Sebastian Möller, Wai-Yip Chan, Nicolas Côté,
Tiago H. Falk, Alexander Raake, and Marcel Wältermann

Speech Quality Estimation

[Models and trends]

© INGRAM PUBLISHING

This article presents a tutorial overview of models for estimating the quality experienced by users of speech transmission and communication services. Such models can be classified as either parametric or signal based. Signal-based models use input speech signals measured at the electrical or acoustic interfaces of the transmission channel. Parametric models, on the other hand, depend on signal and system parameters estimated during network planning or at run time. This tutorial describes the underlying principles as well as advantages and limitations of existing models. It also presents new developments, thus serving as a guide to an appropriate usage of the multitude of current and emerging speech quality models.

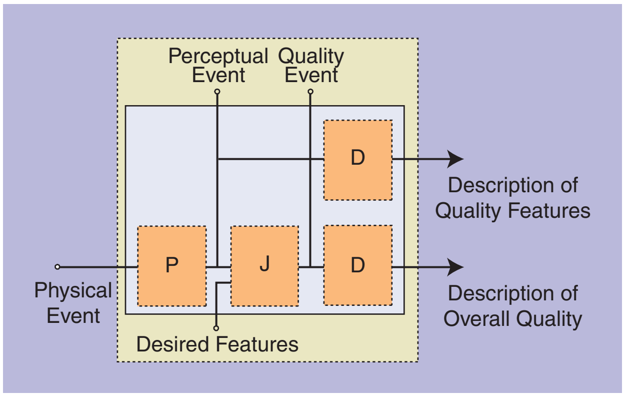
INTRODUCTION

Since the large-scale introduction of telephony networks, efforts have been made to guarantee high-quality and reliable services to human users. Transmission performance was initially measured by informally exchanging phonetically rich phrases between two network terminals, thus quantifying the intelligibility associated with the channel. Later, such informal procedures were replaced by standardized listening-only and conversational tests that provided more stable conditions—and thus smaller confidence intervals—when asking test participants to rate the (perceived) loudness or intelligibility, listening effort, or overall quality of the heard speech samples or conversations [28].

For a participant in a quality judgment experiment, for example, speech quality is regarded as a multidimensional construct, as it is the result of three processes: perception (P), judgment (J), and description (D), as depicted in Figure 1. The perception processes are triggered by a so-called "physical event" (i.e., a sound wave reaching the human ears), which gives rise to a "perceptual event." We use the term "event" to denote an instance of occurrence of a phenomenon in time and space; see [4]. This "perceptual event" can also be described in a multidimensional way, wherein

Digital Object Identifier 10.1109/MSP.2011.942469
Date of publication: 1 November 2011

1053-5888/11/$26.00©2011IEEE



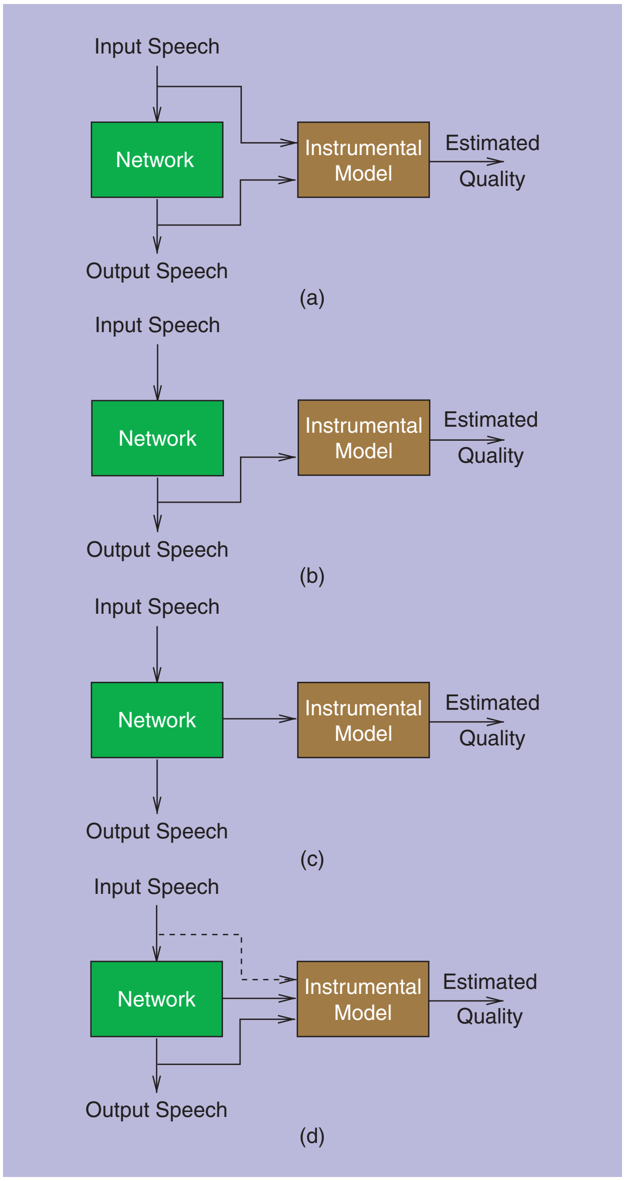
[FIG1] Schematic representation of a participant in a quality judgment experiment; see [55].

features such as loudness, coloration, or noisiness are quantified [55], [60], [61].

The features of the perceptual event are further compared to the desired features of some internal reference [42], [55]. This reference can be formed via repeated telephone usage experiences, but it also reflects numerous context- and situation-dependent factors such as the user's expectations (e.g., free versus paid call), motivation (e.g., urgent call), and experience (e.g., avid mobile phone user); the test setup (e.g., listening-only, conversational) and audio bandwidth (e.g., narrowband versus wideband); as well as environmental factors (e.g., noisy versus quiet environments), to name a few [51]. The result of this comparison is a "quality event" that may be quantitatively described as a judgment of the "overall quality." Unfortunately, both the "perceptual event" and the "quality event" are internal to the perceiving human (denoted by the dotted line around the processes in Figure 1). To quantify salient attributes of these internal events, one has to rely on human test participants expressing their "subjective" judgments in terms of opinion scores. Most commonly, the mean opinion score (MOS) is used where the individual participants' scores are averaged to level out individual factors. As such, subjective methods are time-consuming, laborious, and expensive, thus prompting the development of instrumental or so-called "objective" speech quality estimation methods.

Instrumental models are used to estimate the average user judgment of the quality of a service. Commonly, though not necessarily, individual experiences and requirements are not taken into account by these models. Most models provide an estimate of the "overall quality" judged in a quiet listening-only or conversational context according to standardized test conditions [34], or with the consideration of background noise and its suppression [37]. Other models estimate multiple quality features such as coloration, noisiness, and continuity [61], [7]. The models (Figure 2) base their estimations

1) on signals that can either be measured at the electrical or the acoustic interfaces of the transmission channel of interest (signal based)



[FIG2] Types of instrumenal models. (a) Full-reference signal-based model; (b) reference-free signal-based model; (c) parametric model; (d) protocol-information-based and hybrid model.

2) on parameters that are estimated during the network planning phase (parametric)

3) on parameters collected at run time from network processes and control protocols

4) on a combination of 1) and 2) (hybrid models).

In the last scenario, the speech signal at the network output can be used alone or together with the network input speech signal, as depicted by the dashed line in Figure 2(d). Existing models can provide quality estimates for the classical narrowband (NB) (300–3,400 Hz), wide-band (WB) 50–7,000 Hz), or superwide-band (SWB) 50–14,000 Hz) signals. Table 1 gives an overview of models currently standardized or being discussed by relevant standardization bodies.

SIGNAL-BASED MODELS

Signal-based models employ speech signals transmitted or otherwise modified by speech processing systems to estimate quality. Most models provide quality estimations according to the Absolute Category Rating (ACR) listening quality scale defined in [34], but recently other models have been designed to predict individual quality features [52], [58], [62]. Two types of signal-based models exist: full-reference (also known as "intrusive" or "double-ended") models, which depend on a reference (system input) speech signal and a corresponding degraded (system output) speech signal; and reference-free ("nonintrusive" or "single-ended") models, which depend only on the latter degraded signal.

[TWO TYPES OF SIGNAL-BASED MODELS EXIST: FULL-REFERENCE AND REFERENCE-FREE MODELS.]

The idea of a full-reference model for predicting listening-only quality is simple—assuming that the aim is to transmit a speech signal over a channel without any perceptual degradation, then the perceptually weighted distance between the channel input and output signals should be indicative of the speech transmission quality. It is important that the distance be calculated on a perceptual level, as modern speech transmission channels do not aim at reproducing the exact signal, but only generate a similarly sounding signal at the output.

This underlying principle also points at some principal weaknesses of such models; because a comparison is made with respect to the input signal, this signal also has to reflect all the desired features of Figure 1 to correctly reflect the quality judgment process. Most models, however, predict the judgment made in an ACR test, and not in a paired-comparison paradigm. In an ACR test, the listener has no direct access to the reference input signal; the desired features are induced from the listener's experience by the test context, e.g., by the fact that the test contains only NB, WB, or SWB stimuli. The test context circumscribing the listener judgment is usually accounted for by selecting a model usage mode (NB, WB, or SWB) or separate model varieties for different test contexts.

Existing reference-based models comprise three components: 1) a preprocessing step including a level- and time-alignment of the two speech signals; 2) a perceptual transformation of the speech signals simulating parts of the peripheral human auditory system; and 3) an assessment unit that compares the two perceptually transformed signals. The most widespread full-reference model, the Perceptual Evaluation of Speech Quality (PESQ) [39] provides quality estimates for NB speech signals. It is based on its predecessor, the Perceptual Speech Quality Measure [(PSQM), formerly standardized in [38], but shows an improved performance for packet-switched networks by employing a better time-alignment algorithm and a different perceptual model]. To support burgeoning WB speech services, a WB extension of PESQ, called WB-PESQ [40], was standardized in 2005. However, this model has a limited scope, as it does not cover electroacoustic transducers, voice quality enhancement (VQE), and time-warping algorithms [39]. Therefore, Study Group 12 of the International Telecommunication Union (ITU) developed a new model called Perceptual Objective Listening Quality Assessment (POLQA) [24], which provides quality estimations in both NB and SWB contexts, and which is intended to cover the majority of existing telephone network scenarios.

The POLQA model provides quality estimation for fixed, mobile, and IP-based telephony services, including speech processing systems such as G.711.1, G.718, Skype SILK, Adaptive Multirate AMR-WB+, Advanced Audio Coding AAC LD, Enhanced Variable Rate Codecs (EVRC), and Continuous Variable Slope Delta Modulation (CVSD) codecs, as well as VQE algorithms (e.g., noise reduction, bandwidth extension, and automatic gain control). In its SWB mode, POLQA has a

[TABLE 1] TAXONOMY OF STANDARD OBJECTIVE SPEECH QUALITY PREDICTION MODELS.

QUALITY ASPECT	TYPE OF INPUT	INPUT	AUDIO BANDWIDTH	EXAMPLE
L-OQ	SIGNAL	M: 1 e	NB	P.563 [32], ANIQUE+[46]
		M: 2 e	NB	PSQM (P.861) [38], PESQ (P.862) [39]
			WB	WB-PESQ (P.862.2) [40]
		M: 2 e/a	NB/SWB	P.OLQA (P.863) [24]
	PARAM.	P	NB, WB	P.564 [33]
L-N	SIGNAL	M: 2e	NB	ETSI EG 202 396-3 [10]
L-M	SIGNAL	M: 2e/A	NB/SWB	P.AMD [25]
C-OQ	SIGNAL	M: 1e	NB	P.562 (CALL CLARITY INDEX [31], NONINTRUSIVE E-MODEL)
		M: 2e	NB	PESQM [2]
	PARAM.	E	NB	E-MODEL (G.107) [29]
			WB	WB E-MODEL (G.107) [29]
C-M	PARAM.	E	WB	DIMENSION-BASED WB E-MODEL [27]

Quality aspects: L = listening-only overall quality; L-N = listening-only for speech quality, noise quality, and overall quality; L-M = listening-only with several quality dimensions; C-OQ = conversational overall quality; C-M = conversational with several quality dimensions. Input: M = measurement; P = protocol information; E = offline measurement or estimation; 1 = one signal; 2 = two signals; a: acoustic signal, i.e., talker's speech signal measured by microphone(s); e: electric signal, i.e., talker's speech signal measured anywhere along the network transmission path. Audio bandwidth: NB = 300−3,400 Hz; WB = 50−7,000 Hz; SWB = 50−4,000 Hz. Exemplary models will be discussed in the subsequent text.

wider scope than WB-PESQ: it covers a wider bandwidth, from NB to SWB, and specific degradations such as frequency distortions introduced by user's terminal and nonoptimal listening levels. The assessment unit in POLQA uses an "idealized" signal, instead of the standard input reference signal for comparison; computes six different quality values; and combines them into an overall speech quality estimate. The estimate is computed in the so-called "cognitive" model that simulates high-level cognitive processes. Details on the standardized model can be found in [24].

Signal-based models for assessing the overall quality of a transmission channel provide an estimate of the quality level that can be reached with that particular channel; however, they do not provide insight into *why* a particular channel is good or bad. Thus, further information is necessary to diagnose the sources of poor quality. Such insight can be gained from models that predict multiple quality features. For a variety of modern transmission channels, Wältermann et al. [61] have uncovered three underlying orthogonal perceptual dimensions, i.e., discontinuity, coloration, and noisiness. Additionally, as recently documented in [7], the inclusion of a loudness dimension can also be useful in cases where nonoptimal (i.e., too high or too low) listening levels are present. Other dimensions pertaining to signal and/or background perceptual quality have been proposed in [60]. These dimensions were first determined via psychoacoustic experiments. Subsequently, multidimensional instrumental reference-based quality models were developed to estimate such percepts, such as the four-dimensional model recently documented by Côté [8], among others (e.g., [52], [58], and [59]). Moreover, on the basis of multiple computed quality features, it is possible to estimate the overall quality as a linear combination of the constituent features, as proposed in [62].

Estimation of multiple dimensions is also advantageous when characterizing the quality of noise-suppressed speech. Noise suppression algorithms can introduce unwanted artifacts to the speech signal, such as musical noise. In such situations, listeners can become confused as to which components of a noisy speech signal should form the basis of their ratings of overall quality. To reduce the error variance (or listener uncertainty) in the subjects' ratings of overall quality, the subjective test procedure recommended in ITU-T Rec. P.835 [37] instructs the listener to successively attend to and rate three different components of the noise suppressed speech signal: the speech signal alone (SMOS), background noise alone (NMOS), and the overall quality effect (GMOS). A full-reference model that estimates these three indices has been developed by the European Telecommunications Standardization Institute (ETSI) and is recommended in [10]. Besides an estimation of the speech quality that is based on a clean speech reference signal, the noise impact is estimated by means of the so-called "relative approach" [17], which emphasizes dynamic characteristics of the noise signal.

[REFERENCE-FREE MODELS HAVE GAINED MUCH ATTENTION RECENTLY AS REFERENCE SPEECH SIGNALS ARE NOT READILY AVAILABLE WITH IN-SERVICE NETWORKS.]

Reference-free models have gained much attention recently as reference speech signals are not readily available with in-service networks. Like the participants in ACR-type listening tests, reference-free models assess speech quality without the need to "listen" to a high-quality "clean" version of the target signal. Humans, through their experience, have acquired knowledge of normal and abnormal phenomena in speech sounds. Subjects in ACR listening tests rely only on this prior knowledge (desired features in Figure 1) to judge speech quality. In a similar vein, reference-free models utilize prior knowledge of normal and/or abnormal behavior of speech signal features to estimate speech quality. To represent prior knowledge, existing reference-free models employ models of speech production [19], speech perception [46], speech signal feature likelihood [14], [43], or a combination thereof [48].

In 2004, ITU-T held a competition to standardize a nonintrusive signal-based measure. Two algorithms stood out during this competition; one became the ITU standard P.563 [32] and the other, ANIQUE+, became an American National Standard Institute (ANSI) standard [1], [46]. While these models have been shown to be reliable for many telecommunications scenarios, recent research has suggested that their quality prediction performance is compromised for scenarios involving VQE algorithms (e.g., noise suppression [9] and dereverberation [16]) and wireless-VoIP tandem connections [12]. For both NB and WB reverberant and dereverberated speech, a no-reference quality model termed speech-to-reverberation modulation energy ratio (SRMR) has been recently proposed [16] and is available as open-source software for academic and research purposes. Directives on how to download the SRMR toolbox for MATLAB (Mathworks) can be obtained by contacting Tiago H. Falk. Table 2 summarizes application conditions in which the abovementioned standardized signal-based quality models have been recommended to be used and to be avoided.

The signal-based models presented so far aim at predicting speech quality or its features in a listening-only context. However, a primary goal of telecommunication speech services is to enable users to interact through conversations. Ease of conversation or interaction is however not measured in listening-only quality assessments, where the listeners do not interact with the speaker. In telephony conversations, interactivity is affected by the mouth-to-ear transmission delay. Large delays make it hard to interrupt the speaker, to promptly exchange turn for speaking, and for two simultaneous speakers to quickly return to a single-talker configuration. Also, the annoyance level of echoes [18] increases with delay. Echoes can be caused by reflections of the talker's speech across four-wire-two-wire circuit interfaces or across the acoustic interface (loudspeaker-microphone coupling).

APART FROM PARAMETRIC MODELS LIKE CCI AND THE E-MODEL, SIGNAL-BASED MODELS ALSO HAVE BEEN EXTENDED TO PROVIDE ESTIMATIONS OF CONVERSATIONAL QUALITY.

The ITU-T P.561 standard specifies requirements for in-service nonintrusive measurement devices (INMDs) to measure two-way speech transmission path parameters such as speech and noise levels, echo loss, and path delay [30]. A proprietary method called call clarity index (CCI), described in ITU-T P.562 Annex A, maps these measured parameters to an estimated conversational MOS value [31]. Within this paradigm, researchers have devised signal measurement algorithms to calculate the "planning" parameters in the parametric E-model (described in the section "Parametric Models") for the purpose of estimating listening or conversational quality.

Apart from parametric models like CCI and the E-model, signal-based models also have been extended to provide estimations of conversational quality. In a first step, Appel and Beerends [2] developed a model for talking-only speech quality, called Perceptual Echo and Sidetone Quality Measure (PESQM), which is based on PESQ. It takes into account the impact of degradations such as talker echo, coding artifacts, and additive noise on the talker's perception of his/her own voice. Guéguin et al. [20] proposed a signal-based model of conversational speech quality that combines both PESQ and PESQM with an estimation of the conversational impact of pure delay, the latter being derived from the E-model. None of these approaches has been standardized yet, but the definition and validation of a signal-based model for conversational speech quality is a work item in ITU-T Study Group 12.

PARAMETRIC MODELS

Signal-based models require speech signals as input to the quality estimation method. Thus, at least a prototype implementation or simulation of the transmission channel has to be set up. During the network design process, such signals are commonly not available; instead, the network is characterized by the technical specifications of its constituent elements. Such specifications include, amongst others, the frequency-weighted insertion loss (so-called "loudness rating") and the delay associated with a particular transmission path, the power of signal-correlated or uncorrelated noise inserted by the equipment, the probability that packets get lost or discarded in Internet-Protocol-(IP)-based transmission, as well as the type of speech codec and error concealment techniques used. Most of these specifications can be quantified in terms of planning parameters that enable parametric estimation of speech quality prior to the connection becoming live.

The E-model [44] can be treated as an archetypal parametric model used to estimate the quality associated with a speech transmission channel in a conversational context. Thus, in contrast to models that estimate listening quality, the E-model takes "two-way interaction effects" such as delay and echoes into account. The model features impairment factors that parametrically capture the different types of impairments in a telephone connection, covering the complete transmission chain from the mouth of the speaker to the ear of the listener.

The E-model impairments are grouped into four classes: 1) impairments affecting the basic signal-to-noise ratio of the transmission channel, such as ambient noise or circuit noise; 2) impairments occurring simultaneously with the speech signal, such as a nonoptimal sidetone level, or quantization distortions resulting from pulse-code modulation (PCM); 3) impairments occurring delayed with respect to the speech signal, such as talker and listener echoes, or the conversational impact of pure delay; and 4) impairments resulting from nonlinear and/or time-varying equipment, such as coding distortions or the effects of packet loss.

[TABLE 2] APPLICATION CONDITIONS IN WHICH EXISTING SIGNAL-BASED MODELS ARE RECOMMENDED TO BE USED AND TO BE AVOIDED.

MODEL	RECOMMENDED FOR	LIMITATIONS
PESQ	INPUT LEVELS, TRANSMISSION CHANNEL ERRORS, PACKET LOSS WITH OR WITHOUT CONCEALMENT, BIT-RATES, TRANSCODINGS, NOISE AT SENDING SIDE, TIME-VARYING DELAY, WAVEFORM CODECS, CELP CODECS, OTHER CODECS (CF. [39])	LISTENING LEVELS, LOUDNESS, HYBRID CODECS SUCH AS AMR AND EVRC, TIME-WARPING, NOISE REDUCTION, ECHO CANCELLATION
WB-PESQ	AS ABOVE, BUT WITH WB TRANSMISSION	AS ABOVE WITH WB TRANSMISSION, HYBRID CODECS SUCH AS AMR-WB, G.729.1 AND EVRC-WB (CF. [8])
POLQA	SAME AS PESQ ABOVE, BUT WITH SWB TRANSMISSION, VQE ALGORITHMS, SHORT TIME-WARPING ALGORITHMS, ELECTRO-ACOUSTIC TRANSDUCERS, HYBRID SPEECH CODECS	STRONG TIME-WARPING DISTORTIONS, EVRC CODECS
ETSI EG 202 396-3	NOISE REDUCTION ALGORITHMS, IN NB OR WB TRANSMISSIONS	
P.563	SAME AS PESQ ABOVE AND FOR SHORT- AND LONG-TERM TIME WARPING OF THE SPEECH SIGNAL (CF. [32])	TALKER ECHO, SIDETONES, LOW BITRATE (< 4KBPS) LPC VOCODER TECHNOLOGIES, SINGING VOICE. ALSO, APPLICATION SCENARIOS INVOLVING VQE ARTIFACTS, BITRATE MISMATCH BETWEEN ENCODER AND DECODER, AND AMPLITUDE CLIPPING WERE NOT FULLY VALIDATED DURING THE TIME OF THE STANDARDIZATION
ANIQUE+	SAME AS P.563 ABOVE	SAME AS P.563 ABOVE

In each class, the degradation is quantified in terms of a so-called "impairment factor" that is assumed to be additive on a perceptual scale. Thus, the overall transmission rating of the connection can be expressed by

$$R = Ro - Is - Id - Ie, eff + A, \tag{1}$$

where *Ro* is the basic signal-to-noise ratio of the channel, *Is* is the impairment factor related to the simultaneous degradations, *Id* is the impairment factor related to delayed degradations, and *Ie, eff* is the impairment factor related to nonlinear and time-varying degradations. *A* is called the advantage factor and reflects the quality expectation of the user. Depending on particular circumstances, such as mobile connections or connections to hard-to-reach areas, the user's quality expectation may differ from the norm; roughly speaking, *A* serves to adjust the desired features of Figure 1. The final transmission rating *R* (range: 0 = worst… 100 = best) can easily be transformed to a conversational MOS, which is the average rating on an overall quality scale collected in a conversation test carried out according to [34], following an S-shaped curve defined in [29].

Note that the transmission rating scale *R* can also be deployed for providing information on a purely perceptual level. In other words, the transmission system-based impairment factors described above can be replaced by multiple perceptual-dimension-related impairment factors, such as discontinuity, noisiness, and coloration [27].

Also, the E-model was originally developed for NB speech transmission. With an increase in wideband VoIP usage, the E-model framework has recently been updated to also provide valid predictions for WB transmission [54]. A complete WB-version of the E-model is expected at the end of the current ITU-T study period (2009–2012). In this case, the transmission rating scale is extended to the range 0…129 to reflect the potential quality advantage of WB transmission. Further extensions are currently being discussed for SWB transmission, where maximum values of $R = 179$ have been observed in auditory tests [64]. Due to its general nature and dispensation of signal input, the E-model is especially attractive for network planning. For this such purpose, the E-model is recommended by the ITU-T [29]. A list of recommended and not recommended application conditions of the NB and the WB version of the E-model is given in Table 3.

PROTOCOL-INFORMATION-BASED MODELS

The E-model has also been used for monitoring quality of VoIP in many studies [55], but it often does not provide accurate measurements for individual calls. As a consequence, alternative models based on protocol information as input have been developed. Instead of using the voice payload of the transmitted packets, the models exploit protocol header information such as the time stamps and sequence numbers from RTP [22] headers for delay and packet-loss-related information, and information on the end-point behavior such as dropped packets statistics or PLC information [23]. The main goal of such models is to enable passive network and/or end-point monitoring with a lightweight parametric approach, at the same time avoiding privacy concerns when accessing user-related payload information. The models can be employed at different locations in the service chain; by locating the model and measurement points in the client, the network, or both, solutions adaptable to different architectures are enabled (see the section "Practical Guidance" for more details). Examples of models of this type are described in [5] and [6]. Instead of standardizing an individual method, ITU-T recommends a procedure for validating NB- or WB-listening quality monitoring models by taking PESQ predictions as the reference; this procedure is described in [33]. An update of this procedure is expected with the new POLQA standard [24] for full reference quality prediction. One of the main differences to parametric planning models such as the E-model is that quality is followed and pooled over time, enabling more accurate predictions of per-call quality for the case of nonuniform packet loss. A comparison of parameter-based models employing different ways of temporal integration or pooling can be found, for example, in [53]. Table 4 summarizes the application areas of models that correspond to [33].

HYBRID APPROACHES

Modern communication scenarios can involve tandeming and internetworking of heterogeneous links, thus leading to impairment combinations that compromise the performance of existing quality measurement algorithms. For instance, in

[TABLE 3] APPLICATION CONDITIONS IN WHICH PARAMETRIC MODELS ARE RECOMMENDED TO BE USED AND TO BE AVOIDED.

MODEL	RECOMMENDED FOR	LIMITATIONS
E-MODEL	NB HANDSET TELEPHONY, INCLUDING THE EFFECTS OF OVERALL LOUDNESS, FREQUENCY DISTORTION, QUANTIZING DISTORTION, CODING, BACKGROUND NOISE, CIRCUIT NOISE, NONOPTIMUM SIDETONE LEVEL, TALKER AND LISTENER ECHO, PURE DELAY, RANDOM AND BURSTY PACKET LOSS	DIFFERENT LISTENING LEVELS, NONHANDSET TERMINALS INCLUDING NOISE REDUCTION AND ECHO CANCELLATION [50]
WB E-MODEL	WB TELEPHONY WITH HANDSET AND HEADPHONE LISTENING, INCLUDING CODING DISTORTIONS AND RANDOM PACKET LOSS [29]; FIRST EXTENSIONS PROPOSED FOR OVERALL LOUDNESS, FREQUENCY DISTORTION, AND SEND SIDE NOISE	NONOPTIMUM SIDETONE LEVEL, TALKER AND LISTENER ECHO, PURE DELAY, OTHER TERMINALS INCLUDING NOISE REDUCTION AND ECHO CANCELLATION, BANDWIDTH EXTENSION ALGORITHMS

[TABLE 4] APPLICATION CONDITIONS IN PROTOCOL-INFORMATION-BASED MODELS ARE RECOMMENDED TO BE USED AND TO BE AVOIDED.

MODEL	RECOMMENDED FOR	LIMITATIONS
P.564	APPLICATION RANGE DETERMINED BY INPUT INFORMATION AVAILABLE AT MEASUREMENT POINT AND APPLICATION RANGE OF PESQ	PREDICTIONS FOR EFFECTS NOT ADDRESSED BY PESQ, SUCH AS ECHO OR DELAY, CANNOT BE VALIDATED USING [33]; ECHO AND OTHER INFORMATION AS AVAILABLE E.G., FROM THE CLIENT [23] MUST BE MEANINGFUL

wireless VoIP tandem communications, standard signal-based models such as PESQ [39] and P.563 [32] were shown to be sensitive to varying packet loss rates and PLC strategies [12], [13]. Parametric models, in turn, were shown to be sensitive to acoustic background noise combined with PLC artifacts, as well as noise suppression artifacts combined with speech codec distortions [13].

Hybrid approaches, which can make use of both the signal decoded from the payload and IP connection parameters extracted from protocol header information, have been proposed to overcome the limitations of pure signal and pure parametric/protocol-based approaches. The model developed in [13], for example, made use of IP connection parameters such as codec and PLC type, packet size, and packet loss pattern to determine a "base quality" representative of the transmission link under test. Distortions that were not captured by the connection parameters, such as the acoustic noise type and level, temporal clippings, and PLC and noise suppression artifacts, were computed from perceptual features extracted from the decoded speech signal and used to adjust the base quality accordingly.

Alternately, hybrid approaches can also be taken when input parameters (e.g., the codec algorithm used to generate the speech payload or other codec-related impairment factors) of a parametric model are unobserved or otherwise unavailable. For example, if the E-model is to be used in conjunction with a new type of codec, a corresponding effective equipment impairment factor *Ie, eff* has to be derived. For this purpose, the ITU-T recommends either to carry out listening-only tests, or to rely on signal-based full-reference models. A full-reference model such as PESQ is able to estimate adequate *Ie, eff* values, provided that the estimations are normalized for biases resulting from the test stimuli; the corresponding normalization procedures are defined in [35] for NB and in [36] for WB speech transmission.

In general, hybrid approaches entail fusing diverse information that is more readily or economically accessible, reliable, or timely in specific speech quality estimation applications. For instance, in speech transmission over tandem links, measuring the degradations at each link endpoint and distributing the measurements across network nodes along the transmission path provides more accurate estimates of speech quality than relying only on information at transmission endpoints. Such distributed measurement also provides diagnostics that would enable isolating sources of degradation to specific network elements. With evolving wireline and wireless networks, opportunities abound for embedding into network elements functionalities for distributed quality monitoring, diagnosis, remedy, and assurance.

PRACTICAL GUIDANCE

Given the multitude of available models, it is not always apparent to the practitioner which model to apply for a given purpose. To further complicate the situation, users are often faced with multiple models that are deemed applicable to a specific application. In Table 5, we compile a list of currently recommended or in-use approaches for speech quality prediction, including emerging standards which are still under discussion. It is hoped that the information provided in the table and documented below will assist users and/or researchers in selecting an appropriate model for their specific application.

For network planning purposes, the only currently recommended model is the E-model. Its NB version covers all standard network elements as well as standard handset terminals; first extensions for nonhandset terminals including signal-processing equipment such as noise reduction and echo cancellation have been presented [45], [50], but they are not yet conclusive. For WB, the underlying scale has been extended and *Ie,eff* values are provided in [29]; other types of degradations have been dealt with in [54], but they are not yet recommended by the ITU-T. For SWB, only a transmission rating scale extension has been proposed [64]. First proposals have also been made to predict individual quality features with a perception-based E-model, but they are not yet conclusive [27]. Users should exercise care when considering the nonconcluded elements in their applications and/or research.

For network optimization and maintenance, the upcoming recommended model is the reference-based P.OLQA, published as ITU-T Rec. P.863 in early 2011 [24]. It comes in NB and SWB modes, the latter also covering WB scenarios and acoustic recording conditions. P.863 will, perhaps gradually, replace NB [39] and WB P.862 PESQ [40], which are still widely used in field measurement equipment as well as in the laboratory. P.863 has shown significantly better performance than P.862 on a large variety of databases and a wider range of usage scenarios [26]. Current work in ITU-T SG12 is focusing on the characterization of the new model, as well as on the development of a multidimensional model Perceptual Approaches for Multidimensional (P.AMD) analysis [25], which is able to estimate four to seven quality features, describing discontinuity, [TABLE 5] APPLICATION SCENARIOS OF SPEECH QUALITY PREDICTION MODELS.

APPLICATION	TYPE OF INPUT	SOURCE OF INPUT	TARGET QUALITY PREDICTION	TARGET AUDIO BANDWIDTH	CONSIDERED CHANNEL ELEMENTS AND SCOPE	RECOMMENDED MODEL
PLANNING	PARAMETERS	ESTIMATION	MOS-CQEN	NB	ALL MOUTH-TO-EAR	G.107 [29]
	PARAMETERS	ESTIMATION	MOS-CQEM	WB	ALL MOUTH-TO-EAR	G.107 [29] + EXTENSIONS [54] (SEE NOTE 1 BELOW)
	PARAMETERS	ESTIMATION	MOS-CQEM	SWB	ALL MOUTH-TO-EAR	G.107 [29]+ EXTENSIONS [64] (SEE NOTE 1 BELOW)
	PARAMETERS	ESTIMATION	QUALITY FEATURES	SWB	ALL MOUTH-TO-EAR	G.107 [29] + EXTENSIONS [27] (SEE NOTE 1 BELOW)
OPTIMIZATION	2e SIGNALS	SIMULATION	MOS-LQON	NB	CODECS, TRANSMISSION ERRORS, NOISES AT THE SENDING SIDE	P.862 [39] (SEE NOTE 2 BELOW)
	2e OR 2a SIGNALS	SIMULATION	MOS-LQON	NB	CODECS, TRANSMISSION ERRORS, TERMINALS, TIME-WARPING, ETC.	P.863 [24]
	2e SIGNALS	SIMULATION	MOS-LQOW	WB	SAME AS P.862	P.862.2 [40] (SEE NOTE 2 BELOW)
	2e OR 2a SIGNALS	SIMULATION	MOS-LQOM	SWB	CODECS, TRANSMISSION ERRORS, TERMINALS, TIME-WARPING, NOISES AT THE SENDING SIDE, LISTENING LEVELS	P.863 [24]
	3e SIGNALS	SIMULATION	SMOS, NMOS, GMOS	NB	BACKGROUND NOISE, NOISE REDUCTION	EG 202 396-3 [10]
	2e OR 2a SIGNALS	SIMULATION	FOUR . . . SIX QUALITY FEATURES (COLORATION, NOISINESS, DISCONTINUITY, LOUDNESS)	NB	SAME AS P.863	P.AMD [25]
	2e OR 2a SIGNALS	SIMULATION	FOUR . . . SIX QUALITY FEATURES (COLORATION, NOISINESS, DISCONTINUITY, LOUDNESS)	SWB	SAME AS P.863	P.AMD [25]
MONITORING	1e SIGNAL	MEASUREMENT	MOS-LQON	NB	SAME AS P.862	P.563 [32], ANIQUE+ [1]
MAINTENANCE	2e SIGNALS	MEASUREMENT	MOS-LQOW	WB	SAME AS P.862	P.862 [39] (SEE NOTE 2 BELOW)
	2e OR 2a SIGNALS	MEASUREMENT	MOS-LQOM	NB, WB (SEE NOTE 3 BELOW), SWB	SAME AS P.863	P.863 [24]
	1e SIGNAL	MEASUREMENT	MOS-CQON	NB	SAME AS G.107	P.562 (CCI) [31]
	2e SIGNALS	MEASUREMENT	MOS-CQEM	NB, WB, SWB	SAME AS G.107	G.107 (NONINTRUSIVE)
	PARAMETERS	MEASUREMENT	MOS-LQON	NB	IP NETWORK IMPAIRMENTS ON THE ONE-WAY LISTENING QUALITY	P.564 [33]
	PARAMETERS	MEASUREMENT	MOS-LQOW	WB	IP NETWORK IMPAIRMENTS ON THE ONE-WAY LISTENING QUALITY	P.564 ANNEX B [33]
	2e SIGNALS	MEASUREMENT	MOS-CQO	NB	ECHO, SIDETONE	PESQM [2]

Type of input: 1 = one signal; 2 = two signals; a = acoustic signal or recording; e = electrical signal. Target: MOS = mean opinion score; CQ = conversational quality; LQ = listening quality; E = estimated during network planning; O = objective using signal-based measures; N = narrow-band; M = mixed narrow-, wide- and/or superwide-band; W = wide-band. Audio bandwidth: NB = 300–3,400 Hz; WB = 50–7,000 Hz; SWB = 50–14,000 Hz. Notes: 1) Full WB and SWB models covering all channel elements not yet available; 2) ITU-T Rec. P.862 and related models are no longer recommended by ITU-T SG12, and should be replaced by ITU-T Rec. P.863; and 3) no separate WB mode available; to be used in its SWB mode.

coloration, noisiness, and nonoptimum loudness, and more fine-grained features such as low- and high-frequency coloration, slow- and fast-varying time-localized distortions as well as the level and variation of background noise [41].

For network monitoring, a variety of approaches is conceivable; thus model usage should be guided based on the number of available signals and/or if one-way or two-way communication is sought. For one-way communications, (i.e., listening-only) two methods may be employed. First, if only a single electrical signal is available during network operation, reference-free models such as [32] and [1] should be used (the so-called nonintrusive measurement). Second, if two electrical or acoustically recorded signals are available, reference-based models such as [39] and [24] should be considered. Note that to perform reference-based quality measurement, a clean reference signal needs to be injected into a piece of equipment or a call connection, thus temporarily disrupting the network or call session. For two-way communications, or conversational quality measurement, users have four options. Within a reference-based paradigm, practitioners may employ the clean and processed signals to estimate impairment factors that can be forwarded to parametric models, as is recommended by ITU-T Rec. P.562 [31]. Alternately, users may opt to use the nonintrusive version of the E-model. Third, parameters collected from packet headers can be used for parametric estimation using approaches that can be validated according to [33] (see the section "Protocol-Information-Based Models"). While the abovementioned methods fall within the parametric or hybrid paradigms, pure signal-based models may also be applied to predict conversational quality; one such approach, though not recommended by ITU, is described in [2].

FUTURE TRENDS AND CHALLENGES

Full-reference signal-based models were widely developed in the 1990s, and the last decade witnessed a widespread domination of the PESQ model, mostly due to its high correlation with subjective quality scores. Needless to say, speech transmission systems have evolved over the last decade, employing more complex signal processing algorithms, such as speech enhancement. As expected, PESQ and WB-PESQ performance did not follow suit, leading to the development of its successor—ITU-T Rec. P.863 (POLQA). While the new POLQA model [24] is intended to cover all relevant in-use transmission scenarios and equipment, the model is still restricted to short sentences of approximately 6–20 s, far below the duration of a typical phone call, which ranges between one and two minutes. Nevertheless, a major trend is the development of models that predict "full-length" call quality. The first steps have already been taken [3], [65], and a standard has been put forward [11], though not yet covering WB or SWB transmission. Clearly, further research is still needed before a general-purpose tool is available. One possibility is to investigate temporal-integration strategies that combine multiple short-duration P.863 estimations into a final longer-duration call quality rating. Similar considerations apply for no-reference methods including protocol-header-based and hybrid approaches. Some of these models can already account for longer observation windows but need to be adapted to WB and SWB quality prediction. Here, it will be desirable to achieve equally stable and well-validated models as P.863.

Another major trend pursued by the speech quality measurement community has been the development of reliable multidimensional quality models for enhanced speech (e.g., noise suppressed, bandwidth extension, and dereverberated speech), where unwanted perceptual artifacts, residual noise, and signal-component distortions need to be detected and quantified [16]. The same holds true for the variety of other signal processing equipment used in today's networks and terminal elements (e.g., level-adjustment and echo cancellation); such equipment can be best characterized with dimension-based approaches such as the one described in [25]. Multidimensional quality models can also play a pivotal role in characterizing human-machine communication, such as diagnosing the quality of synthesized speech and of spoken dialogue systems [15], [49]. Quality dimensions of interest can include listening effort, articulation, naturalness, continuity/fluency, and prosody similarity with natural speech, to name a few.

[QUALITY DIMENSIONS OF INTEREST CAN INCLUDE LISTENING EFFORT, ARTICULATION, NATURALNESS, CONTINUITY/FLUENCY, AND PROSODY SIMILARITY WITH NATURAL SPEECH, TO NAME A FEW.]

Atypical speech constitutes a range of novel conditions for most existing standard speech quality models. These models are mostly tested using the speech and listening of healthy adult native speakers of a few selected languages. Model performance may be poor for speech and hearing capabilities that are not represented in the test conditions.

While speech "quality" serves as an essential performance measure in typical telecommunication conversation settings, other measures provide more specifically useful information. Currently, objective measurement of speech "intelligibility" is actively researched. Commonly, "intelligibility" measures the percentage of words or subword units understandable to native speakers with healthy hearing and cognition. Speech intelligibility is a component of speech quality, in the sense that good intelligibility is necessary but not sufficient for good quality. For instance, robotic speech may be perfectly intelligible but not natural and hence not high quality. Speech intelligibility measures are particularly useful in degraded conditions, such as noisy and reverberated speech, talkers with speech impairments, and listeners with hearing impairments. For instance, Hu and Loizou [21] reported that most existing noisy-speech enhancement algorithms improve speech quality but hardly improve speech intelligibility. In some task-oriented, mission-critical applications, it might be preferable to configure speech enhancement processing to maximize speech intelligibility, perhaps at the expense of weighing less other attributes of speech quality such as naturalness. In human-machine communication wherein speech reception is by a machine, "quality" may be measured to enable maximization of machine "intelligence." While a measure of intelligibility may be appropriate for automatic speech recognition, the measure might not be adequate for speaker identification.

Finally, speech and audio quality are an integral part of the user's perception of multimedia quality. Future research directions will strive to develop models that combine audio and video quality information and which better reflect human interaction behavior. Such models may also be able to cover surround sound and three-dimensional video with stereoscopic rendering techniques. It is also possible that future models will incorporate adaptable user-specific parameters to truly represent a user's quality of experience (QoE). A major challenge witnessed today is the lack of publicly available data to develop and test such models. Devising models and/or model design methods that have low subjective-data costs is an ongoing challenge. Crowdsourcing approaches (e.g., Amazon's Mechanical Turk) may contribute to the solution of this challenge [56].

AUTHORS

Sebastian Möller (sebastian.moeller@telekom.de) studied electrical engineering at the universities of Bochum (Germany), Orléans (France), and Bologna (Italy). He received a doctor-of-engineering degree in 1999 and the venia legendi with a book on the quality of telephone-based spoken dialogue systems in 2004. In 2005, he joined Deutsche Telekom Laboratories, TU Berlin, and in 2007, he was appointed professor for quality and usability at Technical University (TU) Berlin. His primary interests are in speech signal processing, speech technology, and quality and usability evaluation. Since 1997, he has taken part in ITU–T Study Group 12, where he is currently corapporteur of Question Q.8/12.

Wai-Yip Chan (geoffrey.chan@queensu.ca), also known as Geoffrey Chan, received his B.Eng. and M.Eng. degrees from Carleton University, Ottawa, and his Ph.D. degree from the University of California, Santa Barbara. He is currently with the Department of Electrical and Computer Engineering, Queen's University, Canada. He has held positions with the Communications Research Centre, Bell Northern Research (Nortel), McGill University, and Illinois Institute of Technology. His research interests are in multimedia signal processing and communications. He is an associate editor of *EURASIP Journal on Audio, Speech, and Music Processing*, and a member of the IEEE Signal Processing Society Speech and Language Technical Committee.

Nicolas Côté (nicote@free.fr) studied audiovisual engineering at the University of Valenciennes, France. He received a master's degree in acoustic and signal processing applied to music signals from the University of Paris VI in 2005. He joined France Télécom R&D in Lannion, France; Deutsche Telekom Laboratories, TU Berlin, Germany, in 2005; and received a doctor-of-engineering degree in 2010 for his work on the integral and diagnostic intrusive prediction of speech quality. He now works as a scientific researcher at the Université de Bretagne Occidentale, in Brest, France. His research interests include quality assessment of speech transmission and sound reproduction systems.

Tiago H. Falk (tiago.falk@ieee.org) received the B.Sc. degree from the Federal University of Pernambuco, Brazil, in 2002, and the M.Sc. and Ph.D. degrees from Queen's University, Canada, in 2005 and 2008, respectively, all in electrical engineering. Since 2010, he has been an assistant professor at the Institut National de la Recherche Scientifique (INRS-EMT) in Montreal, Canada. His research interests are in multimedia quality measurement and enhancement. His work has engendered numerous awards, including the IEEE (Kingston Section) Ph.D. Research Excellence Award, Best Student Paper Awards at ICASSP (2005) and IWAENC (2008) conferences, and the Newton Maia Young Scientist Award.

Alexander Raake (alexander.raake@telekom.de) received his doctoral degree in electrical engineering and information technology from Ruhr-University Bochum, Germany, in 2005, and his electrical engineering diploma from RWTH Aachen, Germany, in 1997. From 1998 to 1999 he was a researcher at EPFL, Switzerland. Between 2004 and 2009 he held postdoc and senior scientist positions at LIMSI-CNRS, France, and Deutsche Telekom Laboratories, Germany, respectively. Since 2009, he has been an assistant professor at Deutsche Telekom Laboratories, TU Berlin. His research interests are in multimedia technology and QoE. Since 1999, he has been active in ITU-T, currently as corapporteur for Q.14/12 on audiovisual quality.

Marcel Wältermann (marcel.waeltermann@telekom.de) studied electrical engineering at Ruhr-University Bochum, Germany. In 2005, he graduated in the area of communication acoustics. After a two-year engagement at Ruhr-University Bochum, he now works as a scientific researcher at Deutsche Telekom Laboratories, TU Berlin, on quality models for transmitted speech on the basis of perceptual dimensions. Further interests include communication acoustics and speech signal processing. He has been corapporteur for Question 8/12 of ITU–T Study Group 12 since 2009.

REFERENCES

[1] *Auditory Non-Intrusive Quality Estimation Plus (Anique+): Perceptual Model for Non-Intrusive Estimation of Narrowband Speech Quality, ATIS-PP-0100005.2006, American National Standards Institute*, 2006.

[2] R. Appel and J. G. Beerends, "On the quality of hearing one's own voice," *J. Audio Eng. Soc.*, vol. 50, no. 4, pp. 237–248, 2002.

[3] J. Berger, A. Hellenbart, R. Ullmann, B. Weiss, S. Möller, J. Gustafsson, and G. Heikkilä, "Estimation of 'quality per call' in modelled telephone conversations," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'08)*, Las Vegas, NV, 2008, pp. 4809–4812.

[4] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA: MIT Press, 1997.

[5] S. Broom, "VoIP quality assessment: Taking account of the edge-device," *IEEE Trans. Audio, Speech Lang. Processing (Special Issue on Objective Quality Assessment of Speech and Audio)*, vol. 14, no. 6, pp. 1977–1983, Nov. 2006.

[6] A. Clark, "Modeling the effects of burst packet loss and recency on subjective voice quality", in *Proc. Internet Telephony Workshop (IPtel'01)*, New York, Apr. 2001, pp. 1–5.

[7] N. Côté, V. Koehl, V. Gautier-Turbin, A. Raake, and S. Möller "An intrusive super-wideband speech quality model: DIAL," in *Proc. 11th Annu. Conf. Int. Speech Communication Association (Interspeech'10)*, Makuhari, Japan, 2010, pp. 1317–1320.

[8] N. Côté, *Integral and Diagnostic Intrusive Prediction of Speech Quality. Berlin*: Springer-Verlag, 2011.

[9] A. Ekman and B. Kleijn, "Improving quality prediction accuracy of P.563 for noise suppression," in *Proc. Int. Workshop for Acoustic Echo and Noise Control, CD_ROM*, 2008.

[10] "Speech processing, transmission and quality aspects (STQ); speech quality performance in the presence of background noise. Part 3: Background noise transmission—Objective test methods," *Europ. Telecomm. Standardization Institute,* Sophia Antipolis, France, *ETSI EG 202 396-3,* 2008.

[11] "Speech processing, transmission and quality aspects (STQ); estimating speech quality per call," *Europ. Telecomm. Standardization Institute,* Sophia Antipolis, France, *ETSI TR 102 506*, 2007.

[12] T. H. Falk and W.-Y. Chan, "Performance study of objective speech quality measurement for modern wireless—VoIP communications," *EURASIP J. Audio, Speech Music Processing*, vol. 2009, Article ID 104382, 11 pages.

[13] T. H. Falk and W.-Y. Chan, "Hybrid signal-and-link-parametric speech quality measurement for VoIP communications," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 16, no. 8, pp. 1579–1589, Nov. 2008.

[14] T. H. Falk and W.-Y. Chan, "Nonintrusive speech quality estimation using Gaussian mixture models," *IEEE Signal Processing Lett.*, vol. 13, no. 2, pp. 108–111, Feb. 2006.

[15] T. H. Falk and S. Möller, "Towards signal-based instrumental quality diagnosis for text-to-speech systems," *IEEE Signal Processing Lett.*, vol. 15, pp. 781–784, 2008.

[16] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech Lang. Processing (Special Issue on Processing Reverberant Speech: Methodologies and Applications)*, vol. 18, no. 7, pp. 1766–1774, Sept. 2010.

[17] K. Genuit, "Objective evaluation of acoustic-quality based on a relative approach," in *Proc. Inter-Noise* 1996, Liverpool, England, paper 1061, pp. 1–6.

[18] J. D. Gibson, *The Communications Handbook.* Boca Raton, FL: CRC, 2002.

[19] P. Gray, M. P. Hollier, and R. E. Massara, "Non-intrusive speech quality assessment using vocal-tract models," *Proc. Inst. Elect. Eng. Vision, Image, Signal Processing*, vol. 147, no. 6, pp. 493–501, Dec. 2000.

[20] M. Guéguin, R. Le Bouquin-Jeannès, V. Gautier-Turbin, G. Faucon, and V. Barriac, "On the evaluation of the conversational speech quality in telecommunications," *EURASIP J. Adv. Signal Processing*, vol. 2008, 2008, Article ID 185248, 15 pages.

[21] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Commun.*, vol. 49, no. 7, pp. 588–601, July 2007.

[22] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, Eds., "RTP: A transport protocol for real-time applications," *IETF RFC 3550*, Internet Engineering Task Force, Freemont, CA, July 2003.

[23] T. Friedman, R. Caceres, iand A. Clark, Eds., "RTCP extended report (XR)," *IETF RFC 3611*, Internet Engineering Task Force, Freemont, CA, Nov. 2003.

[24] ITU, "Perceptual objective listening quality assessment," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.863, 2011.

[25] ITU, "Draft requirement specification for P.AMD (perceptual approaches for multi-dimensional analysis)," Source: Deutsche Telekom AG, ITU-T SG12 WP2 Meeting, 17 Sept. 2010, Berlin, Int. Telecomm. Union, Geneva, Switzerland, ITU-T Contr. COM 12-143, 2010.

[26] ITU, "Performance of the joint POLQA model," Source: Opticom, TNO, SwissQual, ITU-T SG12 WP2 Meeting, Sept. 17, 2010, Berlin, Int. Telecomm. Union, Geneva, Switzerland, ITU-T Contr. COM 12-148, 2010.

[27] ITU, "Perceptual correlates of the E-Model's impairment factors," Source: Federal Republic of Germany (Authors: M. Wältermann and S. Möller), ITU-T SG12 Meeting, Oct. 17–21, Int. Telecomm. Union, Geneva, Switzerland, ITU-T Delayed Contribution D.071, 2005.

[28] ITU, "Handbook on telephonometry," Int. Telecomm. Union, Geneva, Switzerland, ITU-T, 1992.

[29] ITU, "The E-model: A computational model for use in transmission planning," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. G.107, 2009.

[30] ITU, "In-service non-intrusive measurement device—Voice service measurements," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.561, 2002.

[31] ITU, "Analysis and interpretation of INMD voice-service measurements," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.562, 2004.

[32] ITU, "Single-ended method for objective speech quality assessment in narrow-band telephony applications," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.563, 2004.

[33] ITU, "Conformance testing for voice over ip transmission quality assessment models," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.564, 2007.

[34] ITU, "Methods for subjective determination of transmission quality", Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.800, 1996.

[35] ITU, "Methodology for the derivation of equipment impairment factors from instrumental models," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.834, 2002.

[36] ITU, "Extension of the methodology for the derivation of equipment impairment factors from instrumental models for wideband speech codecs," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.834.1, 2009.

[37] ITU, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.835, 2003.

[38] ITU, "Objective quality measurement of telephone-band (300–3400 Hz) speech codecs," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.861, 1996.

[39] ITU, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.862, 2001.

[40] ITU, "Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.862.2, 2005.

[41] ITU, "Multidimensional subjective testing methodology," Source: Rapporteurs of Q.7/12, ITU-T SG12 Meeting, Geneva, Switzerland, Jan. 18–27, 2011, Int. Telecomm. Union, Geneva, Switzerland, ITU-T TD 367(GEN), 2011.

[42] U. Jekosch, *Voice and Speech Quality Perception: Assessment and Evaluation*. New York: Springer-Verlag, 2005.

[43] C. Jin and R. Kubichek, "Vector quantization techniques for output-based objective speech quality," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, May 1996, vol. 1, pp. 491–494.

[44] N. O. Johannesson, "The ETSI computation model: A tool for transmission planning of telephone networks," *IEEE Commun. Mag.*, vol. 35, no. 1, pp. 70–79, Jan. 1997.

[45] N. O. Johannesson, "Echo canceller performance characterized by impairment factors," *ITU-T Speech Quality Experts Group Meeting,* Ipswich, U.K., Sept. 23–27, Int. Telecomm. Union, Geneva, Switzerland, *Doc. IP–16*, 1996.

[46] D.-S. Kim and A. Tarraf, "ANIQUE+: A new American national standard for non-intrusive estimation of narrowband speech quality," *Bell Labs Tech. J.*, vol. 12, no. 1, pp. 221–236, May 2007.

[47] J. Liang and R. Kubichek, "Output-based objective speech quality," in *Proc. IEEE Vehicular Technol. Conf.,* Stockholm, Sweden, pp. 1719–1723, 1994.

[48] L. Malfait, J. Berger, and M. Kastner, "P.563—The ITU-T standard for single-ended speech quality assessment," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 14, no. 6, pp. 1924–1934, 2006.

[49] S. Möller, F. Hinterleitner, T. H. Falk, and T. Polzehl, "Comparison of approaches for instrumentally predicting the quality of text-to-speech systems," in *Proc. 11th Annu. Conf. Int. Speech Communication Association (Interspeech'10)*, Makuhari, Japan, Sept. 26–30, 2010, pp. 1325–1328.

[50] S. Möller, F. Kettler, H.-W. Gierlich, N. Côté, A. Raake, and M. Wältermann, "Extending the E-model to better capture terminal effects," in *Proc. 3rd Int. Workshop on Perceptual Quality of Systems (PQS'10)*, Bautzen, Germany, 2010.

[51] S. Möller, *Assessment and Prediction of Speech Quality in Telecommunications.* Norwell, MA: Kluwer, 2000.

[52] S. R. Quackenbush, T. P. Barnwell, and M. A. Clemens, *Objective Measures of Speech Quality.* Englewood Cliffs, NJ: Prentice-Hall, 1988.

[53] A. Raake, "Short- and long-term packet loss behavior: Towards speech quality prediction for arbitrary loss distributions," *IEEE Trans. Audio, Speech Lang. Processing (Special Issue on Objective Quality Assessment of Speech and Audio)*, vol. 14, no. 6, pp. 1957–1968, Nov. 2006.

[54] A. Raake, S. Möller, M. Wältermann, N. Côté, and J.-P. Ramirez, "Parameter-based prediction of speech quality in listening context—Towards a WB E-model," in *Proc. 2nd Int. Workshop Quality of Multimedia Experience (QoMEX'10)*, June 21–23, 2010, pp. 182–187.

[55] A. Raake, *Speech Quality of VoIP—Assessment and Prediction.* Chichester, West Sussex, U.K.: Wiley, 2006.

[56] F. Ribeiro, D. Florencio, C. Zhang, and M. Seltzer, "CROWDMOS: An approach for crowdsourcing mean opinion score studies," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'11)*, Prague, Czech Republic, 2011, pp. 2416–2419.

[57] D. L. Richards, *Telecommunications by Speech.* London: Butterworths, 1973.

[58] K. Scholz, "Instrumentelle Qualitätsbeurteilung von Telefonbandsprache beruhend auf Qualitätsattributen (Instrumental quality assessment of telephone-band speech based on quality attributes)," Doctoral Dissertation *(Arbeiten über Digitale Signalverarbeitung, no. 32)*. Aachen, Germany: Shaker Verlag, 2008.

[59] D. Sen, "Predicting foreground SH, SL and BNH DAM scores for multidimensional objective measure of speech quality," in *Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing (ICASSP)*, Montreal, May 2004, vol. 1, pp. 493–496.

[60] W. D. Voiers, "Diagnostic acceptability measure for speech communication systems," in *Proc. ICASSP'77,* Hartford, CT, 1977, pp. 204–207.

[61] M. Wältermann, A. Raake, and S. Möller, "Quality dimensions of narrowband and wideband speech transmission," *Acta Acust. United Acust.*, vol. 96, no. 6, pp. 1090–1103, 2010.

[62] M. Wältermann, K. Scholz, S. Möller, L. Huo, A. Raake, and U. Heute, "An instrumental measure for end-to-end speech transmission quality based on perceptual dimensions: Framework and realization," in *Proc. Interspeech* 2008, pp. 22–26.

[63] M. Wältermann, I. Tucker, A. Raake, and S. Möller, "Analytical assessment and distance modeling of speech transmission quality," in *Proc. 11th Ann. Conf. Int. Speech Communication Association (Interspeech'10)*, Makuhari, Japan, 2010, pp. 1313–1316.

[64] M. Wältermann, A. Raake, and S. Möller, "Extension of the E-model towards super-wideband speech transmission," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'10),* Dallas, TX, 2010, pp. 4654–4657.

[65] B. Weiss, S. Möller, A. Raake, J. Berger, and R. Ullmann, "Modeling conversational quality for time-varying transmission characteristics," *Acta Acoust. United Acoust.*, vol. 95, no. 6, pp. 1140–1151, 2008.