mean Opinion Score (MOS) terminology* [7]: Commonly, results obtained on ACR scales are averaged to produce a Mean Opinion Score, MOS. Whereas the entire principle of averaging results on scales which do not show interval or ratio level may be heavily disputed [8, 9], this procedure is still well-accepted because of its simplicity. Unfortunately, the same (ACR) procedure is used in different types of tests and with different types of stimuli, making an interpretation of results difficult. In order to increase transparency, this Recommendation provides a terminology of MOS values obtained in listening-only vs. talking-only vs. conversational situations, and having been obtained by means of subjective tests, signal-based or parametric instrumental prediction models. The recommendation has been updated three times since then, also distinguishing between purely-narrowband (300–3400 Hz), wideband (50–7000 Hz) and mixed-band transmission systems, electrical and acoustic recordings, and lately also addressing audio-visual test methods.
- *ITU-T Rec. P.880: continuous evaluation of time-varying speech quality* [10]: This recommendation

describes a specific subjective test method to be applied to address time-varying transmission characteristics. Instead of asking of a judgment at the end of a speech sample, or at the end of a conversation, test participants are asked to continuously rate the *instantaneous* quality by means of a slider. Whereas the method is the only recommended one so far for time-varying effects, its applicability has been disputed in the visual domain, mainly because of cognitive overload of the test participants which have to perceive and to rate at the same time [11].

The following five Recommendations focus on the *perceptual effects of specific types of equipment*, either in the network or in the terminal:

- *ITU-T Rec. P.830: subjective performance assessment of telephone-band and wideband digital codecs* [12]: This Recommendation provides technical details on speech recordings, experimental parameters and design, and the test procedure for subjective tests involving narrowband and/or wideband codecs. Importantly, it also contains the frequency characteristics for simulating a somehow "standard" narrowband telephone handset by means of an Intermediate Reference System, IRS.
- *ITU-T Rec. P.831: subjective performance evaluation of network echo cancellers* [13]: For evaluating the effects of imperfect network echo cancellers, four different methods are recommended in Rec. P.831: Conversation tests provide a realistic, but not diagnostic assessment; talking-and-listening tests focus on the initial part of a conversation when the canceller converges to a stable state; and two types of third-party listening tests put the listener in the position of the talker, to observe both sides of a conversation and to be able to provide more diagnostic judgments than it would be possible in a standard conversation test. The third-partly listening test types differ with respect to using a Head And Torso Simulator in the set-up or not.
- *ITU-T Rec. P.832: subjective performance evaluation of hands-free terminals* [14]: Also for hands-free terminals specialized test procedures have been developed. These include conversation tests, specific double-talk tests addressing the double-talk behaviour of the terminal (impaired e.g. by level adjustment or echo cancellation), as well as third-party listening-only tests.
- *ITU-T Rec. P.835: subjective test methodology for evaluating speech communication systems that include noise suppression algorithm* [15]: This method focusses on (imperfect) noise suppression algorithms in the network or in the terminal. The idea is to have a trifold listening test procedure, asking listeners to separately rate the speech quality, the quality of the (residual) noise, and the quality of the entire speech sample. This way, diagnostic information for optimizing the settings of the noise suppression algorithm can be obtained. The results of such tests are the target of instrumental algorithms, see Sect. "Instrumental quality prediction methods".
- *ITU-T Rec. P.840: subjective listening test method for evaluating circuit multiplication equipment* [16]: This Recommendation contains mainly technical details which are important when subjectively testing Digital Circuit Multiplication Equipment, DCME. It describes the recording procedure, the system load simulation, the data processing, as well as the test design and procedure.

The following two recommendations focus on *speech technology* used in the respective services:

- *ITU-T Rec. P.85: a method for subjective performance assessment of the quality of speech voice output devices* [17]: Whereas all documents referenced so far address speech communication services between humans, this is the first of two Recommendations addressing a human's interaction with an automatic system. ITU-T Rec. P.85 focusses only on the output side of such a system, in particular when synthesized speech is used. In order to guide the attention of the listener in a realistic way, a primary information-seeking task is given to the listening test participants, and the quality judgment is just solicited as a secondary task. Two types of questionnaires, addressing different aspects of the speech output, are given for collecting the judgments.
- *ITU-T Rec. P.851: subjective quality evaluation of telephone services based on spoken dialogue systems* [18]: The second recommendation focusses on the behaviour of the entire automatic system, which commonly includes the automatic speech recognition, natural language understanding, dialogue management, response generation, and speech output. For this purpose, interaction tests are recommended in which participants have to carry out pre-defined tasks with the system which are presented in terms of (mostly graphical) scenarios. QoE judgments are then solicited on different questionnaires, including pre-experimental, scenario-specific and post-experimental questionnaires.

Whereas the previously-described documents address subjective evaluation methods, the following recommendations focus on *instrumental quality prediction models*. Two Recommendation series address *predictions based on signals*:

- *ITU-T Rec. P.862: perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs* [19]: This long-standing model is the second recommended model for predicting speech quality obtained in a listening-only situation, after its (superseded) predecessor Perceptual Speech Quality Measure, PSQM (former ITU-T Rec. P.861). It is based on a perceptually-weighted difference between the clean input signal and the degraded output signal, which is averaged over time and transformed to a quality estimation. The model mostly addresses the effects of network impairments, such as coding and linear distortions, noise, and time-varying degradations. It models the results of a listening-only ACR test according to ITU-T Rec. P.800, but on a different scale. Whereas the model has been disputed for some inaccuracies, it is still a recommended standard, despite its successor POLQA which has shown better performance in most of the addressed cases, see Sect. "Instrumental quality prediction methods". The reason may be that it is implemented in many technical solutions which are still in use.
- *ITU-T Rec. P.862.1: mapping function for transforming P.862 raw result scores to MOS-LQO* [20]: This recommendation provides a mapping function from the raw values output by PESQ to MOS values obtained in a test according to ITU-T Rec. P.800.
- *ITU-T Rec. P.862.2: wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs* [21]: This Recommendation describes a small update of the PESQ model to deal with wideband speech signals. Compared to PESQ, it mainly uses a different frequency response for the input signals and a different transformation function. Also for this target application, POLQA described in Sect. "Instrumental quality prediction methods" provides better predictions.
- *ITU-T Rec. P.862.3: application guide for objective quality measurement based on recommendations P.862, P.862.1 and P.862.2* [22]: This document describes the range of transmission conditions and measurement set-ups for which the models according to Rec.s P.862, P.862.1 and P.862.2 can be used reliably.
- *ITU-T Rec. P.563: single-ended method for objective speech quality assessment in narrow-band telephony applications* [23]: Whereas the models described in the P.862 Series of Recommendations make use of the (clean) input and the (degraded) output signal of the transmission channel under investigation, the model described in ITU-T Rec. P.563 only uses the degraded output signal. With the help of an artificial reference reconstitution and some adjustment, the model is able to estimate listening-only quality (as obtained in a P.800 test), but with slightly lower accuracy compared to PESQ. The use case for such a model is in non-intrusive monitoring scenarios, where a clean reference might not be available. As its basis PESQ, it only addresses narrowband transmission scenarios.
- *ETSI Guide EG 201 377: specification and measurement of speech transmission quality* [24, 25]: This guide contains in its Part 1 a basic introduction to intrusive quality prediction models which make use of the input and the output signal of the transmission channel under investigation. It addresses general aspects of pre-processing, psycho-acoustic modelling, and distance calculation. In an informative annex, this part also contains brief introductions to classical models like PESQ and its predecessors, as well as the TOSQA model which is sometimes used for predicting speech transmission quality including the terminals. In its Part 3, it contains an introduction to non-intrusive quality prediction, including a list of parameters which can be determined in a non-intrusive way, as well as basic models which may be used for quality prediction. The Part 3 also contains an informative annex with exemplary models.

The final set of Recommendations addresses the *prediction of speech quality from parameters*. These predictions relate to the conversational situation, and include predictions for sub-optimal sidetone, residual talker and listener echo, as well as the effects of pure delay on the conversation flow (to a limited extent).

- *ETSI Technical Report ETR 250: transmission and Multiplexing (TM); speech communication quality from mouth to ear for 3.1 kHz handset telephony across networks* [26]: This lengthy technical report describes the core algorithm and the idea underlying the so-called E-model, a parametric planning tool for narrowband networks. The model has been developed in a working group of ETSI by merging expertise and experiences gained with models from large network operators during the 1970–80 years. It translates a parametric description of network and terminal elements to so-called impairment factors which are expected to be additive on a so-called "psychological" scale, the transmission rating scale R. On this scale, the respective impairments are expected to be additive, by subtracting their corresponding impairment factors from a maximum $Rmax$ value. The model described in the ETSI report has been at the basis of the standardization activities of ITU-T SG12, but has never been updated itself since 1996.
- *ITU-T Rec. G.107: the E-model: a computational model for use in transmission planning* [27]: This

Recommendation contains the current version of the E-model. Since its first establishment in 1998, it has been continuously updated (also after 2005) to reflect the perceptual effects in a more reliable way. It is also at the basis for the wideband version developed later, see Sect. "Instrumental quality prediction methods".

- *ITU-T Rec. P.833: methodology for derivation of equipment impairment factors from subjective listening-only tests* [28]: One particularly important type of equipment which needs to be considered in the E-model are speech codecs, with and without packet loss degradations. For this purpose, the E-model needs a so-called equipment impairment factor, *Ie*, *eff*. ITU-T Rec. P.833 describes a method for deriving such a factor for a new (unknown) codec on the basis of a properly designed P.800 listening-only test. Tabulated values for the equipment impairment factor for standardized codecs are found in Appendix I of ITU-T Rec. G.113 [29].
- *ITU-T Rec. P.834: methodology for the derivation of equipment impairment factors from instrumental models* [30]: Whereas the P.833 method derives *Ie*, *eff* values from subjective tests, the method described in Rec. P.834 uses instrumental models like PESQ for this purpose. Otherwise, the method remains mainly unchanged.
- *ITU-T Rec. G.109: definition of categories of speech transmission quality* [31]: This Recommendation illustrates how *R* values obtained by the E-model may be translated to categories of speech transmission quality to be used in network planning.

Subjective evaluation methods

Whereas Sect. "State-of-the-art" gave an overview of the state-of-the-art for Recommendations focusing on subjective and/or instrumental assessment of speech quality, we will now focus on presenting and discussing the progress which has been made since 2005 for subjective evaluation methods. This includes updated versions of already mentioned documents, as well as new documents dealing with certain subjective methods. In addition, we will not solely discuss already standardized Recommendations, but also current work-items of the ITU that are about to be standardized in the near future. Again, the relevant content of each document is ordered in their logical order, and in groups of Recommendations dealing with similar topics.

The first document is the new *ITU-T Handbook on Practical Procedures for Subjective Testing* [32]. It collects a wealth of practical information which should be considered when carrying out subjective evaluations with test participants. For this aim, it contains sections on the test purpose, experimental design, conversational and listening-only tests, statistical data analysis, and result reporting. In addition, it includes a special section on the design of experiments for speech codec evaluations. Although the information included in this handbook is not new, the practical value of the information aggregation is immense.

The next two documents are an updated and a new Recommendation dealing with the *MOS terminology* and its interpretation.

- *Update P.800.1: Mean Opinion Score (MOS) terminology*: As mentioned in Sect. "State-of-the-art", the P.800.1 Recommendation has been updated three times since its first publication in 2003 [7]. The 2003 version specified whether values of MOS are related to listening quality or conversational quality, and whether they originate from subjective tests, from objective models, or from network planning models. The first update of 2006 [33] added a separation between listening, conversational, and talking MOS values as well as identifiers regarding the bandwidth (narrowband or wideband) and the type of interface (electrical or acoustical). The second update [34] extended the concept to video and audiovisual quality and provided additional identifiers regarding the video resolution. In the last updated and the currently recommended version of the document [35] a section about limitations and important notes regarding the MOS value was added.
- *New P.800.2: Mean Opinion Score interpretation and reporting*: This document, first published in 2013 [36] and slightly updated in 2016 [37], introduces some of the more common types of MOS and describes the minimum information that should be reported to enable a correct interpretation of MOS values. The Recommendation clarifies that MOS values obtained for a particular condition in a subjective experiment can be influenced by a large number of factors, such as scales, test participant instructions, stimulus presentation, equipment, or test preparation.

The following three Recommendations focus on *specific subjective evaluation methods* for certain quality values, such as conversational quality, diagnosis, or intelligibility.

- *New P.805: subjective evaluation of conversational quality* [38]: This document describes procedures for conducting conversation tests to evaluate communication quality. In particular, the recommendation shows examples of scenarios, rating scales, and analysis procedures to evaluate the subjective quality of telecommunication services. Other than passive listening-only test, conversation tests allow the simulation of more realistic situations close to the actual service

usage conditions experienced by two active interlocutors. In addition, while in passive listening tests only limited impairments can be evaluated, conversation tests are designed to assess the effects of impairments that can cause difficulty while conversing (such as delay, echo, or interruptions), and may be used to study overall system effects or specific degradations as well.

- *New P.806: a subjective quality test methodology using multiple rating scales* [39]: Integral MOS values alone do not provide diagnostic information on the reason for possibly low MOS value. On the opposite, the MOS values of two differently degraded speech samples, such as noisy speech and speech chopped by packet loss, could be identical. To analyze degradations in a more diagnostic way, Rec P.806 describes a methodology for evaluating the subjective quality of speech samples using multiple rating scales. In addition to scores for the integral quality and loudness, the methodology yields scores for six perceptual quality attributes of the speech sample (for example a slowly-varying degradation in the speech signal, or a degradation due to the level of background noise).
- *New P.807: subjective test methodology for assessing speech intelligibility* [40]: Apart from the quality and the comprehension, the intelligibility is an fundamental aspect to fully quantify the user's perception of a speech transmission system. Thus, Rec. P.807 describes a subjective testing methodology for assessing speech intelligibility. The method provides a percent correct intelligibility score based on a two-alternative forced-choice task where the stimulus is one of two words from a pair. Half of the test items are rhyming word-pairs (they differ only in the initial consonant) and the other half are alliterative word-pairs (they differ only in the final consonant). In addition to a score for overall intelligibility, the method provides scores for each of six distinctive features: voicing, nasality, sustention, sibilation, graveness and compactness. These scores may be used to diagnose the specific cause of impairments leading to degradation of speech intelligibility.

The next Recommendation is an update for the subjective evaluation method for *speech output devices*.

- *Update P.85: amendment 1: new appendix I: evaluation of speech output for audiobook reading tasks* [41]: The methods and the questionnaires presented in Rec. P.85 are adequate for services providing vocal answers related to telephone directory inquiries, weather forecast, mail order, and similar tasks. However, they are less adequate for services where longer text paragraphs or literature are read through synthetic speech output, as is the case in audiobook reading tasks. For such services, the task of the voice output is not pure information provisioning, but rather to provide an entertaining, emotion-seeking or otherwise "interesting" experience. To this end, a test methodology including the speech material, the rating scales, and the test procedure, is presented.

So far, all presented Recommendations provide methods for assessing the speech quality either in a passive listening-only situation or in an interactive two-party conversation. Since 2012, the following series of Recommendations has been approved to provide standardized methods to evaluate audio and audiovisual quality in a multiparty conference call, or *telemeeting*.

- *New P.1301: subjective quality evaluation of audio and audiovisual multiparty telemeetings* [42]: In a multiparty telemeeting, the term multiparty refers to more than two meeting participants who can be located at two or more than two locations. In this regard, Rec. P.1301 describes subjective quality assessment for telemeeting systems that provide multiparty communication between distant locations, using audio-only, video-only, audiovisual, text-based, or graphical means of communication. The Recommendation focuses on the evaluation of those systems by assessing audio-only, video-only, or audiovisual quality aspects, as well as non-interactive and conversational quality. It provides guidance and an overview of relevant aspects that need to be considered in designing an evaluation protocol.
- *New P.1302: subjective method for simulated conversation tests addressing speech and audio-visual call quality* [43]: Subjective tests with two or more participants to evaluate telemeeting systems are time and money consuming. Thus, having simulated and recorded conversations assessed by one participant minimizes the experimental effort. To this end, Rec. P.1302 describes a subjective method for assessing the quality of simulated speech or audio-visual telephony calls with time-varying transmission conditions. The simulated calls consist of several stretches of speech or audio-visual material which are ordered in a logical sequence. After each stretch, test participants have to answer a content-related question to maintain a more-or-less conversational attention focus, and they have to rate the integral quality of the call at the end of the entire sequence.
- *New P.1311: Method for determining the intelligibility of multiple concurrent talkers* [44]: More than for a two-party transmission system, the intelligibility of multiple talkers using a telemeeting system is an important aspect to fully quantify the user's perception of these systems. In this Recommendation, a method for conducting a listening test that measures the

intelligibility of multiple concurrent talkers in a telemeeting is described. This includes specifications on how to conduct such a test, stimulus design, creation of source material, selection of test conditions, as well as exemplary source material.

- *New P.1312: method for the measurement of the communication effectiveness of multiparty telemeetings using task performance* [45]: As a supplement to the three preceding Recommendations, Rec. P.1312 describes a subjective test method for quantifying the effectiveness of telemeeting systems in conveying information in multiparty conversation scenarios. The method measures the rate at which multiple participants exchange information to assess the effectiveness of communication systems compared to face-to-face communication.

In addition to the mentioned new and updated Recommendations, the ITU is currently working on three work items to standardize new subjective methods regarding the *diagnosis* of speech transmission systems. As described for Rec. P.806, gathering only the integral MOS value does not provide diagnostic information in terms of insights into possible sources of the transmission system for a potentially low MOS value. Thus, the aim of the three work items is to define subjective evaluation methods for the listening-only and the conversational situation able to diagnose the quality of transmitted speech. Two paths are conceivable for this purpose: (1) the identification of the *technical causes* of sub-optimum quality, in terms of characteristics of the signal or the transmitting system which cause the lower quality judgment; or, (2) the identification of *perceptual dimensions* of the transmitted signal—these dimensions can be considered as quality features in a multidimensional space, and the integral quality judgment can be seen as a distance to an optimum point (to the perceptual reference) in this space [46]. The three work-items are presented in the following according to the situation under test.

- *Diagnostic tests in the listening-only situation*: For path (1), ITU-T SG12 has developed a methodology for performing expert annotations after listening to transmitted speech files. This methodology may be proposed as a future P-series Recommendation "Technical Causes Analysis" (P.TCA). Its goal is to find technical causes, such as high attenuation or packet loss, by asking experts to identify perceptual impairments, such as sub-optimum speech level, or clipped speech. The underlying assumption is that most links between technical causes and perceptual impairments are "bi-unique", meaning that a given technical cause always leads to one specific perceptual impairment, and a given perceptual impairment is always caused by one specific technical cause. However, this assumption may be disputed. More precisely, different technical causes may lead to the same perception of the expert (e.g.a too low microphone signal and a too high line attenuation both lead to the expert judging "quiet speech"), and the same technical cause may also lead to different perceptual impairments (such as packet loss leading to "temporal speech clipping" and "quiet speech" in the expert judgment). For a detailed discussion of the assumption, see [47]. The P.TCA framework provides nine global categories of impairments, which are further decomposed into 47 sub-classes. The list of impairments can be found in [48]. Based on this list, expert listeners are asked to identify the most prominent degradations within each evaluated sample on a two-step approach, as described in [49]. First results and analyses of the P.TCA annotation method can be found in [47].

  For path (2), a subjective evaluation method based on semantic differential attributes has been applied and is foreseen for a future Recommendation "Assessment of Multiple Dimensions" (P.AMD) [50]. It aims at identifying and quantifying the perceptual dimensions *coloration*, *discontinuity*, *noisiness*, and *sub-optimum loudness* relevant to the integral speech quality in narrowband, wideband, and super-wideband (50–14,000 Hz) telecommunication scenarios. For information on how the four perceptual dimensions were extracted and defined see [46] or [51]. For the subjective annotation, a procedure similar to what is currently recommended for noisy speech signals is proposed (see ITU-T Rec. P.835). Thus, for the subjective direct scaling each dimension is consecutively rated on a separate continuous scale. The subjective method is described in detail in [51] and [50]. The assessment of these four perceptual dimensions shows parallels to Rec. P.806, where in sum seven perceptual dimensions are assessed. Since the both sets of perceptual dimensions are suitable for a proper diagnosis of speech transmission systems, P.AMD recommends both sets, divided into Set A (four dimensions) and Set B (seven dimensions). A comparison of both sets can for example be found in [52].
- *Diagnostic tests in the speaking and conversation situation*: Common speaking and conversation tests, as described in Rec. P.800 or Rec. P.805, provide valid methods for the integral conversational quality, but do not give insights into reasons for possible quality losses, similar to listening-only tests. In addition, speaking and conversation tests lack analytic ability, since naïve participants concentrate on the speaking or on the conversation flow. To circumvent these problems, again path (1), identifying technical causes, or

path (2), assessing perceptual dimensions, are conceivable. While path (1) has so far not been executed for the speaking or the conversational situation, ITU-T SG12 has recently started the work item "Conversational Quality Subjective" (P.CQS) to follow path (2) [53]. The aim of the work item is to approve a recommendation that describes a test methodology to diagnose the speaking and conversational situation. A potential candidate for this Recommendation as well as first results and analyses of the new candidate test method can be found in [54]. The proposed method specifically allows the participants to perceive each phase of a conversation separately (the listening phase, the speaking phase, and the interacting phase), in addition to a natural conversation, and yields integral conversational quality scores as well as quality scores for each phase. In addition, scores for multiple underlying perceptual dimensions of conversational speech quality are provided. These scores enable to analyze conversational speech quality for diagnosis and optimization. The identification of the perceptual dimensions underlying the conversational situation is presented in [55].

Instrumental quality prediction methods

Besides the advances for subjective evaluation methods, ITU-T SG12 has also been active regarding the progress of instrumental quality prediction methods since 2005. This includes new recommendations and current work items dealing with signal-based quality prediction models as well as updates of the parametric E-Model described in Rec. G.107.

The first Recommendation was approved to provide a baseline for *statistical evaluation*, qualification and comparison of instrumental quality prediction models.

- *New P.1401: methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models* [56]: During the development of an instrumental speech quality model, two fundamental steps are essential. First, one or several valid subjective quality tests have to be designed and conducted. These tests provide subjective quality ratings serving as a ground truth for the instrumental model. The second step is the design and validation of the instrumental quality model. Here, the subjective and the instrumental quality values are compared in terms of correlation and error. Thus, a stable and self-sustained statistical evaluation procedure is required in the development of instrumental quality models, and ITU-T Rec. P.1401 presents guidelines, or a framework, for this purpose. For example, it is recommended to use at least 24 votes per sample in a subjective test to assure a significant correlation with a potential instrumental quality model.

The following recommendations and current work items all describe *signal-based quality prediction models*. They include models aiming at predicting the integral quality, the intelligibility, and others which provide diagnostic information. The models either use the clean input signal and the degraded output signal of the transmission channel for their estimation (so-called full-reference approach), or only the degraded output signal (so-called no-reference approach) for their prediction. While most of these models are supposed to predict the quality in a listening-only situation, one work item develops a diagnostic signal-based instrumental quality model for the conversational situation.

- *New P.863: Perceptual Objective Listening Quality Assessment* [57]: This recommendation describes the successor of the PESQ model, the so-called Perceptual Objective Listening Quality Assessment (POLQA) model. POLQA is an instrumental quality model for predicting integral listening speech quality from narrowband to superwideband telecommunication scenarios as perceived by the user in a Rec. P.800 or Rec. P.830 ACR listening only test. The new POLQA model shows a reduction of the Root Mean Square Error Star (RMSE* [56]) by around 30% compared to the predictions of PESQ. The Recommendation presents a high-level description of the method and advice on how to use it. In 2014, an updated version of Rec. P.863 was approved [58], introducing bug fixes and resolving reported issues from POLQA field deployments.
- *New P.863.1: application guide for recommendation ITU-T P.863* [59]: In order to facilitate the usage of the new POLQA model, this Recommendation gives guidance on how to use POLQA accurately. It also provides important remarks on the speech files to be used in Rec. P.863.
- *Diagnostic full-reference quality estimation for the listening-only situation*: The test method described in Rec. P.835 was shown to provide reliable and valid results. As an instrumental counterpart, ETSI Guide EG 202 396-3 describes a model for predicting the quality of wideband and narrowband speech in noisy environments [60]. In addition, ITU-T SG12 is currently working on an independent instrumental model to predict the subjective ratings of the speech quality, the quality of the noise, and the integral quality. This work item is called "Perceptual Objective Noise Reduction" (P.ONRA) [61]. While the ETSI model is already standardized and used by industry, P.ONRA is still under development.

For predicting the speech quality experienced with super-wideband and fullband terminals in the presence of background noise, ETSI TS 103 281 [62] describes two models addressing the speech quality, background noise quality, and overall quality, as measured according to ITU-T Rec. P.835: A model which is similar to the one of [60], as well as one which is based on a detailed model of human hearing, from the ear canal to the hair cells. The Technical Specification also provides evaluation results comparing model predictions to subjective data. Further, ETSI TS 103 106 [63] describes a modification of the EG 202 396-3 model for being used with mobile terminals, as well as an evaluation of model performance.

Regarding the subjective test method described in P.AMD and Rec. P.806, ITU-T SG12 decided to develop an instrumental model to predict subjective scores for the perceptual dimensions of Set A and Set B under the work item P.AMD [50]. The model is supposed to have two operational modes, one for each set. For Set A, a potential candidate model is the so-called Diagnostic Intrusive Assessment of Listening quality (DIAL) model, presented in [64]. Based on this model, a first overview in terms of a high-level block diagram has already been proposed [65]. In addition, further potential indicators for the prospective model have just recently been presented and show to improve the model [66]. However, the potential candidate model still has to be validated and optimized on more data. For Set B, also a high-level block diagram has been presented, that needs to be validated on more Rec. P.806 data as well [67].

- *No-reference quality estimation for the listening-only situation*: The current standard Rec. P. 563 solely addresses narrowband transmission for no-reference signal-based instrumental quality estimation. Hence, ITU-T SG12 started a new standardization process to provide a no-reference model that is also suitable for wideband and super-wideband speech transmission. The work item is called "Single-ended Perceptual Evaluation of Listening Quality" (P.SPELQ) [68]. The proposed model already shows a high performance on training data, but has problems with some conditions of independent test data. In addition, the model was so far only tested on simulated speech files, and not in field tests with live recordings [69].

  In addition to the no-reference integral instrumental quality estimation, ITU-T SG12 has also started a standardization process for a no-reference diagnostic instrumental quality model, alongside the P.AMD standardization process [50, 70]. The work item is called "Single-ended Assessment of Multiple Dimensions" (P.SAMD). The approach of P.SAMD is to provide individual dimension estimators for each of the dimensions proposed in P.AMD. For Set A, first dimension estimators for noisiness [71], coloration [72], and loudness [73] show promising results. However, the amount of evaluation data is until now quite limited. Thus, further data and validation is needed until P.SAMD can be approved as a Recommendation. For Set B, so far no no-reference dimension estimators have been developed.
- *Quality estimation for the conversational situation*: Alongside the standardization of a subjective diagnostic test method for the conversational situation in P.CQS, ITU-T SG12 also aims at recommending a corresponding instrumental diagnostic conversational quality model. The standardization process is done under the working title "Conversational Quality Objective" (P.CQO) [74]. Based on the proposed subjective method for P.CQS, a first candidate model was presented in [54]. The model uses seven individual dimension estimators to predict the quality of the three conversational phases, and the integral conversational quality. Due to the difficulties to gather conversational data, the model is so far only at a very early development stage and can only provide moderate performance. However, if more data is available, the proposed model makes a promising starting point for an instrumental diagnostic conversational quality model.
- *Instrumental speech intelligibility prediction*: Due to increasing problems in speech intelligibility based on more complex telephony scenarios and non-linear speech processing, the demands for an instrumental method testing speech intelligibility raised. Therefore, ITU-T SG12 opened a work-item under the title Objective Speech Intelligibility (P.OSI). [75] provides a proposal for a benchmark procedure for assessing the performance of an instrumental intelligibility algorithm. In [76, 77], first results of potential candidate models are compared with subjective intelligibility scores (Rec. P.807). The results show that modern telecommunication networks have a serious impact on the intelligibility of speech and that the proposed models allow moderate to accurate predictions.

The following recommendations and work items describe *parametric quality prediction models*. The documents mostly refer to the E-Model and its updates towards more accurate predictions, the wideband transmission context, and diagnosis.

- *Update G.107: the E-model: a computational model for use in transmission planning* [78]: Since 2005, the E-Model has been continuously updated concerning more accurate quality predictions for codecs under dependent packet loss conditions, and to provide an

assessment of delay impairments that adapts better to less delay-sensitive use cases. In addition, an impairment factor framework for wideband speech transmission was included, but has been removed later in favour of a new, stand-alone Recommendation for a wideband E-model.

- *New G.107.1: wideband E-model* [79]: This Recommendation gives the algorithm for the wideband version of the E-model. This is a separate model that uses, within limits, similar concepts and input parameters as the model described in [78]. However, the wideband-E-Model does not cover degradations like non-optimum sidetone levels or quantizing distortions. Furthermore, for some parameter combinations of high importance (e.g., the effects of delay in conjunction with other impairments), wideband-E-model predictions are currently under study.
- *Update G.109: new appendix I—the E-model-based quality contours for predicting speech transmission quality and user satisfaction from time-varying transmission impairments* [80]: This appendix of Rec. G.109 introduces contours that can be used to predict speech transmission quality from time-varying transmission impairments. The quality contours are derived and determined from the E-model by using the rating factor *R* for all possible combinations of packet loss and mouth-to-ear delay.
- *New P.833.1: methodology for the derivation of equipment impairment factors from subjective listening-only tests for wideband speech codecs* [81]: This recommendation describes an extension of the methodology for deriving equipment impairment factors from subjective listening-only tests as described in ITU-T Rec. P.833. It is intended to be applied for determining wideband equipment impairment factors, capturing the degradation introduced by wideband speech codecs. The resulting equipment impairment factors determined by this method are intended to be used on the extended wideband-E-model transmission rating scale.
- *New P.834.1: extension of the methodology for the derivation of equipment impairment factors from instrumental models for wideband speech codecs* [82]: This document is an extension of the method for deriving equipment impairment factors from instrumental models of Rec. P.834. However, instead of using models like PESQ, wideband instrumental models like Rec. P.862.2 or Rec. P.863 are recommended to be used here.
- *Diagnostic quality prediction*: To provide diagnostic information using parametric quality prediction models, it was shown in [51] that three out of the four P.AMD dimensions may also be reliably estimated with parameters which are used by the E-Model. In other words, a dimension-based version of the E-model was developed, called the DNC model (discontinuity, noisiness, coloration). The combination of dimensions towards integral quality was performed using a Euclidean norm of a positive vector describing the respective degradation of each dimension. The integral quality results on a limited set of databases showed that the dimension-based approach could outperform the original E-model. However, the approach needs to be validated on a larger set of independent test data, and is not standardized yet.

A final set of ETSI Standards, Technical Specifications and Technical Reports addresses transmission requirements for different types of terminal equipment:

- *ETSI TS 103 737 to TS 103 740: transmission requirements for wireless terminals* [83–86]: This set of Technical Specifications describes mostly performance requirements, but also minimum quality requirements for wireless terminals, including softphones. It details test configurations, performance parameters and their measurement methodologies, but also the listening-only quality in send and receive direction, as well as the quality of background noise transmission. [83] addresses narrowband handset and headset terminals, [84] narrowband handsfree terminals, [85] wideband handset and headset terminals, and [86] wideband handsfree terminals.
- *ETSI ES 202 737 to ES 202 740: transmission requirements for VoIP terminals* [87–90]: Similar to the above Technical Specifications, this set of Standards describes mostly performance parameters, but also methods to estimate speech quality in the presence of background noise, background noise transmission, as well as listening-only quality in the send and receive direction. It proposes the use of the model described in [60] for assessing speech quality in the case of background noise, as well as the POLQA model for listening-only quality (TOSQA [91] is recommended as an alternative for this purpose) [87], addresses narrowband handset and headset terminals [88], narrowband handsfree terminals [89], wideband handset and headset terminals, and [90] wideband handsfree terminals.
- *ETSI TS 102 924 and TS 102 925: transmission requirements for super-wideband/fullband terminals* [92, 93]: Similar to the above, these specifications recommend the use of POLQA for listening-only speech quality of wideband headset [92] and handsfree [93] terminals.
- *ETSI TR 102 949: wideband and super-wideband speech terminals: perceptually motivated parameters* [94]: This Technical Report discusses parameters which are related to sub-aspects of speech quality for

wideband and super-wideband terminals, namely the loudness and the intelligibility of speech signals. It reports on subjective evaluation methods as well as computational models which might be used for predicting such effects; however, no final methods are recommended in this report. Instead, evaluation results for both aspects are given in an exemplary way.

Emerging QoE evaluation paradigms

The subjective evaluation methods which have been described so far are limited in several ways. First, they are all carried out in a laboratory environment with selected test participants who perform a (more-or-less) artificial test task under controlled conditions. A high degree of control is desirable in order to improve within-test and between-test reliability, but it may negatively affect the validity of the measurement, in case that the test conditions do not reflect real-life usage. Second, they seek conscious judgments of quality (overall quality or individual quality dimensions) of the test participants. Such a judgment situation also does not reflect the real-life, where the question of QoE might only come to the fore when the QoE level drops below a certain threshold, or when it remarkably exceeds what the user is used to. Third, the judgment on quality is given in retrospect, with the exception of the continuous judgment described in ITU-T Rec. P.880, see Sect. "State-of-the-art". While for short usage episodes the memory of the user might be adequate to accumulate the experience and to provide it with a label (from the scale), especially in long-term usage situations, and in situations where the quality level varies over time, such a retrospective judgment might not necessarily be adequate. It comes self-understood that also the instrumental models which aim at predicting the subjective judgments show the same limitations.

In the following sub-sections, we will therefore provide a brief overview on running activities which are underway in ITU-T SG12 to address these shortcomings. Two of the shortcomings have resulted in work items which aim at producing new Recommendations, whereas for the third point this is still unclear.

Crowdsourcing-based evaluation

The limitation of lacking realism due to the laboratory test setting can best be tackled by asking users in real-life situations. Carrying out quality judgment tasks under realistic conditions has since long been a topic of investigation, and methods have formerly been recommended in ITU-T Rec. P.82 (superseded). This fact is still mentioned in Section 6.3 of ITU-T Rec. P.800. In addition, this Recommendation mentions the SIBYL method where a small proportion of a user's ordinary calls inside a company are passed through special arrangements which modify the normal quality of transmission according to a test program. Such a set-up has the advantage that the normal conversational situation, including the content and the purpose of the call, are maintained. However, the set-up is quite difficult and may be limited to companies who can tolerate artificially-introduced degradations for their employees. In addition, the quality of incoming calls can only be further degraded, but not enhanced; thus, the conditions which can be tested will tend to the lower end of the scale.

A better paradigm to test speech quality in real-life situations arised with crowdsourcing platforms who offer online microjobs to a substantial number of registered workers. Workers can opt in to such microjobs which typically last for only a short period of time (minute or several minutes), and receive a payment (usually in the order of several cents or Euros/Dollars) after performing the task, and after the task result has been initially checked by the task provider. With respect to speech quality assessment, such tasks may consist of listening to speech samples online using a computer, and judging their quality. More ambitious tasks include connecting two workers in order to perform a conversation or a talking-and-listening test, as this requires temporal synchronization of the workers.

Whereas it seems that laboratory test tasks may easily be transferred to a crowdsourcing scenario, a detailed analysis reveals many differences the consequences of which are still unknown. First, the test environment in the crowd is mostly uncontrolled. This relates to the equipment used for listening to speech samples (or for recording them from workers, in case of a conversation test), its connection to the computer hardware (including soundcards, level adjustment, etc.), the internet connection (which may be unreliable), as well as the room the task is performed in (reverberation, background noise, etc.). In addition, the worker is not fully under control of the experimenter: it may happen that s/he carries out other parallel tasks, is not focussed on the task, or is simply ignoring it to a large extent. Such behaviour might be counteracted with trapping questions or alike, but cannot fully be excluded. Also the workers themselves are mostly unknown to the experimenter, including potential deficits in perceiving (hearing loss) and understanding the instructions (e.g. due to language problems).

Even if these problems were under control, it is probable that the test task should be organized in a different way. As microtasks are commonly short in duration, an entire test session which would be common for a laboratory test would have to be split into several microtasks for finding a

sufficient number of workers. This way, the range of test conditions listened to by each worker is different from the laboratory, a fact that might impact the results. As computer interfaces differ in their input (mouse, touch-screen) and output (size. resolution) capabilities, the answering format might not be the same for each worker, and it might not be always optimum for the test task under consideration. All these test-specific factors are largely unknown, and a thorough analysis is necessary before coming up with a recommended method. In turn, if such a method becomes available, the results which can be obtained with it would reflect real-life usage situations—including their inherent variability—in a much better way than controlled laboratory conditions can. In addition, it is expected that crowdsourcing tests are quicker and cheaper compared to laboratory tests. This is why ITU-T SG12 is currently working on a new work item P.CROWD which should result in a new Recommendation on how crowdsourcing-based evaluation should be performed for speech services [95, 96].

Physiology-based evaluation

Whether being carried out in the lab or in the crowd, all subjective test methods described so far make use of *conscious* judgments of quality, i.e. it is the explicit task of test participants to judge quality. This paradigm does not reflect real-life service usage, where users are rarely asked for their quality judgment. Thus, putting the focus on the quality may distort the obtained results. In addition, with the exception of the continuous quality judgment task of ITU-T Rec. P.880, all methods solicit quality judgments in retrospect. This will recur to memory effects which have not been fully understood yet.

A different test paradigm would be to collect reactions from test participants *which reflect their momentary quality experience*. Such reactions may be communication behaviour (e.g. backchannels in case of non-intelligible speech), but especially for more subtle degradations physiological reactions from test participants may be promising indicators for experienced quality. Physiological reactions can either stem from the peripheral system (such as skin conductance, muscle movements, e.g. in the face) of from the central nervous system (such as brain activity shown by Electro-Encephalography, EEG, or bloodflow shown by Near Infrared Spectroscopy, NIRS).

A number of investigations have shown that long-term exposition to quality-impaired stimuli resulted in an increase in the relative power of alpha (8–12 Hz) and theta (4–8 Hz) frequency bands located in frontal and parietal-occipital areas, when measured with an EEG. These effects reflect emotional processing and fatigue, respectively [97, 98]. Rapid, short-term (phasic) changes in neuro-electric activity time-locked to the onset of a defined stimulus event manifest in the EEG signal as a specific sequence of event-related potential components. Attentional and cognitive processes are particularly associated with the so-called “P300” component which refers to a positive voltage change occurring approx. 300 ms after the onset of an unexpected and meaningful event. It was shown that a change in P300 amplitude and lag could partially be associated with quality degradations, both in the visual and in the auditory/speech domain [99, 100].

Other physiological metrics have been proposed. For example, skin conductance has shown to correlate with affective arousal [101, 102]. The valence of such an arousal, i.e. whether it is connotated to positive or negative emotions, can e.g. be measured with an Electro-Myogram (EMG) which registers wrinkles around the eyes [103]. Whereas such peripheral measurements can be obtained with relatively low experimental effort compared to central reactions (such as EEG or NIRS), they have not yet been shown to be in a direct relation to perceived QoE. Unfortunately, physiology-related signals are inherently noisy, and their acquisition and analysis requires a significant amount of expertise. This is why ITU-T SG12 has decided to provide a new Recommendation on the test set-up and test procedures to be used with physiology-based evaluations, so that results are meaningful and reproducible [104, 105]. It is expected that the methods to-be-recommended will provide a valuable add-on complementing traditional opinion-test methods regarding *unconscious* and *continuous* indications of perceived QoE.

Episodic and multi-episodic quality evaluation

As most methods currently recommended by ITU solicit one judgment after perceiving a stimulus (listening to a speech probe or carrying out a conversation), the temporal development of QoE during stimulus perception remains uncovered. Temporal changes in QoS *do* however occur, in particular with modern time-varying transmission techniques, such as mobile and IP-based telephony. These QoS changes affect also perception and experience. A number of temporal effects were found in QoE research on short periods from seconds up to several minutes, see. e.g. [106] for a review. For example, the primacy effect and the recency effect describe a more severe impact on a retrospective judgment from phases at the beginning and at the end of an experience episode, respectively. Duration neglect has also been found, i.e. that the length of an episode has a rather small effect on the retrospective judgment. These effects were so far mainly assessed for single usage episodes. E.g., the method defined in ITU-T Rec. P.1302 [43] tries to simulate conversation behaviour of one call of approximately 1 to 2 min by listening to logically-

concatenated speech samples. Instrumental prediction models have been developed as well that consider those effects and improve prediction performance on the time scale of one usage episode, both for speech-only and for video-telephony services [106–109]. None of these models has yet been standardized by ITU-T, but ETSI mentions a model in its Technical Report 102 506 [110].

In addition to the temporal changes *within* one usage episode, QoE might also change *between* multiple usage episodes, e.g. between speech calls. The analysis of such changes is important for speech service providers, as their services are commonly used on a regular basis, namely via a subscription. As usage episodes are commonly separated in time, an ideal way to analyze such *multi-episodic* quality are field tests, in which users are asked to use a particular service repeatedly over e.g. 1 to 2 weeks. During this usage period, the quality of the individual calls is deliberately manipulated according to pre-defined QoS profiles. Users are then asked to judge the quality after each individual call they carried out, and in addition, a judgment summarizing multi-episodic usage is solicited after a certain number of days [111]. With the help of this method, it becomes possible to analyze the temporal effect of individual episodic QoE judgments on the multi-episodic QoE, i.e. the QoE judgment related to the entire service usage period [112]. The execution of such experiments in the field is however difficult, due to technical problems at the test participants' homes, as well as due to the difficultly to motivate test participants to execute such tests over a long period of time. As a consequence, also shorter usage periods (such as an hour) have been used to analyze multi-episodic QoE. Instrumental models predicting the observed effects are still very limited, as first approaches in [112] show.

Concluding remarks

We have presented a review of standardization activities addressing the quality of speech communication which have been undertaken within the last 12 years mostly in Study Group 12 of ITU-T, and some by ETSI, and have highlighted the Recommendations and Recommendation updates which have resulted from these activities. The results of this work show several strategic directions which have been taken by the respective standardization bodies:

- *Transition from narrowband to wideband, super-wideband and fullband transmission*: Whereas nearly all standards available in 2005 addressed narrowband speech transmission, there is a strong tendency towards higher audio bandwidths. The reason is that terminal devices change quickly, and thus a change in the transmission protocol, including a codec with a higher audio bandwidth, can easily be implemented—the network bandwidth not being a determinant factor any more. During the transition, artificial bandwidth extension algorithms become relevant, but we expect that their importance will diminish with the introduction of full IP-based systems.
- *Need for diagnostic information*: Quite a lot of standardization effort is dedicated to obtaining diagnostic information for service optimization, both on the subjective and on the instrumental side (see the work items P.AMD, P.ONRA, P.TCA, P.SAMD, P.CQS, P.CQO). The reason for this tendency is obvious: efficient service provisioning which has the target of highest-possible QoE does not only require efficient measurement and prediction of QoE, but also indicators towards the reason of (potentially) low QoE, in the QoS domain. Whereas the respective algorithms are still in their definition phase, we see a considerable demand for this type of information.
- *Need for better validity of experience metrics*: Whereas most traditional methods start from short speech samples as representatives for a communication service, real usage is interactive, and involves longer periods of exposure to and usage of speech. As a consequence, we observe a tendency to move from listening-only tests to talking-and-listening or conversation tests, from individual (approximately 4–8 s long) speech samples to simulated or real calls, and from single calls to multiple calls, being more representative of service usage than short stretches of speech. These tendencies underline the need for *valid* measurements, reflecting the object of measurement in a better way. Along the same lines we can see the tendency from lab towards crowd experiments (increasing ecological validity), and from post-experience opinion tests towards during-experience physiological measurements (increasing temporal validity). The aimed-for increase in validity may come at the expense of reduced reliability (e.g. due to the unforeseeable structure of conversations) and reduced sensitivity (e.g. due to the focus of attention being on the conversation topic, and not on the quality-rating task). However, we think that the obstacles introduced by the new evaluation methods are not insurmountable, the just require more experience and statistical validation.

As the standardization bodies are steered mostly by industry members, these strategic directions apparently have some industrial relevance, justifying the efforts spent in the standardization process.

In addition to the points raised in Sect. "Emerging QoE evaluation paradigms", we foresee the need for further

work which has not yet started in the respective standardization bodies, but which will put new challenges to the QoE of speech services and its evaluation:

- As conversation test results are heavily influenced by the conversation behaviour of (randomly selected) test participants, it would be desirable to *simulate interaction behaviour* in a realistic way. Such a simulation could complement instrumental prediction models for the conversational situation, and could help to build a fully-automated test cycle for interactive speech services. A realistic simulation would however have to behave like "normal" humans do in a conversation, also in the presence of degradations (muted channels, delay, echo); it is yet unclear how such a simulation could be built, and how realistic simulated interactions are.
- The conversation behaviour also becomes important when the spoken interaction is not between humans, but between humans and machines. The advent of *personal speech-based assistants* (Siri, Cortana, Google Assistant, Alexa, etc.) shows that such services are highly demanded, but methods for evaluating their QoE are still sparse, specially when being used as parts of a telecommunication service.
- The step from two-party conversations to telemeetings already shows that conversations may be multi-party, and that they may be operated using a number of different terminal devices. Whereas handsets, hands-free terminals and headsets are also common for two-party conversations, speaker and listener position, as well as room characteristics, become more important in such situations. We foresee a transition towards spatial audio, presented either with multiple loudspeakers or via headphones, making use of multiple distributed microphones, and including physically present speakers together with remotely presented ones. In such *augmented-reality situations*, concepts like presence, immersion and involvement come to the fore, along with speaker identifiability and intelligibility, and need to be considered when judging on the QoE of the conversation situation.
- Whereas we limited our review to speech-only communication situations, the introduction of all-IP services easily allows to *augment speech with non-speech* audio signals (music, noise), and to augment it with a visual representation. Whereas audio-visual integration has been a topic in QoE research for a long time, diagnostic methods are still sparse in this area.

We expect that some of these gaps will be taken up as work items by ITU-T and ETSI, or by other standardization bodies.

Acknowledgements We would like to thank all colleagues from SG12 at ITU-T and from ETSI for their fruitful collaboration over the last years. We extend our thanks to the anonymous reviewers of an earlier version of this article for their fruitful comments and suggestions.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. ITU-T Recommendation P.10/G.100 (2006) Amendment 3 (12/11), New Definitions for Inclusion in Recommendation ITU-T P.10/G.100, International Telecommunication Union, Nov. 2011
2. ITU-T Recommendation P.10/G.100 (2006) Amendment 5 (07/16), New Definitions for Inclusion in Recommendation ITU-T P.10/G.100, International Telecommunication Union, July 2016
3. Le Callet P, Möller S, Perkis A (eds) (2013) Qualinet white paper on definitions of quality of experience. In: European network on quality of experience in multimedia systems and services (COST Action IC 1003), Mar. 2013
4. ITU-T Handbook on Telephonometry, International Telecommunication Union, 1992
5. ITU-T Recommendation P.800, Methods for Subjective Determination of Transmission Quality, International Telecommunication Union, Aug. 1996
6. ITU-T Recommendation P.810, Modulated Noise Reference Unit (MNRU), International Telecommunication Union, Feb. 1996
7. ITU-T Recommendation P.800.1, Mean Opinion Score (MOS) Terminology, International Telecommunication Union, Mar. 2003
8. Knoche H, de Meer H, Kirsh D (1999) Utility curves: mean opinion scores considered biased. In: 7th Int. workshop on quality of service (IWQoS '99)
9. Möller S (2000) Assessment and prediction of speech quality in telecommunications. Springer, Boston
10. ITU-T Recommendation P.880, Continuous Evaluation of Time-varying Speech Quality, International Telecommunication Union, May 2004
11. Borowiak A, Reiter U, Svensson P (2012) Quality evaluation of long duration audiovisual content. In: IEEE 2012 consumer communications and networking conference (CCNC), pp 737–741
12. ITU-T Recommendation P.830, Subjective Performance Assessment of Telephone-band and Wideband Digital Codecs, International Telecommunication Union, Feb. 1996
13. ITU-T Recommendation P.831, Subjective Performance Evaluation of Network Echo Cancellers, International Telecommunication Union, Dec. 1998
14. ITU-T Recommendation P.832, Subjective Performance Evaluation of Hands-free Terminals, International Telecommunication Union, May 2000
15. ITU-T Recommendation P.835, Subjective Test Methodology for Evaluating Speech Communication Systems that Include Noise Suppression Algorithm, International Telecommunication Union, Oct. 2003
16. ITU-T Recommendation P.840, Subjective Listening Test Method for Evaluating Circuit Multiplication Equipment, International Telecommunication Union, Nov. 2003

17. ITU-T Recommendation P.85, A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices, International Telecommunication Union, June 1994
18. ITU-T Recommendation P.851, Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems, International Telecommunication Union, Nov. 2003
19. ITU-T Recommendation P.862, Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-end Speech Quality Assessment of Narrow-band Telephone Networks and Speech Codecs, International Telecommunication Union, Feb. 2001
20. ITU-T Recommendation P.862.1, Mapping Function for Transforming P.862 Raw Result Scores to MOS-LQO, International Telecommunication Union, Nov. 2003
21. ITU-T Recommendation P.862.2, Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs, International Telecommunication Union, Nov. 2005
22. ITU-T Recommendation P.862.3, Application Guide for Objective Quality Measurement Based on Recommendations P.862, P.862.1 and P.862.2, International Telecommunication Union, Nov. 2005
23. ITU-T Recommendation P.563, Single-ended Method for Objective Speech Quality Assessment in Narrow-band Telephony Applications, International Telecommunication Union, May 2004
24. ETSI Guide ETSI EG 201 377-1 V1.2.1, Speech processing, Transmission and Quality Aspects (STQ); Specification and Measurement of Speech Transmission Quality; Part 1: Introduction to Objective Comparison Measurement Methods for One-way Speech Quality Across Networks, European Telecommunications Standards Institute, Dec. 2002
25. ETSI Guide 201 377-3 V1.1.1, Speech Processing, Transmission and Quality Aspects (STQ); Specification and Measurement of Speech Transmission Quality; Part 3: Non-intrusive Objective Measurement Methods Applicable to Networks and Links With Classes of Services, European Telecommunications Standards Institute, June 2003
26. ETSI Technical Report ETR 250, Transmission and Multiplexing (TM); Speech Communication Quality from Mouth to Ear for 3.1 kHz Handset Telephony across Networks, European Telecommunications Standards Institute, July 1996
27. ITU-T Recommendation G.107, The E-Model, a Computational Model for Use in Transmission Planning, International Telecommunication Union, Mar. 2005
28. ITU–T Recommendation P.833, Methodology for Derivation of Equipment Impairment Factors From Subjective Listening-only Tests, International Telecommunication Union, Feb. 2001
29. ITU-T Recommendation G.113, Transmission Impairments Due to Speech Processing, International Telecommunication Union, Nov. 2007
30. ITU–T Recommendation P.834, Methodology for the Derivation of Equipment Impairment Factors From Instrumental Models, International Telecommunication Union, July 2002
31. ITU-T Recommendation G.109, Definition of Categories of Speech Transmission Quality, International Telecommunication Union, Sept. 1999
32. ITU-T Handbook on Practical Procedures for Subjective Testing, International Telecommunication Union, 2011
33. ITU-T Recommendation P.800.1, Mean Opinion Score (MOS) Terminology, International Telecommunication Union, July 2006
34. ITU-T Recommendation P.800.1, Mean Opinion Score (MOS) Terminology, International Telecommunication Union, Feb. 2016
35. ITU-T Recommendation P.800.1, Mean Opinion Score (MOS) Terminology, International Telecommunication Union, July 2016
36. ITU-T Recommendation P.800.2, Mean Opinion Score Interpretation and Reporting, International Telecommunication Union, May 2013
37. ITU-T Recommendation P.800.2, Mean Opinion Score Interpretation and Reporting, International Telecommunication Union, July 2016
38. ITU–T Recommendation P.805, Subjective Evaluation of Conversational Quality, International Telecommunication Union, Apr. 2007
39. ITU–T Recommendation P.806, A Subjective Quality Test Methodology Using Multiple Rating Scales, International Telecommunication Union, Feb. 2014
40. ITU–T Recommendation P.807, Subjective Test Methodology for Assessing Speech Intelligibility, International Telecommunication Union, Feb. 2016
41. ITU-T Recommendation P.85, A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices—Amendment 1: New Appendix I Evaluation of speech output for audiobook reading tasks, International Telecommunication Union, Mar. 2003
42. ITU-T Recommendation P.1301, Subjective Quality Evaluation of Audio and Audiovisual Multiparty Telemeetings, International Telecommunication Union, Sept. 2012
43. ITU-T Recommendation P.1302, Subjective Method for Simulated Conversation Tests Addressing Speech and Audiovisual Call Quality, International Telecommunication Union, Oct. 2014
44. ITU-T Recommendation P.1311, Method for Determining the Intelligibility of Multiple Concurrent Talkers, International Telecommunication Union, Dec. 2014
45. ITU-T Recommendation P.1312, Method for the Measurement of the Communication Effectiveness of Multiparty Telemeetings Using Task Performance, International Telecommunication Union, Feb. 2016
46. Möller S, Garcia M-N, Wältermann M (2014) Quality of experience: advanced concepts, applications and methods. In: Features of quality of experience. Springer, Heidelberg
47. Köster F, Schiffner F, Möller S, Malfait L (2017) Towards degradation decomposition for voice communication system assessment. Qual User Exp 2(1):4
48. ITU-T Temporary Document TD 650rev1 (GEN/12), Requirement Specifications for P.TCA (Technical Cause Analysis), Rapporteur Q.16/12 (L. Malfait), International Telecommunication Union, ITU-T SG12 Meeting, 2011
49. ITU-T Temporary Document TD 686 (GEN/12), Expert Listening for P.TCA, Rapporteur Q.16/12 (L. Malfait), International Telecommunication Union, ITU-T SG12 Meeting, 2011
50. ITU-T Temporary Document TD 137 (GEN/12), Technical Requirement Specification P.AMD and P.SAMD, Rapporteur Q.9/12 (J. Berger), International Telecommunication Union, ITU-T SG12 Meeting, 2017
51. Wältermann M (2013) Dimension-based quality modeling of transmitted speech. In: T-Labs series in telecommunication services. Springer, Berlin
52. ITU-T Contribution COM 12-212, Proposal for a Framework for P.MULTI and P.AMD, Source: Deutsche Telekom AG (Authors: M. Wältermann, S. Möller), International Telecommunication Union, ITU-T SG12 Meeting, June 2011
53. ITU-T Temporary Document TD 965rev1 (GEN/12), Revised Status Report of Question 7/12, Rapporteur Q.7/12 (P. Usai), International Telecommunication Union, ITU-T SG12 Meeting, 2016

54. Köster F (2017) Multidimensional Analysis of Conversational Telephone Speech, Doctoral Dissertation. Technische Universität Berlin, Fakultät für Elektrotechnik und Informatik
55. Köster F, Guse D, Möller S (2017) Identifying speech quality dimensions in a telephone conversation. Acta Acustica united with Acustica 103(3):506–522
56. Recommendation ITU-T, P.1401, Methods. Metrics and Procedures for Statistical Evaluation, Qualification and Comparison of Objective Quality Prediction Models, International Telecommunication Union, July 2012
57. ITU–T Recommendation P.863, Perceptual Objective Listening Quality Assessment, International Telecommunication Union, Jan. 2011
58. ITU–T Recommendation P.863, Perceptual Objective Listening Quality Assessment, International Telecommunication Union, Sept. 2014
59. ITU–T Recommendation P.863.1, Application Guide for Recommendation ITU-T P.863, International Telecommunication Union, Sept. 2014
60. ETSI Guide 202 396-3 V1.6.1, Speech and multimedia Transmission Quality (STQ); Speech Quality Performance in the Presence of Background Noise; Part 3: Background Noise Transmission – Objective Test Methods, European Telecommunications Standards Institute, Jan. 2017
61. ITU-T Contribution COM 12-210, P.ONRA Requirement Specification, Draft V0.5, Source: Qualcomm, Inc. (Authors: A. Schevciw), International Telecommunication Union, ITU-T SG12 Meeting, Sept. 2014
62. ETSI Technical Specification 103 281 V1.1.1, Speech and multimedia Transmission Quality (STQ); Speech Quality in the Presence of Background Noise: Objective Test Methods for Super-wideband and Fullband Terminals, European Telecommunications Standards Institute, Apr. 2017
63. ETSI Technical Specification 1103 106 V1.4.1, Speech and multimedia Transmission Quality (STQ); Speech Quality Performance in the Presence of Background Noise: Background Noise Transmission for Mobile Terminals—Objective Test Methods, European Telecommunications Standards Institute, Nov. 2016
64. Côté N (2011) Integral and diagnostic intrusive prediction of speech quality. In: T-Labs series in telecommunication services. Springer, Berlin
65. ITU-T Contribution COM 12-197, P.AMD High Level Block Diagram for Set A, Source: Deutsche Telekom AG/Orange (Authors: F. Köster, V. Barriac, S. Möller), International Telecommunication Union, ITU-T SG12 Meeting, Sept. 2014
66. ITU-T Contribution COM 12-24, First Results from the POLQA P.AMD Degradation Decomposition Approach Using Noisiness, Discontinuity, Coloration and Loudness Indicators, Source: OPTICOM GmbH, Rohde & Schwarz GmbH & Co KG, TNO (Authors: C. Schmidmer, J. Beerends, A. Llagostera), International Telecommunication Union, ITU-T SG12 Meeting, Jan. 2017
67. ITU-T Contribution COM 12-224, P.AMD High Level Block Diagram for Set B, Source: Qualcomm, Inc. (Authors: A. Schevciw, W. Lu. D. Sen), International Telecommunication Union, ITU-T SG12 Meeting, Sept. 2014
68. ITU-T Temporary Document TD 437 Rev.1 (GEN/12), Technical Requirement Specification P.SPELQ, Rapporteur Q.9/12 (J. Berger), International Telecommunication Union, ITU-T SG12 Meeting, 2014
69. ITU-T Contribution COM 12-337, P.SPELQ Interim Model Performance Results, Source: Huawei Technologies Co. Ltd. (Authors: W. Xiao, P. Coverdale, C. Ju), International Telecommunication Union, ITU-T SG12 Meeting, Dec. 2015
70. ITU-T Contribution COM 12-387, Proposal for a New Work Item on Diagnostic Single-Ended Quality Indicators, Source: Deutsche Telekom AG (Authors: S. Möller, F. Köester), International Telecommunication Union, ITU-T SG12 Meeting, May 2016
71. Köster F, Mittag G, Polzehl T, Möller S (2016) Non-intrusive estimation of noisiness as a perceptual quality dimension of transmitted speech. In: Proc. 5th international workshop on perceptual quality of systems (PQS 2016), pp 74–78
72. Mittag G, Köster F, Möller S (2016) Non-intrusive estimation of the perceptual dimension coloration. In: Fortschritte der Akustik, DAGA 2016: Plenarvortr. u. Fachbeitr. d. 42. Dtsch. Jahrestg. f. Akust. DEGA
73. Köster F, Cercos-Llombart V, Mittag G, Möller S (2016) Non-intrusive estimation model for the speech-quality dimension loudness. In: Informationstechnische Gesellschaft im VDE (ITG) Conference on Speech Communication, ITG, pp 175–179
74. ITU-T Temporary Document TD 27 (WP 2/12), Proposed Scope for P.CQO, Rapporteur Q.15/12 (J. Pomy), International Telecommunication Union, ITU-T SG12 Meeting, 2005
75. ITU-T Contribution COM 12-7, Benchmark Procedure Proposal for the Assessment of Objective Speech Intelligibility Assessment Methods, Source: TNO, The Netherlands (Authors: J. Beerends), International Telecommunication Union, ITU-T SG12 Meeting, Feb. 2013
76. Pocta P, Beerends JG (2015) Subjective and objective measurement of synthesized speech intelligibility in modern telephone conditions. Speech Commun 71:1–9
77. ITU-T Contribution COM 12-93, Subjective and Objective Measurement of Synthesized Speech Intelligibility in Modern Telephone Conditions, Source: Ministry of Transport, Construction and Regional development of the Slovak Republic; TNO, The Netherlands (Authors: P. Počta, J. Beerends), International Telecommunication Union, ITU-T SG12 Meeting, Feb. 2013
78. ITU-T Recommendation G.107, The E-Model, a Computational Model for Use in Transmission Planning, International Telecommunication Union, June 2015
79. ITU-T Recommendation G.107.1, Wideband E-model, International Telecommunication Union, June 2015
80. ITU-T Recommendation G.109—Amendment 1, Amendment 1: New Appendix I The E-Model-Based Quality Contours for Predicting Speech Transmission Quality and User Satisfaction from Time-Varying Transmission Impairments, International Telecommunication Union, Jan. 2007
81. ITU-T Recommendation P.833.1, Methodology for the Derivation of Equipment Impairment Factors from Subjective Listening-Only Tests for Wideband Speech Codecs, International Telecommunication Union, Apr. 2009
82. ITU-T Recommendation P.834.1, Extension of the Methodology for the Derivation of Equipment Impairment Factors from Instrumental Models for Wideband Speech Codecs, International Telecommunication Union, June 2015
83. ETSI Technical Specification 103 737 V1.2.1, Speech and multimedia Transmission Quality (STQ); Transmission Requirements for Narrowband Wireless Terminals (Handset and Headset) From a QoS Perspective as Perceived by the User, European Telecommunications Standards Institute, July 2017
84. ETSI Technical Specification 103 738 V1.2.1, Speech and multimedia Transmission Quality (STQ); Transmission Requirements for Narrowband Wireless Terminals (Handsfree) from a QoS Perspective as Perceived by the User, European Telecommunications Standards Institute, July 2017
85. ETSI Technical Specification 103 739 V1.2.1, Speech and multimedia Transmission Quality (STQ); Transmission Requirements for Wideband Wireless Terminals (Handset and

Headset) from a QoS Perspective as Perceived by the User, European Telecommunications Standards Institute, July 2017
86. ETSI Technical Specification 103 740 V1.2.1, Speech and multimedia Transmission Quality (STQ); Transmission Requirements for Wideband Wireless Terminals (Handsfree) from a QoS Perspective as Perceived by the User, European Telecommunications Standards Institute, July 2017
87. Final Draft ETSI Standard 202 737 V1.7.1, Speech and multimedia Transmission Quality (STQ); Transmission Requirements for Narrowband VoIP Terminals (Handset and Headset) from a QoS Perspective as Perceived by the User, European Telecommunications Standards Institute, July 2017
88. Final Draft ETSI Standard 202 738 V1.7.1, Speech and multimedia Transmission Quality (STQ); Transmission Requirements for Narrowband VoIP Loudspeaking and Handsfree Terminals from a QoS Perspective as Perceived by the User, European Telecommunications Standards Institute, July 2017
89. Final Draft ETSI Standard 202 739 V1.7.1, Speech and multimedia Transmission Quality (STQ); Transmission Requirements for Wideband VoIP Terminals (Handset and Headset) from a QoS Perspective as Perceived by the User, European Telecommunications Standards Institute, July 2017
90. Final Draft ETSI Standard 202 740 V1.7.1, Speech and multimedia Transmission Quality (STQ); Transmission Requirements for Wideband VoIP Loudspeaking and Handsfree Terminals from a QoS Perspective as Perceived by the User, European Telecommunications Standards Institute, July 2017
91. ETSI Guide ETSI EG 201 377-1 V1.3.2, Speech and multimedia Transmission Quality (STQ); Specification and Measurement of Speech Transmission Quality; Part 1: Introduction to Objective Comparison Measurement Methods for One-way Speech Quality Across Networks, European Telecommunications Standards Institute, Oct. 2009
92. ETSI Technical Specification 102 924 V1.1.1, Speech and multimedia Transmission Quality (STQ); Transmission Requirements for Superwideband/Fullband Headset Terminals from a QoS Perspective as Perceived by the User, European Telecommunications Standards Institute, Mar. 2013
93. ETSI Technical Specification 102 925 V1.1.1, Speech and multimedia Transmission Quality (STQ); Transmission Requirements for Superwideband/Fullband Handsfree and Conferencing Terminals from a QoS Perspective as Perceived by the User, European Telecommunications Standards Institute, Mar. 2013
94. Technical Report ETSI, 102 949 V1.1.1, Speech and multimedia Transmission Quality (STQ). Wideband and Superwideband Speech Terminals; Perceptually Motivated Parameters, European Telecommunications Standards Institute, Sept. 2014
95. ITU-T Contribution COM 12-242, Proposal for a New Recommendation P.CROWD: Subjective Methods for Assessing Audio Quality With a "Crowdsourcing" Approach, Source: Orange Labs (Authors: L. Gros, C. Quinquis), International Telecommunication Union, ITU-T SG12 Meeting, May 2015
96. ITU-T Contribution COM 12-386, Comments on P.CROWD, Source: Deutsche Telekom AG (Authors: S. Möller, T. Hoßfeld, B. Naderi), International Telecommunication Union, ITU-T SG12 Meeting, June 2016
97. Antons J-N, Köster F, Arndt S, Schleicher R, Möller S (2013) Changes of vigilance caused by varying bitrate conditions. In: Fifth Int. workshop on quality of multimedia experience 2013 (QoMEX 2013), Klagenfurt, pp 148–151
98. Kroupi E, Hanhart P, Lee JS, Rerabek M, Ebrahimi T (2014) EEG correlates during video quality perception. In: 22nd European signal processing conference (EUSIPCO 2014), pp 2135–2139
99. Antons J-N (2015) Neural correlates of quality perception for complex speech signals. Springer, Cham
100. Sebastian A (2016) Neural correlates of quality during perception of audiovisual stimuli. Springer, Cham
101. Cacioppo JT, Tassinary LG, Berntson GG (eds) (2000) Handbook of psychophysiology. Cambridge University Press, New York
102. Backs RW, Boucsein W (eds) (2000) Engineering psychophysiology: issues and applications. Lawrence Erlbaum Associates, Mahwah
103. Lassalle J, Gros L, Coppin G (2011) Combination of physiological and subjective measures to assess quality of experience for audiovisual technologies. In: 2011 Third international workshop on quality of multimedia experience, pp 13–18
104. ITU-T Contribution COM 12-272, Proposal to Work on a Recommendation on the Use of Physiological Measures as an Additional Test Method for Speech Quality Assessment, Source: Deutsche Telekom AG (Authors: S. Möller, J.-N. Antons, S. Arndt), International Telecommunication Union, ITU-T SG12 Meeting, May 2015
105. ITU-T Contribution COM 12-307, Structure Proposal for the Work Item on the Use of Physiological Measures as an Additional Test Method for Speech Quality Assessment (P.PHYSIO), Source: Deutsche Telekom AG (Authors: J.-N. Antons, S. Arndt, S. Möller), International Telecommunication Union, ITU-T SG12 Meeting, Jan. 2016
106. Weiss B, Guse D, Möller S, Raake A, Borowiak A, Reiter U (2014) Quality of experience: advanced concepts, applications and methods. In: Temporal development of quality of experience. Springer, Heidelberg
107. Gros L, Chateau N (2001) Instantaneous and overall judgments for time-varying speech quality: assessment and prediction. Acta Acustica united with Acustica 87(3):367–377
108. Clark AD (2001) Modeling the effects of burst packet loss and recency on subjective voice quality. In: Proc. 2nd IP-telephony workshop, New York
109. Belmudez B, Lewcio B, Möller S (2013) Call quality prediction for audiovisual time-varying impairments using simulated conversational structures. Acta Acustica united with Acustica 99(5):792–805
110. ETSI Technical Report 102 506 V1.4.1, Speech and multimedia Transmission Quality (STQ); Estimating Speech Quality per Call, European Telecommunications Standards Institute, Aug. 2011
111. Möller S, Bang C, Tamme T, Vaalgamaa M, Weiss B (2011) From speech quality to service quality: a study on long-term quality integration in audio-visual speech communication services. In: Proc. 12th Ann. Conf. Of the Int. Speech Communication Assoc. (Interspeech 2011), Firenze
112. Guse D (2016) Multi-episodic Perceived Quality of Telecommunication Services, Doctoral Dissertation, Fakultät für Elektrotechnik und Informatik, Technische Universität Berlin