Objective Assessment of Speech and Audio Quality—Technology and Applications

Antony W. Rix, John G. Beerends, Doh-Suk Kim, *Senior Member, IEEE*, Peter Kroon, *Fellow, IEEE*, and Oded Ghitza

***Abstract*—In the past few years, objective quality assessment models have become increasingly used for assessing or monitoring speech and audio quality. By measuring perceived quality on an easily-understood subjective scale, such as listening quality (excellent, good, fair, poor, bad), these methods provide a quick and repeatable way to estimate customer experience. Typical applications include audio quality evaluation, selection of codecs or other equipment, and measuring the quality of telephone networks. To introduce this Special Issue, this paper provides an overview of the field, outlining the main approaches to intrusive, nonintrusive and parametric models and discussing some of their limitations and areas of future work.**

***Index Terms*—Audio quality, intrusive and nonintrusive testing, objective models, quality assessment, speech quality.**

I. Introduction

IN CONTEMPORARY audio and speech communications systems, low bit-rate coding has become ubiquitous, ranging from data rates below 4 kbit/s for some military and satellite applications, to around 128 kbit/s for music storage or digital radio. However, low bit-rate codecs, and processing elements such as echo cancellation and noise reduction in telephone networks, are highly nonlinear-, time-, and signal-dependent processes, and their effect on perceived quality is both significant and difficult to predict.

Until the 1990s, the standard way to measure the quality of these processes was to conduct a subjective test, which gives the subject group's *mean opinion score* (MOS) of the quality of each condition under test. Section II of this paper provides further details of the typical procedure. However, using human subjects in a controlled environment is expensive and slow, so while subjective tests are the ideal way to make substantial system decisions like the selection of a codec for an international standard, they are unsuitable for day-to-day quality evaluations.

The goal of objective measurement is to estimate subjective MOS automatically based on measurements of a system. Section II also discusses methods to evaluate the accuracy of an objective model's estimates, by comparison with subjective test results.

With the computing power available today, all of the methods in current use can be applied in real-time—for parametric voice over IP (VoIP) assessment models, a single, low-cost test probe can assess thousands of simultaneous calls, while the more computationally intensive signal-based models can still typically support several parallel channels from a single digital signal processor (DSP). This means that it is practical to use objective quality measures to optimize networks for quality, capacity or cost, or monitor networks based on customer experience.

An important factor in the development of accurate objective measures has been the use of models of human perception. Researchers in psychophysics have constructed models of several large-scale properties of the peripheral auditory system, including the perception of loudness, frequency, and masking [1]. The ways in which such techniques are used, and the strength of the perceptual analogy, vary considerably with the type of objective measure.

Intrusive models, discussed further in Section III, compare an original test signal with a degraded version that has been processed by a system. Such models have also been termed *comparison-based*, or *full-reference*. Most recent intrusive models work by transforming both signals using a perceptual model to reproduce some of the key properties of hearing, then computing distance measures in the transformed space and using these to estimate MOS (see, for example, [2]–[7]).

Nonintrusive models can be used in several configurations, introduced in Sections IV and V. *Nonintrusive signal-based models* (also known as *no-reference* or *single-ended* models), which are in their infancy compared to intrusive models, estimate MOS by processing the degraded output speech signal of a live network (see, for example, [8]–[11]). Several signal-based methods focus on models of speech production or speech signal likelihood, although many exploit some aspects of perception such as noise loudness.

In contrast, *nonintrusive parametric models* generally have no sound signal to process (and so make limited use of perceptual techniques), but instead estimate MOS from measured properties of the underlying transport and/or terminal, such as echo, delay, speech levels and noise [12], VoIP network characteristics [13], [14], or cellular radio reception measures [15]. Parametric models are also widely used for network planning to construct MOS estimates based on tabulated values such as the codec type, bit-rate, delay, packet loss statistics, etc. [16].

This last approach requires a full characterization of the system under test and is, therefore, sometimes referred to as

Manuscript received March 2, 2006, revised June 30, 2006. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rainer Martin.

A. W. Rix is with The Technology Partnership, plc., Cambridge SG8 6EE, U.K. (e-mail: awr@iee.org).

J. G. Beerends is with TNO Information and Communication Technology, 2600 GB Delft, The Netherlands (e-mail: j.g.beerends@telecom.tno.nl).

D.-S. Kim is with Lucent Technologies, Whippany, NJ 07981 USA (e-mail: dsk@lucent.com).

P. Kroon is with Agere Systems, Allentown, PA 18109 USA (e-mail: kroon@agere.com).

O. Ghitza is with Sensimetrics Corporation, Somerville, MA 02144 USA (e-mail: oded@sens.com).

Digital Object Identifier 10.1109/TASL.2006.883260

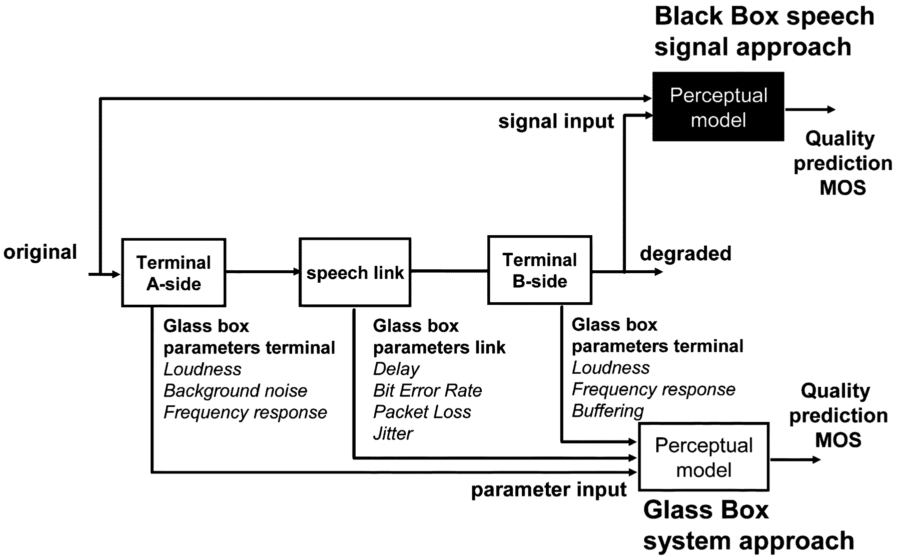

Fig. 1. Overview of the black box signal approach and the glass box system parameter approach. In a mixed approach, parameters can also be estimated from the signal input.

a *glass box* approach. Methods for which no knowledge of the system under test is required are referred to as *black box* approaches. Most real-world quality measurement systems use mixed approaches, requiring assumptions such as the type of codec and the audio terminals. An overview is given in Fig. 1.

It is important to note the limitations of objective models. They are trained on subjective test data that necessarily covers only a restricted set of conditions and is subject to voting errors, and models are influenced by other variables such as the signal being assessed. This limits their accuracy [7], [17]. These issues, and other factors in the application of objective models, are discussed in Section VI. Finally, some areas of interest for further work are highlighted in Section VII.

II. Subjective Quality

A subjective test begins with high-quality recorded material that is representative of the content that the systems under test will be used to process. These *original signals* (also termed *reference signals*) are processed through a wide range of conditions. In addition to the systems under test, standard reference conditions (*anchors*) are used to provide comparison with other subjective tests and ensure that the test is reasonably balanced. Both processed and original clips are then played, in randomized order, to subjects in a controlled environment using calibrated presentation equipment such as telephone handsets or headphones, and the subjects are asked to give their opinion of the quality or amount of degradation.

A. *Listening Tests for Telephony*

The most common approach used in the ITU-T and other telecommunications bodies is described in ITU-T P.800 and related recommendations [18], [19]. Typically, each original recording consists of speech sentence pairs of around 5 s to 8 s duration from a single talker, and recordings from two male and two female talkers are used to evaluate each condition under test. Noise may be added to simulate a noisy environment. The signals are processed through a filter modeling the handset send path, then a speech encoder and other processes such as packet loss that can occur in the network, followed by the decoder. Finally, the material is presented to 24 to 32 naïve subjects using a standardized telephone receiver, and subjects vote on the quality of each clip, most frequently using the five-point absolute category rating (ACR) listening quality (LQ) scale [Table I(a)], [18]. For the ACR method to work well, and to be comparable to intrusive models, it is important that the original signals are of high quality and are included as a condition in the test.

TABLE I
Opinion Scales

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

Other opinion scales have also been used for telephony subjective tests but are less common than ACR LQ. In particular for assessing performance in noise, test cases may involve a comparison between the (possibly noisy) original signal and the degraded output, using the degradation category rating (DCR) or comparison category rating (CCR) methods [18]. An advantage of these methods is that they fix the point of zero audible distortion, but due to the larger number of stimuli, they are much slower to conduct than ACR tests. Recently, ITU-T has approved a new subjective method specifically for evaluating the quality of noise reduction systems. After hearing each test item (with no reference), subjects are asked to give three separate votes on different aspects of the quality: the speech signal, the background, and the overall quality [20]. The overall quality rating is similar to a ACR LQ MOS score, while the other ratings provide possible insights on more detailed signal quality aspects.

The diagnostic acceptability measure (DAM), which also categorizes distortion according to several different labels, was widely used during the early development of quality assessment models and codecs [4]. Researchers have continued to examine the dimensions of subjective quality perception using methods such as multidimensional scaling (MDS) [21], [22].

B. *Audio Quality Testing*

Evaluating high quality audio codecs is challenging since distortions are often much more difficult to detect. The subjective test methods of ITU-R BS.1116 focus on small impairments, and use comparison-based methods rather than the ACR approach [23]. A range of speech and music clips is normally chosen, often narrowed down during pre-test evaluations to a small critical subset. Fewer subjects are used—often about ten—but they are frequently selected on the basis of having acute hearing and are trained to detect specific categories of codec distortion such as pre-echo. Unlike telephony speech quality testing, repeated listening, comparison with the original reference signal, and the use of a hidden reference, are common and consequently testing takes much longer. Often, the original and degraded signals are synchronously looped and subjects are free to switch between a hidden reference, the original and the degraded signal. Subjects listen using high-quality equalized headphones, and vote on a continuous opinion scale focusing on impairments [Table I(b)], [23]. The resultant quality score is termed subjective difference grade (SDG).

With the trend towards lower bit-rate audio with more audible distortion, methods in between P.800 and BS.1116 have also emerged. Listening quality can be used to evaluate audio, either using the P.800 methods (though with wideband headphones rather than telephone handsets) or the MUSHRA procedure [18], [24], [25]. These approaches are normally quicker than a BS.1116 test, though they typically need more subjects, and while they may be less able to discriminate very small differences it can be argued that they are more representative of real users' opinions.

C. *Conversation Quality*

The methods outlined above involve subjects listening passively. This lack of interaction makes them unsuitable to measure processes that only affect, or emerge, in interpersonal communication [16], [26], [27]. In general, a user's view of the quality of a conversation over a telephone connection is built up from three distinct attributes [28].

- Listening quality—how does the subject perceive the voice from the other side of the link (noise, distortion)
- Talking quality—how does the subject perceive his/her own voice (echo, sidetone, background noise switching)
- Interaction quality—how well can both parties interact with each other (delay, double-talk distortions).

Talking and interaction quality are difficult to assess because they strongly depend on properties of the voices and the conversation. Interaction quality depends on delay and the talk-spurt length used in a conversation [27]. The varying voice spectrum can affect how echo is perceived. While in subjective listening tests a large set of subjects can judge exactly the same physical signal, this is not possible in subjective conversational tests, making it more difficult to find relations between subjective and objective measurements.

In general, conversational tests use pairs of subjects, talking over a test network while performing some kind of interactive task, before voting (independently), normally using the quality scale [Table I(a)], [27]. This allows tests to take account of all of the properties of the network from each talker's mouth to ear, including sidetone and handset acoustics, echo, delay, and level impairments [16], [18]. However, conversational tests are relatively rare because they are slower and more expensive and complex compared to listening tests. Simpler and faster methods, such as double-talk tests and talking quality tests [26], are available, but model only some attributes of conversation quality and are not widely used.

D. *Performance Assessment of Models*

Because objective models are designed to be used alongside, or in place of subjective tests, their accuracy is evaluated by comparison to subjective test data. ITU-T P.800.1 [29] defines terminology to assist this:

- MOS-LQS—subjective MOS derived using an ACR LQ subjective test;
- MOS-LQO—objective estimate of MOS-LQS, typically from an intrusive or signal-based nonintrusive model;
- MOS-LQE—parametric estimate of MOS-LQS, typically from the E-model.

Equivalent terms for conversation quality are also defined (MOS-CQS etc). For BS.1116 audio quality, the corresponding terms to MOS-LQS and MOS-LQO are SDG and objective difference grade (ODG) [6].

With ACR listening quality, comparison between MOS-LQS and MOS-LQO can be difficult because there are often substantial variations between subjects, in particular from different countries due to either cultural or language differences. Combined with the tendency of subjects to vote depending on the range of quality of the conditions already heard—which often varies from test to test—there can be differences as large as 1.0 LQ MOS for the same network condition in different tests [30]. Similar issues can arise with ACR conversation quality tests. More generally, although objective quality should be monotonically related to MOS, the relationship is not necessarily linear; indeed, the variation between subjective tests is also often nonlinear. This is accounted for by using normalized objective quality to MOS mappings prior to computation of performance measures, as first introduced in [31]. Several authors have applied the logistic function [5], [32]. The logistic function is now less favored because its flatness outside the central range can conceal large prediction errors at high and low quality. The current most common approach is to use a monotonic polynomial fit [7], [33].

The preferred method of performance assessment in recent standardization work, used for example in the selection of P.862 and P.563, is as follows. The variation between tests is eliminated by applying a monotonic polynomial to mapping from objective scores onto the subjective scale for each MOS test. This function typically is fitted for minimum squared error with a gradient descent method, but is forced to be monotonic by using a cost constraint. It is important that the mapping function is monotonic because otherwise the rank order of predictions is not preserved. The main measure of an objective model's performance is the Pearson correlation coefficient. The residual error distribution may also be calculated as this gives an indication of

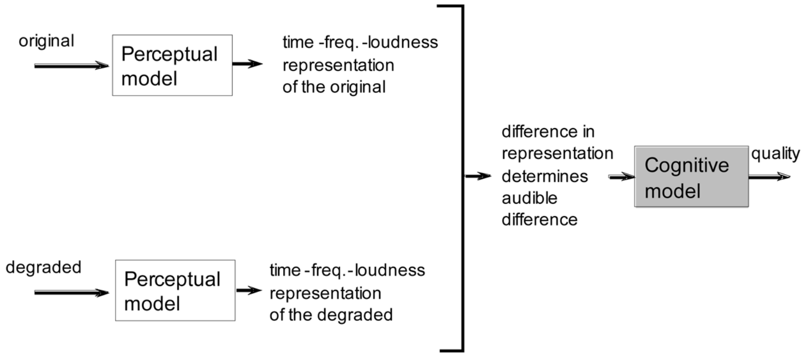
original
Perceptual model
time -freq. -loudness representation of the original
degraded
Perceptual model
time -freq. -loudness representation of the degraded
difference in representation determines audible difference
Cognitive model
quality

Fig. 2. Overview of the intrusive modeling approach.

the likelihood that a model's predictions lie within a given maximum error. For speech quality testing, to reduce variability due to the talker and random voting error, the measures are computed on the condition averages of MOS-LQS and MOS-LQO (i.e., the average MOS-LQS, and the average MOS-LQO, for all test cases representative of each given network condition) [7], [11], [33].

Many subjective tests are required to verify that a model will remain accurate with the very wide range of conditions that can be encountered in practical systems such as telephone networks. For example, the selection of P.862 was based on 22 subjective tests that were known to the authors in model training, and eight unknown tests run by independent laboratories, containing about 1300 conditions in total [33]. Readers are encouraged to be skeptical of the accuracy of any proposed objective models where few subjective tests are reported or where there is no independent validation of the results.

The need for data poses a significant problem, because relatively few subjective tests are in the public domain due to problems with cost, confidentiality, and the desire of laboratories to keep their original material unknown by codec developers or competitors. For speech quality testing, ITU-T has made available ten tests conducted by leading laboratories, split into three different sets of conditions [34]. For audio testing, EBU publishes a CD-ROM containing a wide range of critical, high-quality source items [35]. Several subjective testing and codec development laboratories have made other tests available to academic researchers—see, for example, [36]–[38].

III. Intrusive Objective Measures

Intrusive assessment is based on the use of known, controlled test signals which are processed through the condition under test. Both the original and processed signals are available to the model, and it is also typically assumed that the original signal is itself of near-perfect quality. (There has been little research into the case of capturing in-service signals at the input and output of a system for use with a comparison-based model, because of the lack of control over the quality of the original.) The problem in this case is to estimate the MOS of the degraded signal from the differences between the original and degraded signals.

The development of intrusive models has been closely related to that of low bit-rate speech and audio codecs. From the 1970s, many researchers have applied perceptual methods to speech codecs to allow coding distortions to be optimized for minimum audibility rather than mean squared error, leading to an improvement in perceived quality [40]. This concept was extended by Brandenburg to create the noise-to-mask ratio (NMR) measure, which uses a perceptual masking model to compare the level of the coding noise with the original signal [3]. Several other waveform difference measures have also been considered; see, for example, [4] or the references in [5].

The problem with using waveform difference to derive a quality measure is that substantial changes can be made to a signal waveform that give a large waveform difference, but little or no audible distortion: for example, waveform inversion and phase distortion.

These processes have minimal effect if quality predictions are made on the basis of differences in a transformed space that matches as closely as possible a hypothetical representation of the signal in the brain or peripheral auditory system. For intrusive quality assessment, this concept was introduced by Karjalainen in 1985 in the auditory spectrum distance (ASD) model [2]. It has become the most successful approach in intrusive quality assessment, used, for example, in ITU-T P.861 and P.862 and ITU-R BS.1387 [5]–[7]. In this approach, the audio signal is mapped from the time domain to a time frequency representation using the psychophysical equivalents of frequency and intensity, i.e., psychophysical frequency and loudness. (The more advanced methods model masking as the result of a combination of smearing and compression [6], [41]) In these models, the original and degraded signal are first filtered with the appropriate transfer function of the loudspeaker, handset, or headphone. Then the time signals are mapped to the time frequency domain and then smeared and compressed resulting in two time-frequency-loudness density functions. Next, these density functions are passed to a cognitive model that interprets the difference function, often with substantial additional processing, taking into account other properties such as the signals' instantaneous loudness. This cognitive model is trained using a large set of training data and validated on unknown data. An overview of this modeling is given in Fig. 2.

Although new psychoacoustic models, that closely follow our internal auditory processing steps, are capable of predicting many aspects of psychoacoustic data (see, for example, [42]), the most advanced models of human perception do not necessarily result in the best quality prediction model. Several authors have reported improved performance using more detailed models of perception [43]–[46]. On the other hand, Beerends found that the optimal time-frequency smearing for measuring speech quality does not coincide with the well known time-frequency smearing as found in psychoacoustics experiments [47]. Furthermore, Ghitza [48] reported that the correlation between subjective scores and objective predictions, using an advanced psychoacoustic model, was lower than with a simple approach as given in [49]. He suggested that a reason may be the lack of a good model of central auditory processing (i.e., beyond periphery), which governs the way humans judge the distance between acoustic stimuli. Beerends tried to incorporate this idea into an integrated approach towards measuring speech and music codecs using a more advanced cognitive model [50]. In the current standardized models for measuring audio quality [6] and speech quality [7], no satisfactory combination of advanced psychoacoustic and cognitive modeling is used.

A. *Audio Quality*

Intrusive models for audio quality assessment have generally focused on subjective tests conducted on a similar basis to BS.1116, with emphasis on distortions that are inaudible or just noticeable to expert listeners [23].

Several models developed during the early 1990s were submitted to a competition run by ITU-R from 1994–1996 [41], [51]–[55]. The first of these, Beerends and Stemerdink's perceptual audio quality measure (PAQM) was the most successful but did not meet the ITU-R's requirements. To achieve improved accuracy, PAQM was integrated with NMR and the other submitted models. Furthermore, the model of Thiede [53], which also incorporated a frequency response equalization process, was further enhanced and proved to be the best overall predictor of perceived audio quality. In the final perceptual evaluation of audio quality (PEAQ) standard both models are included and known as the simple and advanced method [6], [55]. The PEAQ advanced model in particular makes use of a wide range of perceptual transformations, representing the signals in terms of modulation, specific loudness, excitation, and excitation equalized for linear filtering and slow gain variation.

Recent work in audio quality assessment has included a new approach to measuring both linear and nonlinear distortions in the perceptual transform (excitation) domain [45], as well as assessment of features that can be used for the prediction of multichannel spatial audio fidelity [56]. Improvements to PEAQ, including new distortion parameters and a new cognitive model, have been proposed [57]. The limitation of PEAQ to a maximum of two channels has been addressed by the development of an expert system to assist with the optimization of multichannel audio systems [58].

B. *Speech Quality*

Since the 1980s, many intrusive speech quality estimation models have been proposed (see, for example, [2], [4], [31]–[33], [36], [37], [43], [44], [48], [49], [59]–[64]). The majority of these compute MOS-LQO for telephone-bandwidth speech conditions.

Following a competition run by ITU-T in 1994–1996, where several approaches were compared, a simple approach known as the perceptual speech quality measure (PSQM) was found to be the most accurate model with both known and unknown subjective MOS data. It was adopted as the first standard perceptual model, ITU-T P.861, in 1996 [5]. Unlike [2], PSQM did not take account of temporal masking. However, PSQM improved on earlier models in its silent interval processing, giving less emphasis to noise in silent periods than during speech, and its use of *asymmetry weighting*. Asymmetry weighting models the increased disturbance due to adding new, uncorrelated, time-frequency components to a signal (positive error) compared to attenuating or deleting components (negative error), as deletion is often concealed by the brain. These changes, combined with optimizations for increased correlation with MOS-LQS, led to a perceptual model that is highly specialized to the task of speech quality assessment, but that differs significantly from the peripheral auditory models in the literature.

PSQM and other early models were trained on subjective tests of speech codecs, and as a result they were found to perform poorly when used to assess some types of telephone network. Particular issues were found with packet loss and variable delay, both of which are common in VoIP, or small amounts of linear filtering that are unavoidable in analog connections. Intrusive models can measure large false errors, and can give highly inaccurate scores if these properties are not taken into account. Because of this, from the mid-1990s, researchers began to use larger subjective test databases, containing conditions representative of these network properties, to develop models that would be more accurate when used in the field.

The perceptual analysis measurement system (PAMS), developed by Hollier, Rix, and others [61], [62], extended the Bark spectral distortion (BSD) model of Wang *et al.* [31] for assessment of telephone network conditions. It includes a multiscale time alignment algorithm to account for VoIP variable delay [63], uses a filterbank rather than a fast Fourier transform (FFT), performs partial equalization of linear filtering, and maps multiple distortion measures to estimate subjective MOS on both the listening quality and listening effort opinion scales [18], [62]. Other authors working on BSD have proposed masking models [36], [64] or the use of phaseless coherence to eliminate the effect of linear filtering [37].

Focusing on the need for an intrusive model for assessment of telephone networks (including VoIP and mobile) as well as speech codecs, ITU-T held a second competition from 1998 to 2000. PAMS and an extended version of PSQM were the two models with the highest overall performance. To meet the ITU-T requirements, the time alignment of PAMS was integrated with the PSQM perceptual model, including improvements such as partial frequency response equalization, a simple masking model, and asymmetry weighting to model noise perception, and the model was retrained across a larger data set. The combined model, referred to as perceptual evaluation of speech quality (PESQ), was standardized as ITU-T P.862 in 2001, replacing P.861 [7], [33], [63]. The average correlation between PESQ score and MOS-LQS for the 22 subjective tests used in training was 0.935. The average correlation for eight unknown tests used for model validation was also 0.935. For an extended data set of 40 subjective tests including the training and validation sets, 93.5% of conditions were, after mapping, within 0.5 MOS, and 99.9% of conditions were within 1.0 MOS after mapping [33].

Note that these conditions represent essentially all data known at the time of standardization, including a substantial number of cases for which PESQ was found not to be as accurate as required and which are, therefore, excluded from its scope [7].

PESQ has been criticized, extended, or improved by several authors. Because PESQ scores are on an arbitrary scale that is not representative of typical subjective tests, a mapping from PESQ score to MOS-LQS, averaged across many different laboratories and subjective tests, was standardized by ITU-T to allow PESQ to give MOS-LQO figures that are on a 1–5 MOS scale typical of ACR LQ tests [30], [65].

PESQ has been modified to optimize performance for speech codecs operating below 4 kbit/s [66], which are outside the scope of P.862 [7]. The effect on PESQ of measurement conditions such as signal level has also been studied, and it has been noted that measured quality drops significantly if too high or low a level is used or if the signal spectrum is poorly matched, issues that are discussed further in Section VI [17], [67]. Finally, experiments have been conducted with PEAQ, PESQ and other models to change their input filter banks, which are all roughly based on the Bark scale, to the equivalent rectangular bandwidth (ERB) scale, though with PESQ this was found to produce little improvement [1], [68].

C. *Extension of Speech Quality Models*

Following the standardization of PESQ, work has continued to extend the scope of intrusive assessment beyond traditional telephony speech quality. A wideband version of PESQ, replacing its telephone handset input filter with a simple high-pass filter, has recently been standardized by ITU-T for assessment of wideband speech (50–7000 Hz) [24]. Models such as PESQ have been considered for use in assessing the quality of noise reduction algorithms, which pose an interesting problem because their complicated processing can improve or degrade quality depending on the signal conditions and subjective measurement method, and also it is not clear what reference signal should be used. New objective models have been proposed to address this, by taking as inputs not only the clean (noise-free) original and processed degraded signals, but also an intermediate signal, the noisy original signal prior to the noise reduction process [69], [70]. These new models estimate P.835 subjective quality using information such as the effect of the noise reduction on the instantaneous loudness of both speech and noise [20].

A major focus of recent research has been to include the acoustic path at the send and/or receive handsets, as well as hands-free terminals, to allow assessment of both handset and network. This is important because the acoustics and signal processing in the handset can have a substantial effect on overall quality. To allow acoustic measurements to be made, a head-and-torso simulator (HATS) models a user's mouth or ears [71]. Beerends, Berger, Goldstein, and Rix collaborated on a submission known as acoustic assessment model (AAM), which extended PESQ with improved level, time and frequency response alignment, temporal and frequency masking, and a binaural cognitive model. The model offered improved performance compared to PESQ for assessment of telephone networks at digital or analog electrical interfaces, particularly in the worst case [72], [73]. However, the lack of standards for subjective testing of acoustic network measurements with environmental noise, combined with concerns in the ITU-T about the need for a new model, led to AAM being shelved, and a new selection process is under way (see Section VII).

IV. Nonintrusive Measurement

Nonintrusive estimation of speech quality from speech waveforms is a challenging problem in that the estimation of speech quality has to be performed with the speech signal under test only, without using a reference. Considering that reference speech signals are not presented together with the speech under test to human listeners in ACR tests, nonintrusive estimation is more analogous to the situation of subjective ACR MOS tests. However, the lack of reference means that the compensation of variability in speech caused by different speakers and different utterances is quite limited in nonintrusive models compared to intrusive models.

Nevertheless, there has been substantial progress in this area, together with the growing need to monitor the speech quality of in-service networks, where intrusive models cannot be applied as reference speech signals uttered by end users are not controlled and may not be available to an objective model.

The first nonintrusive signal-based model in the literature was proposed in 1994 by Liang and Kubichek [8], and the approach that it uses, to estimate the difference between the measured signal and some ideal space of speech signals has been followed by several other authors. In this model, reference centroids are first trained from the perceptual linear prediction (PLP) coefficients [74] of nondegraded speech signals, and then the time-averaged Euclidean distance between degraded PLP coefficients and the nearest reference centroid is calculated as an indication of speech quality degradation. Various distortion measures commonly used in vector quantization (VQ) were explored to improve the performance of model [75], and an approach based on hidden Markov model (HMM) was also proposed [76]. Recently, the idea to measure the deviation of degraded speech from the statistical model trained on clean speech was expanded by Falk *et al.*, in which Gaussian mixture models (GMMs) were used to model the PLP feature vectors for clean speech. In addition to the clean reference speech signals, degraded speech signals were used in obtaining the multivariate adaptive regression splines to map the output of GMM to ACR LQ [38], [39].

A related approach has been to assume that most speech quality degradation caused by speech processing systems in telecommunication networks cannot be produced by biological human speech production systems due to the limited motor mechanism of the human vocal tract. Gray considered a model based on the parameterization of a vocal tract model which is sensitive to telecommunication network distortions [9]. Beerends and Hekstra also proposed a model based on the integration of a speech production model for detecting signal parts that cannot be produced by human vocal tracts and the PESQ intrusive model for estimating the impact of those signal parts [77].

In contrast to the direct utilization of a speech production model, Kim proposed an auditory model for nonintrusive quality estimation (ANIQUE) in which both peripheral and central levels of auditory signal processing are modeled to extract the perceptual modulation spectrum. The modulation spectrum is then related to the mechanical limitation of speech production systems to quantify the degree of naturalness in speech signals [10], [78].

ITU-T held a competition from 2002 to 2004 to standardize a nonintrusive signal-based model. In terms of network conditions, the scope of the model was set slightly wider than P.862, in particular including subjective tests with a broader range of acoustic inputs, network measurements, and noise types, and with talkers speaking in noisy conditions. Two proposals were submitted, with the ANIQUE model [10] being narrowly beaten by a combined model, known as single-ended assessment model (SEAM), which was based on three different models including those of Gray and Beerends and Hekstra [9], [77]. SEAM was adopted as ITU-T P.563 in 2004 [11], [79].

In P.563, a set of key parameters are extracted for the analysis of 1) vocal tract and unnaturalness of speech, 2) strong additive noise, and 3) interruptions, mutes, and time clipping. Based on these parameters, the intermediate speech quality is estimated for each distortion class, and the overall quality is obtained by a linear combination of intermediate speech quality with 11 additional signal features.

The average correlation of SEAM MOS-LQO with MOS-LQS was 0.88 over the set of 24 subjective tests used for training and validation. This is lower than the correlation of PESQ over the same data set (0.93), due not least to the lack of reference available to SEAM, but does indicate that the model has good correlation with subjective test data. While good progress has been made in nonintrusive assessment in recent years, there is clearly still scope for improvement and field experience of using this model.

V. Parametric or Glass Box Methods

Computational models have been widely used for many years for planning telecommunications networks without conducting subjective tests. The approach has more recently been applied to nonintrusive measurements of network parameters such as echo and delay, and to real-time assessment of VoIP systems where the dominant distortions, packet loss, jitter, and the codec, can be accurately modeled by a small number of statistical measures.

A. *E-Model*

The E-model is a telecommunication transmission planning model that was originally developed by ETSI for predicting the overall conversational quality of voice links [16]. The E-model presupposes that all parameters of the voice link that has to be assessed are known. In this glass box approach (see Fig. 1), the system under test is decomposed into a set of factors which affect the conversational quality. Within the telecommunication industry, a large set of commonly found contributing factors, such as loudness, background noise, low bit-rate coding distortions, packet loss, delay, echo, etc., have been quantified regarding their impact on the conversational speech quality. The primary output of the E-model is a quality rating factor R on a 0–100 scale. An invertible mapping exists between R and conversational MOS-CQE. Three main impairment factors are distinguished in the model: impairments which occur more or less simultaneously with the voice signal, impairments caused by delay and, the equipment impairment factor representing impairments caused by low bit rate codecs and errors such as packet loss. A method has been standardized by ITU-T for estimating equipment impairment factors using subjective tests or objective models such as PESQ [80].

The E-model was designed for evaluating networks that may not yet exist, and it is the only standardized model available for this purpose. However, several of the simplifying assumptions on which it is based—for example, linearity and order independence—are known to be wrong in some circumstances. As a result, it is recommended only for use as a planning tool. Nevertheless, by measuring certain parameters on the voice link it can also be used in voice quality monitoring [12], [13], [81]. While the current E-model applies only to telephone bandwidth (300–3400 Hz) speech signals, work is under way to extend it for wideband (50–7000 Hz) speech transmission systems [16], [82].

B. *Parametric Quality Measures of Specific Network Types*

For traditional telecommunications networks that are subject to minimal channel errors or coding distortions (typically those with only mu-law or A-law coding at 64 kbit/s), conversation quality is often dominated by talker echo from analog connections, round-trip delay, noise, and changes to the speech level. In-service nonintrusive measurement devices (INMDs) allow these network parameters to be measured, typically at trunk or international switching centers. Two models have been standardized to allow these objective parameters to be used to estimate conversational MOS: the E-model and the call clarity index (CCI) [12]. The E-model approach combines the measured parameters with a set of default assumptions, using the existing E-model framework to estimate R. CCI contains a functional mapping specifically derived to compute MOS-CQO from INMD measures of the speech and noise levels, echo loss, and delay.

A similar approach can be applied to other types of speech or audio processing. Subjective test data is used to train a functional mapping from the objective parameters to MOS. The resultant mapping is only applicable to the specific network types and the range of conditions exercised in the training data. It is also possible to use an intrusive model instead of subjective tests, although systematic inaccuracies in the intrusive model will be reflected in the parametric model. Examples of this approach are given in [14], [15].

One application of recent interest is to estimate the quality of in-service wireless networks from parametric measures of link-level properties such as bit-rate, residual bit error rate, and speech codec frame erasure [15], [83]. A challenge here is to gather enough data to adequately model the very wide range of types of error that can occur in current mobile channels, with variation in the data rate, channel signal-to-noise ratio, forward error correction, velocity and multipath profile, and the type and level of interference.

C. *Parametric Measurement of VoIP Quality*

Since the late 1990s, VoIP has become more and more common in telephony, leading to a number of quality issues including increased delay, packet loss, and the widespread use of compression as low as 5.3 kbit/s. Loss of packets due to network loss and delay jitter is particularly important to network operators because it is load-dependent and difficult to characterize, even in networks that use traffic management.

Two approaches to parametric VoIP quality monitoring have been proposed. To allow real-time monitoring in low-power edge devices and on network trunks that may carry very large numbers of active calls, it is not possible to process the speech waveform; instead, the models compute distortion parameters from the real-time protocol (RTP) transport that encapsulates the voice stream, and estimate round-trip delay from control protocol parameters.

Clark described a method based on the Gilbert model, a Markov chain model that is commonly used for modeling burst errors in a range of communications channels. From estimates of the Markov model parameters computed from the RTP stream, this model derives an equipment impairment factor, one of the inputs to the E-model [13]. Clark has also proposed modifying the E-model to take account of time-varying perception of quality, which is discussed in Section VI.

Broom has criticized this approach on the grounds that there are large variations between VoIP devices in the implementation of jitter buffers and error concealment. He has developed a proprietary model, based on multiple parameters extracted from the packet stream, that is calibrated for a specific VoIP device (such as an IP phone or gateway). This is achieved by making thousands of intrusive speech quality measurements of the device under test, using a network emulator to vary the operating conditions, and then training a numerical model to predict the PESQ scores for each condition [14].

Both of these models were presented to an ITU-T process to standardize a VoIP parametric model. No winner was selected; instead, a new ITU-T standard, P.564, is currently being considered by ITU-T study group 12. Previously known under the working title of P.VTQ (voice transmission quality), this is expected to recommend a method of performance assessment similar to the calibration process of [14]. Third parties could then use this performance assessment method to determine the accuracy of a VoIP objective model.

VI. Applications

A. *Use of Objective Models*

Objective models such as PEAQ (BS.1387), PESQ (P.862) and SEAM (P.563) are trained to predict the results of subjective tests, and as such, they are generally limited to the scope of the training data and cannot be guaranteed to perform well outside this [6], [7], [11]. For example, as mentioned above, the scope of PESQ does not include very low bit-rate audio codecs with data rates below 4 kbit/s because few of these were used in its training [7], and it been found to correlate relatively poorly with MOS when assessing such codecs [66].

Subjective test data imposes a limit on models' accuracy. Random errors in voting, combined with systematic bias due to factors such as the speech signals used or the balance of other conditions in the test, mean that each condition in a telephony MOS-LQS test typically has a 95% confidence interval between 0.1–0.3 MOS. Training using large numbers of subjective tests can allow models to improve on this a little, but the cost of subjective tests means that models are usually trained with at most a few tens of tests. As a result, accuracy for a given condition is normally on the order of 0.1 MOS. In practice, this seldom presents a problem because even expert listeners struggle to distinguish differences in quality of this magnitude in an ACR context.

Although the expectation is that continued progress will be made in the field of objective estimation of audio and speech quality, this practical limit to achievable accuracy will be difficult to surmount. As a result, for the most critical assessments such as the selection of a new codec, where the difference between candidates may well be smaller than this limit, carefully designed subjective tests (possibly used in combination with objective testing) are highly advisable.

An advantage of standardized models is that they are publicly available, peer reviewed, and independently tested. As a result, important conditions that can cause accuracy problems with a given model are often already known, and it is advisable for users to check the scope of a model before using it extensively. These problems may be due to issues in subjective test data—for example, if test results strongly conflict—or systematic bias with a specific type of distortion or test signal, and they can result in errors as large as 0.5–1.0 MOS in extreme cases. Some commercial implementations of models such as PESQ include automatic diagnosis of certain failure conditions, which can assist the operator, but this is not a substitute for reading the standard and its associated guidelines, or for listening to some of the conditions under test [7], [17].

The accuracy and repeatability of intrusive models depend to a great extent on how they are used, and mistakes remain relatively common. The test signals must be carefully chosen to be free from spurious noises, and have typical (and near-optimal) level and spectral content—for telephony, this usually involves prefiltering recordings with a send filter like the IRS—otherwise, they may cause codecs to behave poorly due to clipping or increased coding distortion, or lift the noise floor [17]. Speech signals should contain both speech and silent intervals and have a balanced phonetic content representative of the languages used in the applicable market [7]. Audio signals must be clean and cover a good range of content types, and the operator should determine whether to test using critical material such as [35], typical material, or a combination of both. Finally, the test equipment used to inject and record the signals must be of sufficiently high quality not to introduce additional distortion.

Signal dependence should also be considered because both speech and audio codec quality can vary significantly with signal content. With audio codecs, this is often of specific interest and so measurements are sometimes considered independently, although for overall comparisons it is normal to average results across all pieces of material for a given condition under test. Speech quality models, including PSQM, PESQ, and SEAM, have not generally been validated for talker or utterance (content) dependency, and they are designed to be used by averaging the objective scores for several different speech samples, typically taken from at least two male and two female talkers, to represent the quality of the condition [7].

Similar issues also affect nonintrusive models. The idealized conditions of subjective tests are a long way from the real signals obtained from measurement points in live networks, terminals, or other edge devices—key points of application of SEAM and other nonintrusive models. In particular, care needs to be taken if signals could have nonoptimal levels or spectral content, or where the signals are captured upstream of echo cancellers and as a result may contain high levels of network or acoustic echo.

B. Time-Varying Quality

Most of the models introduced above have focused on estimating short-term MOS, typically measured in subjective tests using 8 s samples. This timescale is long enough for subjects to form an opinion of quality, but short enough to allow time variation often to be ignored, and to minimize the total test length. However, a typical duration of a phone call or audio track is on the order of 2 min, and several researchers have studied how timescales may affect quality perception.

Gray found weak evidence, using 30 s speech paragraphs, that the first part of the speech sample had greatest weight on overall MOS, and termed this the *primary* effect [84]. Rosenbluth's study, which used 60 s of concatenated short-term speech samples, found the opposite, known as the *recency* effect because the last part of the long-term sample had greatest weight [85]. Several more recent tests have shown weak or no evidence for either the primary or recency effects [86], [87]. As these tests differed substantially in their design, but none was large in terms of number of subjects, speech samples or processing conditions, it cannot be said that there is clear evidence for either primary or recency effects.

However, in all of the tests, mean short-term MOS was a good first-order predictor of long-term MOS. This suggests that it is reasonable to average short-term objective quality scores over the duration of a call to obtain an estimate of overall MOS. Often this is complemented by also considering the worst-case short-term quality. A recent study by Raake indicates that poor worst-case quality does have a significant impact on long-term quality perception [81]. The use of mean and worst-case measures together avoids the possibility that periods of sustained poor quality could be outweighed by good quality—an issue with any single long-term MOS metric—where in practice a user may hang up if quality remains very poor for more than 10 s to 20 s.

VII. Future Work

A major trend in intrusive measurement of speech quality is the extension towards wideband signals. The new wideband PESQ (P.862.2) [24] has only been validated on a limited set of distortions, and further evaluations of this model are expected. ITU-T is preparing a call for proposals for a new model to replace or complement PESQ, with the working title P.OLQA (Objective Listening Quality Assessment). It also proposed to take into account the multiple dimensions of speech quality using multidimensional modeling [88], [89]. It is expected that the new method would be able to assess speech quality between telephone bandwidth (300–3400 Hz) and full audio bandwidth (20–20 000 Hz). Furthermore, the current E-model is being extended towards the use of wideband speech signals (50–7000 Hz) [82].

In audio quality assessment, PEAQ does not correlate well with wideband speech databases, and it was approved on the basis of near-transparent quality audio [6], while the mass-market use of audio codecs that has evolved in subsequent years is generally at a lower quality level. Although at present there are no plans for a replacement to this model, recent research indicates that improvements are possible [57], and some of the ideas in PEAQ could prove useful for P.OLQA. Other active areas of research are the quality assessment of multi-channel audio (e.g., 5.1 surround sound), and automotive audio quality [56], [58].

Conversation quality assessment, either intrusive or nonintrusive, is one area that remains little studied. This may be due in part to the difficulty and cost of conversational subjective tests, or to the improvement of echo cancellers which can minimize one of the dominant conversational impairments. This lack of research is a concern because most telephone systems are used conversationally. A concept for intrusive conversation quality assessment using two or more test nodes that emulate the behavior of human talkers has already been proposed, but has yet to be implemented [90].

Nonintrusive measurement methods, both signal-based and parametric, are relatively new, with the first generation of models designed for network assessment only emerging in the last five years. As these methods start to become widely used, it is highly desirable that experience on their strengths and weaknesses is published. The practical use of nonintrusive models such as [11] in devices such as mobile handsets, and in networks subject to two-way traffic including double-talk and echo canceller artifacts remains challenging.

Although multimedia is outside the scope of this Special Section, speech or audio quality is an important part of multimedia communications, and perceptual methods have also been applied to video quality assessment. Initial models for combining measures of audio and video quality to give an overall audiovisual MOS have been published, but they are based on limited data, and this remains a fruitful area for further research in both subjective and objective domains.

References

[1] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 4th ed. Norwell, MA: Academic, 1997.

[2] M. Karjalainen, "A new auditory model for the evaluation of sound quality of audio system," in *Proc. IEEE ICASSP*, 1985, pp. 608–611.

[3] K. Brandenburg, "Evaluation of quality for audio encoding at low bit rates," in *Proc. 82nd Audio Eng. Soc. Conv.*, 1987, preprint 2433.

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

[5] "Objective quality measurement of telephone-band (300–3400 Hz) speech codecs," 1998, ITU-T P.861.

[6] "Method for objective measurements of perceived audio quality," 1999, ITU-R BS.1387.

[7] "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2001, ITU-T P.862.

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

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

[10] D.-S. Kim, "ANIQUE: An auditory model for single-ended speech quality estimation," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 821–831, Sep. 2005.

[11] "Single-ended method for objective speech quality assessment in narrow-band telephony applications," 2004, ITU-T P.563.

[12] "Analysis and interpretation of INMD voice-service measurements," 2000, ITU-T P.562.

[13] A. Clark, "Description of VQMON algorithm," 2003, ITU-T del. cont. COM12-D105.

[14] S. Broom, "VoIP quality assessment: Taking account of the edge-device," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 6, pp. 1977–1983, No. 2006.

[15] M. Werner, T. Junge, and P. Vary, "Quality control for AMR speech channels in GSM networks," in *Proc. IEEE ICASSP*, 2004, pp. 1076–1079.

[16] "The E-model, a computational model for use in transmission planning," 2002, ITU-T G.107.

[17] "Application guide for objective quality measurement based on recommendations, P.862, P.862.1 and P.862.2," 2005, ITU-T P.862.3.

[18] "Methods for subjective determination of transmission quality," 1996, ITU-T P.800.

[19] "Subjective performance assessment of telephone-band and wideband digital codecs," 1996, ITU-T P.830.

[20] "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," 2003, ITU-T P.835.

[21] J. Hall, "Application of multidimensional scaling to subjective evaluation of coded speech," in *Proc. IEEE Workshop Speech Coding*, 2000, pp. 20–22.

[22] V.-V. Mattila, "Perceptual analysis of speech quality in mobile communications," Ph.D. dissertation, Tampere Univ. Technol., Tampere, Finland, 2001.

[23] "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," 1998, ITU-R BS.1116.

[24] "Wideband extension to recommendation for the assessment of wideband telephone networks and speech codecs," 2005, ITU-T P 862.2.

[25] "Method for the subjective assessment of intermediate audio quality," 2003, ITU-R BS.1534-1.

[26] S. R. Appel and J. G. Beerends, "On the quality of hearing one's own voice," *J. Audio Eng. Soc.*, vol. 50, pp. 237–248, 2002.

[27] N. Kitawaki, "Pure delay effects on speech quality in telecommunication," *IEEE J. Sel. Areas Commun.*, vol. 9, no. 4, pp. 586–593, May 1991.

[28] J. G. Beerends, "A subjective/objective test protocol for determining the conversational quality of a voice link," 2003, ITU-T cont. COM 12-55.

[29] "Mean opinion score (MOS) terminology," 2003, ITU-T P.800.1.

[30] A. W. Rix, "Comparison between subjective listening quality and P.862 PESQ score," in *Proc. Meas. Speech Qual. Net. (MESAQIN)*, 2003, pp. 17–25.

[31] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Sel. Areas Commun.*, vol. 10, no. 5, pp. 819–829, Jun. 1992.

[32] S. Voran, "Objective estimation of perceived speech quality—Part I: Development of the measuring normalizing block technique," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 4, pp. 371–382, 1999.

[33] J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, "Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment, Part II—Psychoacoustic model," *J. Audio Eng. Soc.*, vol. 50, no. 10, pp. 765–778, 2002.

[34] "ITU-T coded-speech database," 1998, Supp. 23 to P series rec., ITU-T.

[35] SQAM [Online]. Available: http://sound.media.mit.edu/mpeg4/audio/sqam

[36] W. Yang, M. Dixon, and R. Yantorno, "A modified Bark spectral distortion measure which uses noise masking threshold," in *Proc. IEEE Work. Speech Coding Telecom.*, 1997, pp. 55–56.

[37] S. W. Park, S. K. Ryu, Y. C. Park, and D. H. Youn, "A Bark coherence function for perceived speech quality estimation," in *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, 2000, vol. 2, pp. 218–221.

[38] T. H. Falk, Q. Xu, and W.-Y. Chan, "Non-intrusive GMM-based speech quality measurement," in *Proc. IEEE ICASSP*, Philadelphia, PA, 2005, pp. 125–128.

[39] T. H. Falk and W.-Y. Chan, "Single-ended speech quality measurement using machine learning methods," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1935–1947, Nov. 2006.

[40] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *J. Acoust. Soc. Amer.*, vol. 66, no. 6, pp. 1647–1652, 1979.

[41] J. G. Beerends and J. A. Stemerdink, "A perceptual audio quality measure based on a psychoacoustic sound representation," *J. Audio Eng. Soc.*, vol. 40, no. 12, pp. 963–974, 1992.

[42] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration," *J. Acoust. Soc. Am.*, vol. 102, no. 5, pp. 2906–2919, 1997.

[43] M. Hansen and B. Kollmeier, "Using a quantitative psycho-acoustical signal representation for objective speech quality measurement," in *Proc. ICASSP*, 1997, pp. 1387–1390.

[44] M. Hauenstein, "Application of Meddis' inner hair-cell model to the prediction of subjective speech quality," in *Proc. IEEE ICASSP*, 1998, pp. 545–548.

[45] B. C. J. Moore, C.-T. Tan, N. Zacharov, and V.-V. Mattila, "Measuring and predicting the perceived quality of music and speech subjected to combined linear and nonlinear distortion," *J. Audio Eng. Soc.*, vol. 52, no. 12, pp. 1228–1244, 2004.

[46] R. Huber and B. Kollmeier, "PEMO-Q—A new method for objective audio quality assessment using a model of auditory perception," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1902 –1911, Nov. 2006.

[47] J. G. Beerends and J. A. Stemerdink, "The optimal time-frequency smearing and amplitude compression in measuring the quality of audio devices," in *Proc. 94th Audio Eng. Soc. Conv.*, preprint 3604.

[48] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 1, pp. 115–132, Jan. 1994.

[49] J. G. Beerends and J. A. Stemerdink, "A perceptual speech quality measure based on a psychoacoustic sound representation," *J. Audio Eng. Soc.*, vol. 42, no. 3, pp. 115–123, 1994.

[50] J. G. Beerends, "Measuring the quality of speech and music codecs, an integrated psychoacoustic approach," in *Proc. 98th Audio Eng. Soc. Conv.*, 1995, preprint 3945.

[51] B. Paillard, P. Mabilleau, S. Morisette, and J. Soumagne, "PERCEVAL: Perceptual evaluation of the quality of audio signals," *J. Audio Eng. Soc.*, vol. 40, no. 1/2, pp. 21–31, 1992.

[52] C. Colomes, M. Lever, J. B. Rault, and Y. F. Dehery, "A perceptual model applied to audio bit-rate reduction," *J. Audio Eng. Soc.*, vol. 43, no. 4, pp. 233–240, 1995.

[53] T. Thiede and E. Kabot, "A new perceptual quality measure for bit rate reduced audio," in *Proc. 100th Audio Eng. Soc. Conv.*, 1996, preprint 4280.

[54] T. Sporer, "Objective audio signal evaluation—Applied psychoacoustics for modelling the perceived quality of digital audio," in *Proc. 103rd Audio Eng. Soc. Conv.*, 1997, preprint 4512.

[55] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, and B. Feiten, "PEAQ-The ITU standard for objective measurement of perceived audio quality," *J. Audio Eng. Soc.*, vol. 48, no. 1/2, pp. 3–29, 2000.

[56] S. George, S. Zielinski, and F. Rumsey, "Feature extraction for the prediction of multichannel spatial audio fidelity," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1994–2005, Nov. 2006.

[57] J. Barbedo and A. Lopes, "A new cognitive model for objective assessment of audio quality," *J. Audio Eng. Soc.*, vol. 53, no. 1/2, pp. 22–31, 2005.

[58] S. Zielinski, F. Rumsey, R. Kassier, and S. Bech, "Development and initial validation of a multichannel audio quality expert system," *J. Audio Eng. Soc.*, vol. 53, no. 1/2, pp. 4–21, 2005.

[59] A. De and P. Kabal, "Auditory distortion measure for speech coder evaluation—Discrimination information approach," *Speech Commun.*, vol. 14, pp. 205–229, 1994.

[60] ——, "Auditory distortion measure for speech coder evaluation—Hidden Markovian approach," *Speech Commun.*, vol. 17, pp. 39–57, 1995.

[61] M. P. Hollier, M. O. Hawksford, and D. R. Guard, "Error activity and error entropy as a measure of psychoacoustic significance in the perceptual domain," *Inst. Elect. Eng. Proc. Vision, Image, Signal Process.*, vol. 141, no. 3, pp. 203–208, 1994.

[62] A. W. Rix and M. P. Hollier, "The perceptual analysis measurement system for robust end-to-end speech quality assessment," in *Proc. IEEE ICASSP*, 2000, vol. 3, pp. 1515–1518.

[63] A. W. Rix, M. P. Hollier, A. P. Hekstra, and J. G. Beerends, "Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment, Part I—Time-delay compensation," *J. Audio Eng. Soc.*, vol. 50, no. 10, pp. 755–764, 2002.

[64] B. Novorita, "Incorporation of temporal masking effects into Bark spectral distortion measure," in *Proc. IEEE ICASSP*, 1999, vol. 2, pp. 665–668.

[65] "Mapping function for transforming P.862 raw result scores to MOS-LQO," 2003, ITU-T P.862.1.

[66] J. Holub, M. D. Street, and R. Smid, "Intrusive speech transmission quality measurements for low bit-rate coded audio signals," in *Proc. 115th Audio Eng. Soc. Conv.*, 2003, preprint 5954.

[67] V.-V. Mattila and A. Kurittu, "Practical issues in objective speech quality assessment with ITU-T P.862," in *Proc. 117th Audio Eng. Soc. Conv.*, 2004, preprint 6209.

[68] P. Kozlowski and A. Dobrucki, "Proposed changes to the methods of objective, perceptual based evaluation of compressed speech and audio signals," in *Proc. 116th Audio Eng. Soc. Conv.*, 2004, preprint 6085.

[69] J. Salmela and V.-V. Mattila, "New intrusive method for the objective quality evaluation of acoustic noise suppression in mobile communications," in *Proc. 116th Audio Eng. Soc. Conv.*, 2004, preprint 6145.

[70] V. Gautier-Turbin and N. Le Faucheur, "A perceptual objective measure for noise reduction systems," in *Proc. Meas. Speech Qual. Net. (MESAQIN)*, 2005, pp. 81–84.

[71] "Head and Torso Simulator for Telephonometry," 1996, ITU-T P.58.

[72] T. Goldstein, H. Klaus, J. G. Beerends, and C. Schmidmer, "Draft recommendation P.AAM—An objective method for end-to-end speech quality assessment of narrow-band telephone networks including acoustic terminal(s)," 2003, ITU-T cont. COM12-C64.

[73] T. Goldstein and A. W. Rix, "Perceptual speech quality assessment in acoustic and binaural applications," in *Proc. IEEE ICASSP*, 2004, vol. 3, pp. 1064–1067.

[74] H. Hermansky, "Perceptual linear prediction (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, pp. 1738–1752, 1990.

[75] C. Jin and R. Kubichek, "Vector quantization techniques for output-based objective speech quality," in *Proc. IEEE ICASSP*, Atlanta, GA, 1996, pp. 491–494.

[76] G. Talwar and R. Kubichek, "Output based speech quality measurement using hidden Markov models," in *Proc. Int. Signal Process. Conf.*, Dallas, TX, 2003.

[77] J. G. Beerends, P. Gray, A. P. Hekstra, and M. P. Hollier, "Call for proposals for a single-ended speech quality measurement method for non-intrusive measurements on live voice traffic," 2000, ITU-T cont. COM12-C11.

[78] D.-S. Kim, "A cue for objective speech quality estimation in temporal envelope representations," *IEEE Signal Process. Lett.*, vol. 11, pp. 849–852, 2004.

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

[80] "Methodology for the derivation of equipment impairment factors from instrumental models," 2002, ITU-T P.834.

[81] A. Raake, "Short- and long-term packet loss behavior: Towards speech quality prediction for arbitrary loss distributions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1957–1968, Nov. 2006.

[82] S. Moller, A. Raake, N. Kitawaki, and A. Takahashi, "Impairment factor framework for wideband speech codecs," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1969–1976, Nov. 2006.

[83] A. Karlsson, G. Heikkila, T. B. Minde, M. Nordlund, B. Timus, and N. Wiren, "Radio link parameter based speech quality index-SQI," in *Proc. IEEE Int. Conf. Elect., Circuits Syst.*, 1999, vol. 3, pp. 1569–1572.

[84] P. Gray, R. E. Massara, and M. P. Hollier, "An experimental investigation of the accumulation of perceived error in time-varying speech distortions," in *Proc. 103rd Audio Eng. Soc. Conv.*, 1997, preprint 4588.

[85] J. H. Rosenbluth, "Testing the quality of connections having time varying impairments," 1998, ITU-T del. cont. COM12-D64.

[86] N. Chateau, "Continuous assessment of time-varying subjective vocal quality and its relationship with overall subjective quality," 1999, ITU-T cont. COM12-94.

[87] L. Gros, N. Chateau, and S. Busson, "A comparison of speech quality judgments in laboratory and in real environment," in *Proc. 114th Audio Eng. Soc. Conv.*, 2003, preprint 5738.

[88] D. Sen, "Predicting foreground SH, SL and BNH DAM scores for multidimensional objective measure of speech quality," in *ICASSP*, 2004, pp. 493–496.

[89] J. G. Beerends, P. C. H. Oudshoorn, and J. M. van Vugt, "Speech degradation decomposition using a P.862 PESQ based approach," 2004, ITU-T cont. COM 12-C4.

[90] A. W. Rix, A. Bourret, and M. P. Hollier, "Modelling human perception," *BT Tech. J.*, vol. 17, no. 1, pp. 24–34, 1999.

Antony W. Rix received the M.Eng. degree in electrical and information science from the University of Cambridge, Cambridge, U.K., in 1996, and the Ph.D. degree from the University of Edinburgh, Edinburgh, U.K., in 2004 for his thesis on perceptual techniques in audio quality assessment.

From 1996 to 2000, he was a Researcher studying perceptual quality assessment at BT Laboratories, Suffolk, U.K., where he developed PAMS and worked with J. G. Beerends and others to combine PAMS and PSQM to produce PESQ (ITU-T P.862). He cofounded Psytechnics in December 2000, as a spinoff from BT Laboratories, with backing from 3i. Here, he contributed to the development of several new objective models including AAM, intrusive and nonintrusive video and audiovisual models, and applications of this technology in quality monitoring in both networks and handsets. In September 2004, he moved to The Technology Partnership plc. (TTP), Cambridge, U.K., where he is a Senior Consultant focusing on wireless communications, DAB/DMB and DVB-H. His current interests include mobile TV devices, delivery systems and standards, and network planning, and he remains active in audio and video quality assessment. He has written or coauthored over 50 papers or standards contributions, and 10 patents

Dr. Rix received an Industrial Fellowship from the Royal Commission for the Exhibition of 1851, to support his Ph.D. research, and his work on PAMS gained awards from BT and the British Computer Society. He is a member of the IEE and AES. He is a Guest Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING Special Issue on Objective Quality Assessment of Speech and Audio.

John G. Beerends received the degree in electrical engineering from the Polytechnic Institute (HTS), The Hague, The Netherlands, in 1975, the M.Sc. degree in physics and mathematics from the University of Leiden, Leiden, The Netherlands, in 1984, and the Ph.D. degree from the Technical University of Eindhoven, Eindhoven, The Netherlands, in 1989. The main part of his doctoral work, which deals with pitch perception, was published in the *Journal of the Acoustical Society of America*.

From 1986 to 1988, he worked on a psychoacoustically optimized loudspeaker system for the Dutch loudspeaker manufacturer BNS. The system was introduced at the Dutch consumer exhibition FIRATO in 1988. In 1989, he joined the KPN Research Laboratory, Leidschendam, The Netherlands, where he worked up to December 2002 on audio and video quality assessment, audiovisual interaction, and on audio coding (speech and music). The work on audio quality, carried out together with Jan Stemerdink, led to several patents and two measurement methods for objective, perceptual, assessment of audio quality. The first method, for measuring speech codec quality, was standardized in 1996 as ITU-T Recommendation P.861 (PSQM. Perceptual Speech Quality Measure). The second method, for measuring music codec quality was integrated into ITU-R Recommendation BS.1387 (1999, PEAQ, Perceptual Evaluation of Audio Quality). From 1996 to 2001, he worked with A. Hekstra on the objective measurement of the end-to-end quality of video and speech. The work on speech quality, partly carried out together with researchers from British Telecom, resulted in ITU-T recommendation P.862 (2001, PESQ, Perceptual Evaluation of Speech Quality). In January 2003, he joined TNO Information and Communication Technology, Delft, The Netherlands, where he continues his research into the end-to-end quality of audio, video and multimedia services. He is the author or coauthor of more than 60 papers/ITU contributions, and 20 patents.

Dr. Beerends received an AES fellowship award in 2003 for his work on audio and video quality measurement. He is a Guest Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING Special Issue on Objective Quality Assessment of Speech and Audio.

Doh-Suk Kim (M'97–SM'05) received the B.S. degree in electronics engineering from Hanyang University, Seoul, Korea, in 1991, and the M.S. and Ph.D. degrees in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Taejon, in 1993 and 1997, respectively.

From 1993 to 1996, he was a Contractor with the Systems Engineering Research Institute (SERI), Taejon, working on feature extraction for robust speech recognition in noisy environments. In 1997, he served as a Postdoctoral Fellow at KAIST. From November 1997 to October 1998, he worked on objective speech quality assessment and its application to speech coding at the Acoustics and Speech Research Department, Bell Laboratories, Lucent Technologies, Murray Hill, NJ, as a Postdoctoral Member of Technical Staff. In November 1998, he joined the Human and Computer Interaction Lab, Samsung Advanced Institute of Technology (SAIT), Yongin, Korea, and engaged in research on phase perception of speech and low-bit-rate speech coding. Since June 2001, he has been with the Voice and Data Quality and Performance Analysis Group, Lucent Technologies, Whippany, NJ. His research interests include auditory psychophysics, objective assessment of speech quality, speech coding, and robust speech recognition.

Dr. Kim was the recipient of the Samsung Humantech Thesis Prize Silver Award in 1996, and is Guest Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING Special Issue on Objective Quality Assessment of Speech and Audio, to be published in 2006. Dr. Kim has published more than 50 papers and standard contributions, and holds six patents.

Peter Kroon (F'96) received the M.S. and Ph.D. degrees in electrical engineering from the Delft University of Technology, Delft, The Netherlands.

The regular-pulse excitation speech coding technique described in his Ph.D. dissertation forms the basis of the GSM full rate coder. In 1986, he joined Bell Laboratories, Murray Hill, NJ, where he has worked on a variety of speech coding applications, including the design and development of the 4.8 kbit/s secure voice standard FS1016 and the ITU-T 8 kbit/s speech coding standard G.729. From 1996 to 2000, he supervised a research group at Bell Labs, Lucent Technologies, working in the areas of speech and audio coding design and communications technology. In 2000, Dr. Kroon became director of Media Signal Processing Research, Agere Systems, Allentown, PA, a spin off from Lucent Technologies, where he was responsible for research and development of media processing for satellite radio, VoIP and cellular terminals. In 2003 he moved to the Mobility business unit of Agere, where he is chief multimedia architect responsible for algorithmic design, and software and hardware integration of multimedia components for cellular phones. Dr. Kroon has published more than 50 papers, and holds 12 U.S. patents.

Dr. Kroon received the 1989 IEEE SP Award for authors less than 30 years old, for his paper on regular pulse coding. He served as Member of IEEE Speech Committee (1994–1996), General Chair, IEEE Speech Coding Workshop 1997, Associate Editor of IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1997–2000), Member at Large, IEEE Signal Processing Society Boards of Governors (2001–2003), and as a Guest Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING Special Issue on Objective Quality Assessment of Speech and Audio (2006).

Oded Ghitza received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from Tel-Aviv University, Tel-Aviv, Israel, in 1975, 1977, and 1983, respectively.

From 1968 to 1984, he was with the Signal Corps Research Laboratory of the Israeli Defense Forces. From 1984 to 1985, he was a Bantrell Postdoctoral Fellow at the Massachusetts Institute of Technology, Cambridge, and a consultant with the Speech Systems Technology Group, Lincoln Laboratory, Lexington, MA. From 1985 to early 2003, he was with the Acoustics and Speech Research Department, Bell Laboratories, Murray Hill, NJ, where his research was aimed at developing models of hearing and at creating perception based signal analysis methods for speech recognition, coding, and evaluation. Since early 2003, he is with Sensimetrics Corporation, Somerville, MA, where he continues to acquire and model basic knowledge of auditory perception for the purpose of advancing speech, audio and hearing-aid technology.

Dr. Ghitza is an Elected Fellow of the Acoustical Society of America, "For contributions to signal processing techniques for speech" (1998), and is a Guest Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING Special Issue on Objective Quality Assessment of Speech and Audio.