

Chapter 12

Speech Communication

Nicolas Côté and Jens Berger

Abstract The goal of any speech service is the transmission and/or processing of speech signals. In this chapter we discuss the Quality of Experience (QoE) of speech communication systems, including networks, speech processing applications and terminals. We then give an overview of the methods employed to quantify and further estimate the QoE of speech communication services with a focus on diagnostic instrumental models. Such models provide indications on either the technical causes of degradations or the quality features impacted by a component in the speech communication system.

12.1 General Overview of Speech Communication

12.1.1 *Quality of Experience in the Context of Speech Communication*

As defined by Hardy [7], a voice [speech] service corresponds to a voice interaction through a telecommunication system.¹ Two types of speech services exist, namely (1) speech communication services, which imply a conversation between a talker and a listener (or several listeners in case of teleconferencing systems) in a near

¹ In the literature, the terms “voice service” and “speech service” are mostly used interchangeably. Here, we will refer to “voice” when the characteristics of the human voice are addressed, and to “speech” when both the signal carrier and the referred content are of interest.

N. Côté (✉)
Institute of Electronics, Microelectronics and Nanotechnology, ISEN Department,
Lille, France
e-mail: nicolas.cote@isen.fr

J. Berger
SwissQual AG, a Rohde & Schwarz Company, Zuchwil, Switzerland
e-mail: jens.berger@swissqual.com

“real-time” manner, and (2) streaming services (e.g. recorded messages stored on a device). These services replace the air path between two interlocutors having a face-to-face conversation. Since the success of any service depends on its QoE, the quality assessment of the corresponding speech communication system or speech processing application is required for both the developers and the telecommunication providers.

12.1.2 Factors of Speech Communication QoE

Even if the quality of the transmitted speech is a factor determining the QoE of speech communication systems, user’s satisfaction encloses many different aspects. According to the theoretical framework of QoE introduced in Chap. 4, the physical factors influencing the QoE are grouped into three categories: human influence factors, context influence factors and the system influence factors. The “human influence factors” here correspond to the talker’s difficulties to produce an acoustic message (e.g. aphonia) and the listener’s difficulties to understand this message (e.g. hearing impairments). Since humans can use speech services in very diverse situations, especially with the massive introduction of mobile terminals, the last category, “context influence factors”, covers many heterogeneous environments (in terms of time and place). The “system influence factors” include all technical characteristics, physical equipment and computer programs, of the speech service. Section 12.2 describes mainly both subcategories, “network related system” and “device-related system” of the more general system influence factor category.

12.1.3 Features of Speech Communication QoE

The perceived quality of telephone systems has been studied for many decades [5, 9, 28, 40]. In these studies, auditory tests have been carried out where subjects had to judge the perceived quality of transmitted speech. It resulted from these studies that speech quality, like other perceptual magnitudes, is by nature a “multidimensional” object. Researchers introduced many quality features of speech signals: intelligibility, clearness, brightness, loudness, naturalness, nearness, spaciousness, etc. For instance, a good intelligibility of the transmitted and/or processed speech is a prerequisite for a maximum quality rating of the speech service. However, a perfect intelligibility of the talker’s message at the listener’s side is not sufficient to achieve high quality. For instance, the transmitted bandwidth can be restricted to the usual telephone bandwidth, while the intelligibility remains almost perfect.

According to Möller et al. [25], the QoE space of a speech communication service covers aspects of both speech perception and service usage. In Chap. 5, the QoE features were classified in terms of four levels from perception to service usage. In the field of speech communication, the first level of quality features called “level of

direct perception” corresponds to the perception by the ear of the acoustic wave and the transmission of the resulting auditory information to the central nervous system. In a conversational situation, when two conversation partners interact, the QoE of the speech service is influenced by several other features classified in terms of the “level of interaction”. For instance, this level includes the naturalness of the interaction between two interlocutors during a phone call. The third level of QoE features, the “level of the usage instance”, includes all features related to the physical and social environment at the talker’s and listener’s side. For instance, the background noise or the room reverberation at the listener’s side has an influence on the listening effort and thus on the QoE of the whole speech communication system [24]. Another example is the advantage of mobility with cordless terminals and mobile telephony. The last category of QoE features, called “level of service”, covers aspects like stability over the entire duration of the communication, call set-up duration or interruptions of the connection. This organization of quality features in four layers shows that quality features are related to both instantaneous and multi-episodic experiences of the service. All of these features play a role in the long-term acceptability of the service and the averaging process is relatively complex (see Chap. 10).

Since many speech quality features exist in the literature, several authors developed perceptual spaces based on few orthogonal quality features referred to as “speech quality dimensions”. The following section summarizes the speech quality spaces proposed in the literature. Wälterman et al. [41] combined two auditory methods to derive a speech quality space composed of the three following dimensions:

- **Discontinuity**: this dimension reacts to degradation in the time domain, i.e. an unpredictable variation over time of the signal.
- **Noisiness**: this dimension is affected by the amount of unwanted information added to the speech message (either noise or a second talker).
- **Coloration**: this dimension can be affected by the two following elements: (1) a deviation from a reference timbre (e.g. *dark* or *bright*) and (2) a bandwidth restriction.

These three dimensions are of the type vector model. In other words, the origin of the space defines the highest quality and the space is defined by positive values only.

However, all speech stimuli employed in Wältermann et al. [41] were adjusted to a fix listening level of 79 dB SPL. Consequently, Côté [2] proposed to include a fourth quality feature to the perceptual space; **loudness**. Indeed, loudness is considered as the main feature of speech services QoE [5]. A loudness impairment is introduced in the case of non-optimal listening level, that is, an attenuation or an amplification introduced by the entire communication system. Loudness thus is a feature of the “ideal-point model” type. The three perceptual dimensions described by Wältermann et al. [41] are considered as orthogonal. However, the perceptual dimension “loudness” can be correlated with the other dimensions. The loudness summation effect shows that the bandwidth of a sound has an impact on its perceived loudness [4]. In Côté et al. [3], the authors showed the converse effect; the coloration due to a speech communication system has an influence on the optimal listening level.

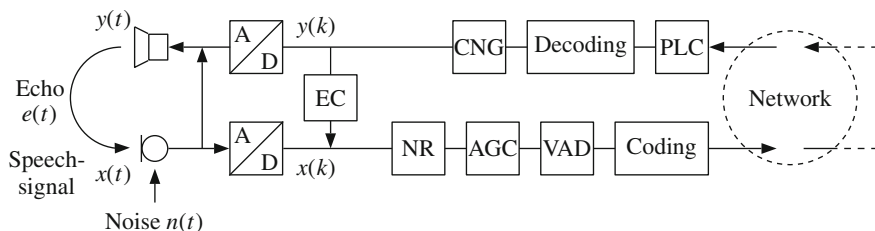


Fig. 12.1 Elements composing a speech communication system. A/D refers to analog to digital conversion, AGC to automatic gain control, EC to echo compensations, CNG to comfort noise generation, PLC to packet-loss concealment, VAD to voice activity detection and to NR noise reduction

These above described four features should reflect the whole perceptual quality space of transmitted speech. However, Sen [35] proposed a 5-dimensional space including noisiness and two sub-dimensions for each feature, coloration and discontinuity: slowly-varying and rapidly-varying discontinuities and low-frequency and high-frequency distortions. ITU-T Study Group 12 compares these two speech quality spaces within the work item P.AMD.

12.2 Speech Communication Systems

Nowadays, speech communication systems are composed of a multitude of components. The section below give an overview of the most important elements composing a speech communication system and their possible impact on its QoE. A typical example of such a system is depicted in Fig. 12.1. From the four types of speech processing systems described by Rabiner [30], only speech coding is introduced in the present chapter. Speech synthesis is covered by Chap. 13, speech recognition and speaker verification technologies are not covered by the present book.

Firstly, a telephone user talks and produces an acoustic signal, $x(t)$. This signal is received by the microphone of the talker's handset. However, this handset also receives sound from the environment, $n(t)$, produced by the sound sources surrounding the telephone user. The microphone converts the acoustic signal into an electrical signal, which is digitalized (i.e. sampled and quantized in $x[k]$, where k is the sample index) and *pre*-processed in order to remove the undesired signals (i.e. background noise, reverberation and echo). Then, this processed signal is encoded with a (low bit-rate) speech codec and sent to the transmission network. During the transmission to the handset of the conversation partner, the signal passes through several gateways and nodes. At the listener's side, a continuous electrical signal is decoded with the help of several digital "post-"processing algorithms. Then, the loudspeaker of the listener's handset converts the processed electrical signal into an acoustic signal,

$y(t)$. The listener's acceptability of the whole communication service based (mainly, but among other aspects) on the perception of the transmitted acoustic signal, $y(t)$, is the subject of the present chapter.

12.2.1 Telephony Networks

The traditional analog telephone network, referred to as Public Switched Telephone Network (PSTN), has been optimized for an almost perfect intelligibility of the speech message. For instance, the bandwidth of the transmitted speech corresponds to the transmission of the frequencies between 300 and 3,400 Hz that enables a comprehensibility of almost all phonemes. This bandwidth is, nowadays, referred to as Narrow-Band (NB). The PSTN is based on a "circuit-switched" network: the two interlocutors are connected by a physical circuit. In such a network, all physical parameters are well controlled to ensure a stable Quality of Service (QoS): the network accessibility is guaranteed and preserved over the whole call. During the last two decades, the deregulation of the telecommunications market led to heterogeneous transmission systems and speech processing algorithms. The first main transition was the introduction of digital transmission, the ISDN (Integrated Services Digital Network), which resulted in a decrease of circuit noise. Then, mobile phone networks, GSM (Global System for Mobile) and UMTS (Universal Mobile Telecommunication System) networks, have been broadly set up all over the world. The users of mobile telephony services are able to move from a quiet (house, office) to a noisy environment (street, train station) during a phone call. However, these networks are highly dependent on the characteristics of the radio channel between the mobile phone and the antenna. This air path leads to interferences, producing bit errors and frame losses, and handovers between two cells, two codecs and/or even two bandwidths, producing discontinuities in the transmitted signal. These quality variations in mobile networks result in a perceived instability of the communication system (see Chap. 27).

In addition to mobile telephone networks, speech communications over computer networks have been introduced. The Voice over IP (VoIP) protocol is based on a discontinuous transmission of packets of data, and the network is consequently referred to as a "packet-switched" network. Nowadays, the packet-switched network is the most widely used transmission path, because of its enhanced flexibility compared to the circuit-switched network. For instance, large audio bandwidths can be transmitted such as Wideband (WB, i.e. 50–7,000 Hz), Super-Wideband (S-WB, i.e. 50–14,000 Hz) and Full-Band (FB, i.e. 20–20,000 Hz) bandwidths. These wider bandwidths introduce less coloration of the speech compared to the narrow telephone bandwidth and thus increase the QoE. For instance, a comparison of clean WB and NB transmissions shows an increased quality of 29% in the WB case [27, 29]. However, VoIP transmissions may increase several quality impairments. For instance, these wider bandwidths may increase the influence of the environmental noise at the talker's side, and the packetization process lengthens the

overall transmission delay. A long transmission delay may introduce an audibility of the talker's own voice (echo) and reduces talking quality and double talk capability.

A packet-switched network may introduces discontinuities in the transmitted speech message, too. This annoying degradation appears more frequently than in a circuit-switched network. These discontinuities have two origins: (1) the bit-rate allocation is not guaranteed over the whole call and (2) the packets can take different transmission paths that lead to a time-varying transmission delay. This variation in transmission delay is referred to as "jitter". To generate a continuous signal, a buffer is placed in front of the decoder. The size of this de-jitter buffer (e.g. 120 ms) defines the tolerated lengthening of transmission delay between two consecutive packets. However, the size of the jitter buffer increases the overall transmission delay and, thus, may affect the conversation effectiveness. In case the speech segment may be lost during the transmission or arrives too late to synthesize a continuous signal, an algorithm "reconstructs" the missing packets. This algorithm called Packet-Loss Concealment (PLC) reduces the discontinuities in the speech signal. Nowadays PLC algorithms uses time-scale modifications of the speech signals (also known as "time-warping") which enable a smooth reconstruction of the waveform and avoid any discontinuity in the speech signal.

12.2.2 User Interfaces

The physical interface between the customers and the transmission system can be a handset, a headset or a Hands-Free Terminal (HFT). Such acoustic terminals have an influence on the speech coloration. The timbre modification of the talker's voice is introduced by the electro-acoustic properties of the two transducers (microphone and loudspeaker). Therefore, QoE of user interfaces is determined by their sending and receiving frequency response characteristics. In addition, loudness is a main parameter for all acoustic interfaces. According to the "orthotelephonic reference position" [12], the output signal loudness of such acoustic terminals should be equivalent to the perceived loudness of two interlocutors having a face-to-face conversation at one-meter distance.

Nowadays, the handset manufacturers introduce new services to user terminals in order, for instance, to increase the mobility of the user. For instance, screens with haptic feedback are included in modern mobile phones. Place for transducers is consequently reduced and causes challenges for their acoustic design. Although they enable a greater mobility, these terminals include several digital processing systems such as noise reduction algorithm that may degrade the transmitted speech signal [26].

Table 12.1 Characteristics of NB speech coding algorithms

Codec	Codec type	Frame length (ms)	Bit-rate (kbits)	I_e
G.711	PCM	0.125	64	0
G.726	ADPCM	0.125	40	2
–	–	–	16	50
G.729	CS-ACELP	10	8	10
GSM-FR	RPE-LTP	20	13	20
GSM-EFR	ACELP	20	12.2	5

The value of I_e is expressed on the NB quality scale of the E-model [13], ranging from 0 to 100 [15]

Table 12.2 Characteristics of WB speech coding algorithms

Codec	Codec type	Frame length (ms)	Bit-rate (kbits)	$I_{e,WB}$
G.722	ADPCM	0.125	64	13
–	–	–	48	31
G.722.1	MLT	20	32	13
–	–	–	24	19
G.722.2	CELP	20	23.85	8
–	–	–	23.05	1
–	–	–	14.25	10
–	–	–	6.6	41

The value of $I_{e,WB}$ is expressed on the WB quality scale of the E-model [14], ranging from 0 to 129 [15]

12.2.3 Speech Coding

A speech coding algorithm is a system that reduces the network rate used to transmit the speech signal. The speech coder produces a compressed signal from the input speech signal, referred to as the *bitstream*. After transmission over the network, the aim is to get a synthesized speech signal as similar as possible to the original speech. The impact of the speech codec on QoE depends on three physical characteristics: (1) the bit-rate expressed in *kbits*, (2) the frame length expressed in *milliseconds* (typical ranges of frame length are 5–30 ms), and (3) the paradigm employed by the coding algorithm. Tables 12.1 and 12.2 present the characteristics of several NB and WB speech coding algorithms. Almost all speech codecs have a flat band-pass within the allowed transmitted bandwidth (NB, WB or S-WB) and a low quantization noise resulting in a perfect intelligibility of the coded speech. However, they introduce audible non-linear degradations that decrease their perceived quality and affect automatic speech and/or speaker recognition algorithms. The parameter called “equipment impairment factor” (I_e), used in the E-model [13], quantifies the degradation introduced by the coding–decoding process. In addition, the coding–decoding process introduces a delay which impacts the conversation effectiveness. Nowadays,

speech codecs use a simple model of human auditory perception [17], are scalable from NB to WB [18], and some modern codecs also allow for coding of both speech and audio signals [16].

12.2.4 Voice Quality Enhancement

Voice Quality Enhancement (VQE) algorithms are integrated into the network or even directly into the terminal to reduce the new impairments introduced by mobile or VoIP networks. These algorithms are, for examples, echo cancellation, noise reduction, de-reverberation and automatic gain control, see Fig. 12.1. Echoes of the talker's own voice is introduced either by an acoustic feedback at the listener's side or by an impedance mismatch at the interconnection between two networks. As already mentioned, the latter effect is exacerbated in packet-based networks due to longer transmission delay. Therefore, echo cancellation techniques are needed if the delay exceeds 15 ms. Noise reduction is another VQE algorithm that has been widely introduced in mobile terminals. It reduces the environmental noise at the talker's side transmitted by the network. This algorithm complemented by a de-reverberation algorithm and an echo canceller separates the desired signal components from the undesired ones. However, noise reduction algorithms based on spectral subtraction reduce the noise level but simultaneously introduce musical noises on the speech signal [33]. Therefore, Möller et al. [26] proposed to describe the speech degradations resulting from imperfect noise reduction and echo cancellation by two additional equipment impairment factors *Inr* and *Iec*.

12.3 Speech Communication QoE Measurement Methods

The following sections introduce the measurement methods employed to quantify and further estimate the QoE of speech communication services, i.e. speech transmitted through a network and/or processed by speech processing systems. However, voice and speech quality measurement methods are employed in very diverse scientific fields: medicine (e.g. the evaluation of voice-related problems), linguistics (e.g. cultural comparisons) or speech communication. Each field has its own assessment paradigm.

12.3.1 Auditory Methods

The most accurate auditory measurement method would be an assessment by customers in natural environments. In practice, such "in-field" tests are hardly implemented, and speech services QoE is assessed with artificial auditory quality

Table 12.3 5-point scales

Quality of the speech [22]	Score	Impairment [11]
Excellent	5	Imperceptible
Good	4	Perceptible, but not annoying
Fair	3	Slightly annoying
Poor	2	Annoying
Bad	1	Very annoying

tests carried out in laboratories where the perception process is “directed” by an experimenter. Many different auditory test methods are employed by the academic laboratories and the speech service industries. For instance, listening-only experiments are carried out to gather the most important QoE features. Their realism is lower than that of conversational tests, since only the transmission system influence factors are assessed. The P-Series of Recommendations published by the ITU–T describe a general framework of speech communication measurement methods. In a listening quality test (referred to as listening-only test by the ITU–T), the listeners rate on a measurement scale a set of short speech samples (4–8 s) transmitted by different speech communication systems. The most widely used measurement scale is the 5-point integral quality scale presented in Table 12.3 (left column [22]). Such methodologies are not suited to compare speech stimuli with small impairments. Consequently, high-quality speech processing systems are assessed by methodologies used in the audio world and published by the ITU–R organization [10, 11] (see Table 12.3, right column).

Most of the ITU–T and ITU–R auditory methods quantify the quality of a speech service with a single value. This value is often used as an estimation of the overall speech service QoE. In addition to these methods, more complex auditory test methods give diagnostic information about the assessed processing conditions. Such quality tests rely on either a multi-scale rating process or a multidimensional analysis of the auditory results. For instance, Voiers [39] developed a specific multidimensional scaling method called Diagnostic Acceptability Measure (DAM) which assesses quality features of speech samples. More recently, Wältermann [40] developed a similar method to assess the three speech quality dimensions discontinuity, noisiness and coloration. However, such multidimensional tests are expensive and time-consuming since the listeners are trained beforehand (experienced), and they employ several rating scales for each speech stimulus (see also Chap. 5).

12.3.2 Instrumental Methods

Auditory methodologies rely on judgments by test subjects who are asked to give their opinion about the quality of a speech stimulus. Since auditory tests are costly and time-consuming, instrumental methods have been developed. Instrumental methods

have different applications such as the daily monitoring of transmission networks (e.g. VoIP) or the optimization of processing systems (e.g. speech codecs). They provide either a single estimated value that possibly represents the quality of the speech communication system (integral models), or a decomposition of the quality into several quality features (diagnostic models). In the following sections, we review the reliable models employed to predict the different aspects of speech communication QoE. Many building blocks have been developed such as the ITU-T Rec. P.863 [23] model which estimates the listening quality of transmitted speech. However, a tool that covers all aspects of the QoE and predicts the overall QoE of speech communication services in a single value is not available yet.

Richters and Dvorak [31] proposed a performance model based on seven quality criteria (speed, accuracy, availability, security, simplicity and flexibility) for each function of the service (sales, connection, billing, technical support, etc). This model is employed to assess the QoS of speech communication services and covers many aspects of the service usage. More recently, Möller et al. [25] organized all QoS parameters of speech communication services in a theoretical model which covers the four levels of QoE-features (perception, interaction, situation and service). For an example of an exhaustive evaluation of a speech communication service with such quality criteria, see Chen et al. [1].

Many models have been developed and standardized to estimate the quality of transmitted speech in a listening-only situation. Takahashi et al. [37] classified them in three different groups: parameter-based models that use parameters describing the elements of the system (e.g. ITU-T Rec. G.107 [13]), signal-based models that use the transmitted or processed speech signal (e.g. ITU-T Rec. P.863 [23]), and the packet-layer models that use information about the service operation (e.g. ITU-T Rec. P.564 [21]). For instance, the well-known Perceptual Evaluation of Speech Quality (PESQ) model includes a robust time-alignment algorithm useful for VoIP variable delay [32]. The PESQ is now superseded by a new listening-only signal-based model, called POLQA [23], that represents an intrusive speech quality model suitable for NB to S-WB connections, electro-acoustic interfaces and VQE algorithms. Most of these models provide an integral estimation of the quality. Recently, diagnostic models have been developed in order to either indicate (1) the technical causes of a single impairment or (2) describe the communication system QoE on few speech quality features. In the former case, diagnostic models provide useful information to system designers and operators that help them for maintenance purposes. For instance, the ITU-T Rec. P.502 [20] describes standard methods to assess each element of user terminals and network components. The corresponding test signals are described in a separate standard [19]. These methods assess characteristics such as (1) the frequency response, the sidetone, the harmonic distortion and the loudness ratings of the user terminals, and (2) the echo loss, the double talk capabilities and the background noise of the transmission networks. Even though this first type of diagnostic models provides an exhaustive evaluation of the physical equipments, they do not help telecommunication providers to design a voice service optimized for their specific needs. The second type of diagnostic models describe a voice service in a simple quality space. They help the end-user to choose a voice service based on its cost

and its QoE. The benefit of such a diagnostic model has initially been investigated by Quackenbush et al. [28]. More recently, two sets of quality-feature estimators have been developed from the perceptual quality space derived by Wältermann et al. [41]. Côté [2] improved estimators initially developed by Scholz et al. [34] and Huo et al. [8] into a signal-based model called Diagnostic Instrumental Assessment of Listening-quality (DIAL). This model provides values of the four dimensions “coloration”, “discontinuity”, “noisiness” and “loudness”. Wältermann [40] proposed a diagnostic parametric model based on the E-model [13]. In parallel, Sen and Lu [36] derived four estimators for temporally localized (slow-jitter and fast-jitter) distortions and frequency localized (low-pass and high-pass) distortions.

So far, no instrumental model has been standardized for the estimation of the speech quality in a conversational situation. However, Guéguin et al. [6] proposed such a tool that combines estimations from three other models: PESQ and PESQM for listening- and talking-only situations, respectively, and the E-model for a delay impairment factor introduced by the transmission delay. Long-term quality estimation has been studied by more researchers. For instance, Weiss et al. [42] was able to estimate a long-term listening-only speech quality score (up to 2 min.) based on PESQ estimations for 4–8 s stimuli.

12.4 Conclusions and Future Trends in Speech Communication QoE

In this chapter we presented both the technical elements and quality features which are relevant for the Quality of Experience of speech communication systems. We reviewed the auditory and instrumental methods suitable for speech quality assessment with a focus on diagnostic instrumental models that provide one output per QoE dimension.

Over the last decades, instrumental models have been developed on either speech or music databases. The former ones estimate the QoE of speech services such as telephony, whereas the latter ones are dedicated to audio devices such as loudspeakers or headphones [38]. The new standard model POLQA [23] has been developed for the QoE estimations of speech communication systems only. However, both streaming and telephony services now employ similar packet-based networks. This new usages encourage the researchers to develop a common model that works with both types of input signals.

Current speech quality models do not cover the influence of the listener’s acoustic environment. Indeed, listening through a handset in a noisy environment involves binaural hearing which is not covered by current models. Even though many studies have been published over the last two decades, the effects of binaural hearing are still unclear and difficult to include in such quality models.

Further work is thus expected in the development of reliable instrumental methods. However, such instrumental methods require, at first, auditory test results. Therefore, the community of researchers who works in the field QoE would appreciate collaborations with voice service providers and developers of speech processing systems to get access to databases including specific impairments and/or listening contexts.

References

1. Chen K, Huang C, Huang P, Lei C (2006) Quantifying skype user satisfaction. In: Proceedings of the conference on applications, technologies, architectures, and protocols for computer communications (SIGCOMM), pp 399–410. Pisa
2. Côté N (2011) Integral and diagnostic intrusive prediction of speech quality. Springer, Berlin
3. Côté N, Gautier-Turbin V, Möller S (2007) Influence of loudness level on the overall quality of transmitted speech. In: Proceedings of the 123rd AES convention, 7175, New York
4. Fastl H, Zwicker E (2007) Psychoacoustics: facts and models, 3rd edn. Springer, Berlin
5. Fletcher H, Galt RH (1950) The perception of speech and its relation to telephony. *J Acoust Soc Am* 22(2):89–151
6. Guéguin M, Le Bouquin-Jeannes R, Gautier-Turbin V, Faucon G, Barriac V (2008) On the evaluation of the conversational speech quality in telecommunications. *EURASIP J Adv Signal Process*. Article ID 185248
7. Hardy W (2003) VoIP service quality: measuring and evaluating packet-switched voice. McGraw-Hill, New York
8. Huo L, Wältermann M, Heute U, Möller S (2008) Estimation of the speech quality dimension “discontinuity”. In: Proceedings of the 8th ITG-Fachbericht-Sprachkommunikation, Aachen
9. IEEE Standards Publication 297 (1969) Recommended practice for speech quality measurements. Institute of Electrical and Electronics Engineers, New York
10. ITU-R Recommendation BS.1116-1 (1997) Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. International Telecommunication Union, Geneva
11. ITU-R Recommendation BS.1284-1 (2003) General methods for the subjective assessment of sound quality. International Telecommunication Union, Geneva
12. ITU-T Handbook on Telephonometry (1992) International Telecommunication Union, Geneva
13. ITU-T Recommendation G.107 (2011) The e-model, a computational model for use in transmission planning. International Telecommunication Union, Geneva
14. ITU-T Recommendation G.107.1 (2011) Wideband e-model. International Telecommunication Union, Geneva
15. ITU-T Recommendation G.113 (2007) Transmission impairments due to speech processing. International Telecommunication Union, Geneva
16. ITU-T Recommendation G.718 (2008) Frame error robust narrow-band and wideband embedded variable bit-rate coding of speech and audio from 8–32 kbit/s. International Telecommunication Union, Geneva
17. ITU-T Recommendation G.722.1 (2005) Low-complexity coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss. International Telecommunication Union, Geneva
18. ITU-T Recommendation G.729.1 (2006) Based embedded variable bit-rate coder: an 8–32 kbit/s scalable wideband coder bitstream interoperable with G.729. International Telecommunication Union, Geneva
19. ITU-T Recommendation P.501 (2012) Test signals for use in telephonometry. International Telecommunication Union, Geneva
20. ITU-T Recommendation P.502 (2000) Objective test methods for speech communication systems using complex test signals. International Telecommunication Union, Geneva

21. ITU-T Recommendation P.564 (2007) Conformance testing for voice over IP transmission quality assessment models. International Telecommunication Union, Geneva
22. ITU-T Recommendation P.800 (1996) Methods for subjective determination of transmission quality. International Telecommunication Union, Geneva
23. ITU-T Recommendation P.863 (2011) Perceptual objective listening quality assessment. International Telecommunication Union, Geneva
24. Jung O (2012) Assessment of conversational speech quality inside vehicles, concerning influences of room acoustics and driving noises. *Acta Acustica Acustica* 98(3):461–474
25. Möller S, Berger J, Raake A, Wältermann M, Weiss B (2011) A new dimension-based framework model for the quality of speech communication services. In: Third international workshop on quality of multimedia experience (QoMEX), pp 107–112
26. Möller S, Kettler F, Gierlich HW, Poschen S, Côté N, Raake A, Wältermann M (2012) Extending the e-model for capturing noise reduction and echo canceller impairments. *J Audio Eng Soc* 60(3):165–175
27. Möller S, Raake A, Kitawaki N, Takahashi A, Wältermann M (2006) Impairment factor framework for wideband speech codecs. *IEEE Trans Audio Speech Lang Process* 14(6):1969–1976
28. Quackenbush S, Barnwell T, Clements M (1988) Objective measures of speech quality. Prentice Hall, Englewood Cliffs
29. Raake A (2006) Speech quality of VoIP—Assessment and prediction. Wiley, Chichester
30. Rabiner L (1995) The impact of voice processing on modern telecommunications. *Speech Commun* 17(3–4):217–226
31. Richters JS, Dvorak CA (1988) A framework for defining the quality of communications services. *IEEE Commun Mag* 26(10):17–23
32. Rix A, Hollier M, Hekstra A, Beerends J (2002) Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment, part i-time alignment. *J Audio Eng Soc* 50(10):755
33. Scalart P, Filho J (1996) Speech enhancement based on a priori signal to noise estimation. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP-96), vol 2, pp 629–632
34. Scholz K, Wältermann M, Huo L, Raake A, Möller S, Heute U (2006) Estimation of the quality dimension “directness/frequency content” for the instrumental assessment of speech quality. In: Proceedings of the 9th international conference on spoken language processing (ICSLP), Pittsburgh, pp 1523–1526
35. Sen D (2004) Predicting foreground SH, SL and BNH DAM scores for multidimensional objective measure of speech quality. In: IEEE international conference on acoustics, speech, and signal processing (ICASSP’04), vol 1, pp 493–496
36. Sen D, Lu W (2012) Objective evaluation of speech signal quality by the prediction of multiple foreground diagnostic acceptability measure attributes. *J Acoust Soc Am* 131(5):4087–4103
37. Takahashi A, Yoshino H, Kitawaki N (2004) Perceptual QoS assessment technologies for VoIP. *IEEE Commun Mag* 42(7):28–34
38. Thiede T, Treurniet W, Bitto R, Schmidmer C, Sporer T, Beerends J, Colomes C (2000) PEAQ—The ITU standard for objective measurement of perceived audio quality. *J Audio Eng Soc* 48(1/2):3–29
39. Voiers WD (1977) Diagnostic acceptability measure for speech communication systems. In: IEEE international conference on acoustics, speech, and signal processing (ICASSP’77), Hartford, pp 204–207
40. Wältermann M (2013) Dimension-based quality modeling of transmitted speech. Springer, Berlin
41. Wältermann M, Raake A, Möller S (2010) Quality dimensions of narrowband and wideband speech transmission. *Acta Acustica Acustica* 96(6):1090–1103
42. Weiss B, Möller S, Raake A, Berger J, Ullmann R (2009) Modeling call quality for time-varying transmission characteristics using simulated conversational structures. *Acta Acustica Acustica* 95(12):1140–1151