On the Quality Assessment of Sound Signals

A. A. de Lima*, F. P. Freeland*, R. A. de Jesus*, B. C. Bispo*, L. W. P. Biscainho*, S. L. Netto*‡,
A. Said†, A. Kalker†, R. Schafer†, B. Lee†, and M. Jam†
*PEE/COPPE, Federal University of Rio de Janeiro - Rio de Janeiro, Brazil
†HP Labs - Palo Alto, USA
‡Contact address: PEE/COPPE/UFRJ, POBox 68504
Rio de Janeiro, RJ, Brazil, 21941-972. Email: sergioln@lps.ufrj.br

***Abstract*— This paper constitutes an introduction to the field of quality evaluation of sound (speech and audio) signals. The need for such an assessment is inherent to modern communications: VoIP, mobile phone, or teleconference systems require meaningful measures of performance, which may ultimately assure good service or profitable business. A brief survey on subjective and objective evaluation methods is provided. Recent developments as well as new topics to be investigated are also addressed. Experiments are conducted to illustrate how to validate quality assessment methods.**

I. Introduction

The concept and measure of quality can be as comprehensive as human needs and imagination. In the realm of sound, for example, one can be anywhere between:

- The minimum required quality to allow intelligible voice communication; and
- The highest possible quality towards audio fidelity for aesthetic enjoyment.

The many gradations and crossings in-between are determined by the target application. Along the second half of the 20^{th} Century, analog telephony established an acceptable standard for intelligible conversation, while high-end home audio pursued the utmost perfection in sound recording and reproduction. Modern technology provides control over (and thus, scalability of) quality level: Now, while speech quality in VoIP and telepresence services reaches unprecedented clarity, portable audio popularizes quite low-fidelity music reproduction.

Whatever the application considered, sound quality assessment is necessarily linked to human hearing—historically the first way to perform audio and speech quality assessment. Of course, serious measurements require systematic *subjective* methods [1]. Technical organizations and scientific societies have made consistent efforts towards standardization since the 1990s. Today, most published regulations can be found among ITU (International Telecommunication Union) Recommendations, which cover speech- as well as audio-oriented procedures for quality evaluation.

Subjective quality assessment assumes several constraints, among which the following can be mentioned:

- The number of subjects should be sufficiently large to produce meaningful statistics;
- Subjects should share equal control characteristics;
- Environmental conditions should be the same;
- The test procedure should be repeatable.

These issues can render subjective tests too expensive and time-consuming. Hence, automatic methods for evaluating sound quality are often desirable. Traditional measurements like the signal-to-noise ratio (SNR) fail to emulate human opinion. Psychoacoustic phenomena like absolute hearing threshold or masking between close stimuli must be included in perceptual modeling of the human auditory system—which, ultimately, provides clues on how subjective assessment works. Therefore, perceptual measurements [2] have been the preferred alternative in *objective* evaluation of sound signals.

In the following, a brief tutorial is presented on the subject of quality measurement of sound signals. For that purpose, the paper is organized as follows: Sections II and III discuss the main standardized methods for, respectively, speech and audio quality assessment. In Section IV, new trends are pointed out. Section V brings some experiments on validation of quality assessment methods, whereas conclusions are drawn in Section VI.

II. Quality Evaluation of Speech

A. Subjective testing

In the 1990s, the quality assessment of speech signals was made through subjective tests, where a group of listeners was asked to score the general quality of a given set of signals. These subjective tests were standardized in ITU-T P.800.1 [3]. Perhaps the most popular subjective test described in this norm is the absolute category rating (ACR). In this test, a speech stimulus is played to the subject, who is asked to answer the question "What is your opinion on the quality of speech?" using a discrete 1–5 scale, the meaning of which is provided in Table I. Typically, each signal has a 5 s to 8 s duration, comprising two sentences, separated by approximately 0.5 s of silence, of a single speaker. A total of about 50 speech samples are evaluated by 24–32 listeners, and the average result indicates the so-called *mean opinion score* (MOS) for that ensemble of signals.

TABLE I
MOS scale.

Score	ACR Listening Quality
1	Bad
2	Poor
3	Fair
4	Good
5	Excellent

Other listening-opinion tests also discussed in the ITU-T P.800.1 are the degradation category rating (DCR) and the comparative category rating (CCR). These two tests employ the original (reference) and modified versions of each speech sample,. The tests differ in the order where the signals are presented: In the DCR, the degraded signal immediately follows the original version, whereas in the CCR, the order is random. Also, the two tests use distinct discrete scales, with different numbers of output levels, to compare the quality of the second speech sample with respect to the first one: from 1 ("annoyingly degraded") to 5 ("inaudibly degraded") in the DCR, and from -3 ("much worse") to $+3$ ("much better") in the CCR.

B. Objective testing

Nowadays, the standard algorithm for objective evaluation of speech quality is the ITU-T P.862 *perceptual evaluation of speech quality* (PESQ) method [4]. This algorithm has evolved from previous

978-1-4244-1684-4/08/$25.00 ©2008 IEEE

ones (namely, the so-called PSQM99 and PAMS tests) and is basically composed by three successive stages: Pre-processing, perceptual modeling, and cognitive modeling.

The pre-processing block performs level alignment, filtering, and time alignment of the reference and degraded signals. These procedures improve correlation of both signals, assuring a fairer comparison of their pre-processed versions.

The perceptual modeling stage includes time-frequency mapping, frequency warping, and loudness mapping to emulate processing performed by the human hearing system on both signals.

The cognitive modeling determines two measures of noise disturbance which are combined to generate the PESQ score between the two speech signals. The PESQ output x can then be mapped onto the MOS scale using the relationship

$$y = 0.999 + \frac{4.000}{1 + e^{Ax+B}}, \tag{1}$$

with $A = -1.4945$ and $B = 4.6607$.

The PESQ method is suited for 300-3400 Hz telephone signals. A wideband PESQ (W-PESQ) version is defined in recommendation ITU-T P.862.2 [5], where the PESQ input filter is replaced by another one with a wider 50-7000 kHz bandwidth, representing headphone characteristics. Mapping the W-PESQ score onto the MOS scale also follows equation (1), but in this case with $A = -1.3669$ and $B = 3.8224$.

C. Non-intrusive testing

The PESQ method introduced above employs both the original and corrupted versions of a speech signal to objectively evaluate their similarity. A recent trend in subjective quality evaluation of speech signals attempts to perform such a task in an open loop, that is, based solely on the corrupted signal. This approach is commonly referred to as the non-intrusive method [6]. The ITU-T standard for this single-ended assessment is the P.563 algorithm [7] which is implemented in three stages: Pre-processing, distortion estimation, and perceptual mapping.

The first stage normalizes the signal level, filters the input signal imposing a frequency response similar to a telephone terminal, and detects intervals of silence, to discriminate speech from noise frames.

The distortion estimation is performed by three parallel blocks. The first one employs a tube model for the human vocal track and detects speech distortion by identifying unacceptable variations in the tube structure. The second block performs signal reconstruction to allow a PESQ-like procedure to evaluate speech degradation. The third block estimates specific degradations such as robotization and additive noise.

The third stage in the P.563 algorithm compares the three outputs from the previous stage to previously set thresholds for each type of disturbance. The significant distortions are then weighted together to determine an overall level of quality for the signal at hand. Despite its non-intrusive nature, P.563 results are highly correlated to subjective MOS levels, as verified in Section V.

III. Quality Evaluation of Audio

'High' audio quality implies accurate reproduction of sound in general; this must be the idea behind audio quality assessment methods. Furthermore, this very generality makes the definition of no-reference tests for audio quite difficult.

A. Subjective testing

In this section, two standards defined by the ITU-R for subjective evaluation of audio quality [1] are described. Both are intended for evaluation of mono (regarding basic audio quality), stereo (regarding either basic audio quality or stereophonic image quality), and multichannel (regarding either basic audio quality, or front image quality, or impression of image quality) audio systems, specially codecs.

The ITU-R BS.1116 [8] method targets small impairments. Tests are performed by at least 20 expert listeners, who are presented to triads (A,B,C) of signals: A is the reference signal, B and C are either the signal under test or an identical copy of A. Listeners grade the impairments of B and C against A, according to a continuous scale from 1 ('very annoying') to 5 ('imperceptible'). The corresponding subjective difference grade (SDG), ranging from -4 to 0, between the test and reference signals can also be defined. The number of different triads should be made 1.5 times the number of systems, but at least 5.

The ITU-R BS.1534 [9] recommendation targets intermediate impairments (grades 1–3 in BS.116). Tests are performed by at least 20 experienced listeners. They are presented the reference signal and a set containing a maximum of 15 unidentified versions of it, including the original and a lowpass version. Listeners grade the impairments of each signal in the test set against the reference, using a continuous scale from 0 to 100, divided into five equal intervals, ranging from 'bad' to 'excellent'.

Overall, signals employed in both tests should consist of natural (and neutral) audio capable of stressing the system under test, with durations between 10 and 25 s. Results should be reported with respective confidence intervals. The sometimes disputed requirement of expert or at least experienced listeners is essential to keep the variance of results reasonably low.

B. Objective testing

The ITU-R BS.1387 [10] is the objective counterpart of BS.1116 and BS.1534. It is a blend of individual contributions from several researchers. The resulting method, commonly referred to as the perceptual evaluation of audio quality (PEAQ) method, is described in two versions: Basic (less complex, for real-time applications) and advanced (a more complex, but more accurate procedure).

The PEAQ is an intrusive method, which receives at its inputs both the reference and the signal under test in stereo 16-bit PCM at 48 kHz, pre-aligned in time. The algorithm performs level adjustment of both signals, and maps them into a time-frequency domain. In a warped frequency scale, based on auditory critical bands, a perceptual model (including hearing threshold, masking etc.) is computed for each signal. These models are pre-processed and then mutually compared under several criteria, yielding the so-called set of model output variables (MOVs). These features are input to a neural network with predefined weights, which delivers an objective difference grade (ODG) between 0 ('imperceptible impairment') and -4 ('very annoying impairment') that is expected to emulate the SDG defined before. Their reported correlation is stronger than 0.85.

An arguable issue in PEAQ is the pre-trained neural network. It implies some specialization on the (coded) signals included in the training set. Moreover, defects 'unknown' to the net can yield unexpected results, thus preventing the use of the method for other impairments than coding artifacts.

IV. Recent Research and Future Trends

Besides the standardized methods presented in preceding sections, several alternative approaches can be found in the recent

literature. LCQA [11] is a non-intrusive method for monitoring of telephone-band speech quality over a network, from per-frame speech coding parameters. Rnonlin [12] and PEMO-Q [13] are both psychoacoustics-based intrusive methods for evaluation of speech and audio along the entire audible spectrum (the former approaching particularly nonlinear distortions). An interesting solution for nonintrusive assessment of speech and audio quality, based on the insertion of watermarks in the DWT (Discrete Wavelet Transform) representation of the signal, is described in [14]; the method has been systematically validated for telephone-band speech, only.

A careful search in the literature shows that several important issues in sound quality assessment require further investigation, especially concerning audio signals in full-audio band. Such issues may include, for instance, development of real-time non-intrusive methods; evaluation of high-quality speech in telepresence systems; or quality assessment of restored non-coded audio.

V. Experiments on Quality Assessment of Sound

This section illustrates how objective methods for audio quality assessment and their implementations can be systematically validated.

A. Speech Signals

Recently, the entire ITU-T portfolio has been made available to the general public. Using such data, the behavior of ITU-T standards P.862 (PESQ), P.862.2 (W-PESQ), and P.563 can be easily validated.

The database ITU-T P.Sup23 [15] has a set of approximately 200 non-corrupted speech signals with corresponding degraded versions, all sampled at 16 kHz. In all cases, signal degradation considered at least one coding/decoding cycle with the ITU-T G.729 speech codec. In such a case, since the G.729 requires an 8-kHz input signal, each original signal was at least once downsampled to 8 kHz, coded/decoded by the G.729, corrupted by a specific impairment procedure, and upsampled back to 16 kHz to generate its degraded counterpart. Due to the downsampling/upsampling procedure, all original signals have an 8-kHz bandwidth, whereas their degraded versions only have a 4-kHz bandwidth.

The P.Sup23 database includes three sets of corrupted signals corresponding to different experiments: In Experiment 1, additional coding/decoding cycles with the G.729 and/or other codecs are performed; In Experiment 2, noise is added to the coded signal; In Experiment 3, channel errors are emulated in the coded version of the signal. The corresponding ACR-MOS results from 24 listeners for Experiments 1 and 3 are provided in the database (which also includes CCR results for Experiment 2), and can then be used to validate the performance of some quality assessment methods for speech signals.

For instance, the objective evaluation of each degraded signal of Experiment 1 with respect to its original version, using PESQ, W-PESQ, and P.562 methods, is shown in Figure 1 along with the corresponding MOS result provided in P.Sup23. Figure 2 compares the MOS with the results obtained by the PESQ, W-PESQ, and P.563 algorithms for all degraded signals in Experiment 3.

Table II shows the correlation factor between the subjective MOS and the result from each ITU-T method in Experiments 1 and 3. From this table, one can notice that the outputs from the three quality assessment methods, particularly the PESQ algorithm, are strongly correlated to the subjective MOS in both experiments. Experiment 3 resulted in lower correlation due to its highly nonlinear characteristics. The lower correlation factor of the P.563 method results from its non-intrusive nature, whereas the average behavior of the W-PESQ could be explained by the difference between the bandwidths of original (8 kHz) and degraded (4 kHz) signals, which can affect the W-PESQ performance.

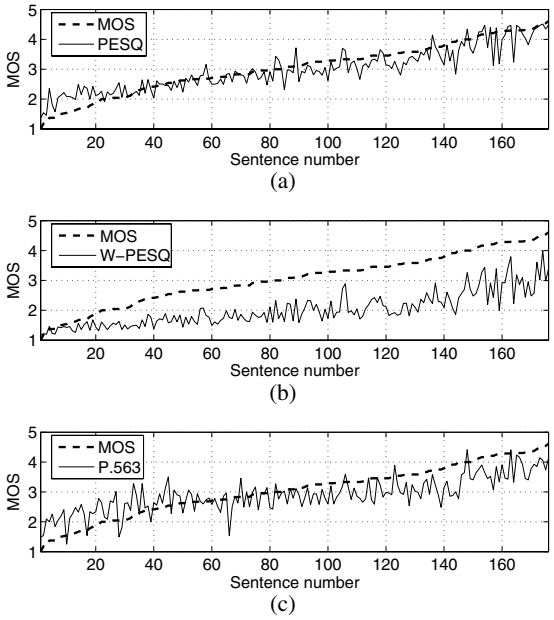

Fig. 1. Objective evaluation and MOS for each English sentence in Experiment 1 of ITU-T P.Sup23: (a) PESQ; (b) W-PESQ; (c) P.563.

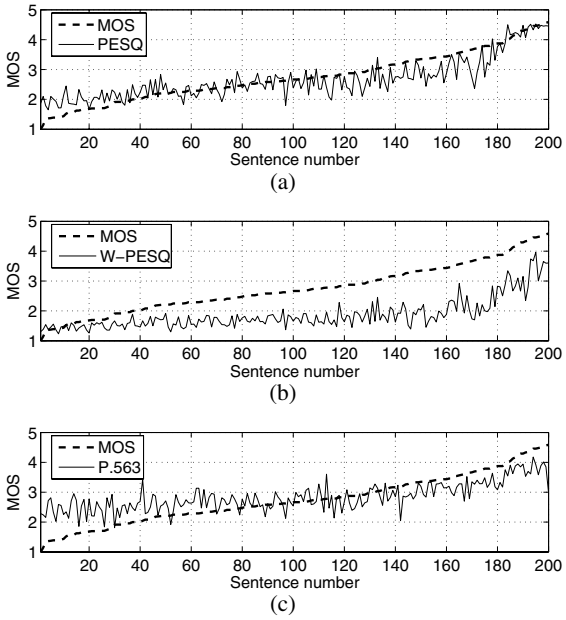

Fig. 2. Objective evaluation and MOS for each English sentence in Experiment 3 of ITU-T P.Sup23: (a) PESQ; (b) W-PESQ; (c) P.563.

B. Audio Signals

In this section we compare the performance of three versions of the ITU-R BS.1387 (PEAQ) algorithm for quality assessment of audio signals. For that purpose, we use 16 pairs of reference/degraded sound (13 of audio and 3 of speech) signals, whose PEAQ reference scores are provided in the same ITU-R recommendation. All signals are

TABLE II
CORRELATION FACTORS BETWEEN SUBJECTIVE MOS AND THE OBJECTIVE RESULTS FROM ITU-T METHODS.

Method	Experiment 1	Experiment 3
PESQ	0.91	0.87
W-PESQ	0.84	0.80
P.563	0.79	0.77

sampled at 48 kHz, as needed for PEAQ, and the codecs used for degradation are explicitly defined in the recommendation.

The tested algorithms were: Dr. Kabal's basic PEAQ implementation[1] (implementation A); Dr. Gottardi's basic PEAQ version[2] (implementation B); and an advanced PEAQ version referred to as implementation C.

In Figure 3, the ODGs provided in the ITU-R BS.1387 recommendation are compared with the results generated by each algorithm.

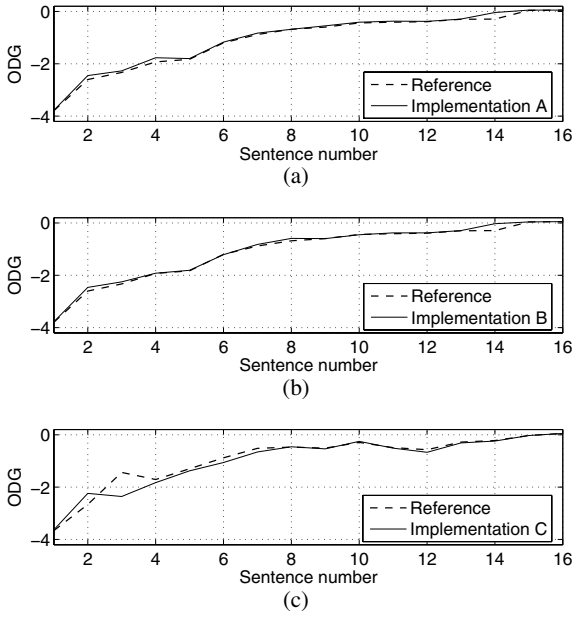

Fig. 3. Comparison of basic PEAQ reference version with other implementations: (a) Implementation A; (b) Implementation B; (c) Implementation C.

Table III shows the correlation factor and the mean squared error (MSE) between the output of each implementation and the PEAQ grades included in the ITU-R recommendation. From this table, one verifies that all three algorithms present quite reliable performance for these signals.

TABLE III
CORRELATION FACTOR (CF) AND MSE OF THREE DISTINCT PEAQ IMPLEMENTATIONS FOR SOUND SIGNALS PROVIDED IN ITU-R BS.1387 RECOMMENDATION.

Implementations	CF	MSE
A (Basic)	0.998	0.007
B (Basic)	0.998	0.009
C (Advanced)	0.962	0.076

[1]http://www-mmsp.ece.mcgill.ca/Documents/Software/index.html

[2]https://sourceforge.net/project/showfiles.php?groupid=89506&packageid=93837&releaseid=182742

VI. CONCLUSION

A brief survey on the subject of quality evaluation of sound (speech and audio) was provided using a framework based on several ITU recommendations. The methods discussed here included the P.800 (which describes the so-called MOS scale), P.862 (PESQ), P.862.2 (W-PESQ), and P.563 for speech; and BS.1116, BS.1534, and BS.1387 (PEAQ) for general audio signals. Recent results were addressed and future trends pointed out. Experiments were carried out to illustrate the recommendable methodology to compare objective and subjective evaluation methods, as well as to validate different implementations of a given method.

ACKNOWLEDGMENT

The authors would like to thank the CNPq Brazilian agency for supporting this work.

REFERENCES

[1] S. Bech and N. Zacharov, *Perceptual Audio Evaluation*. Chichester, UK: John Wiley, 2006.

[2] J. G. Beerends, "Audio quality determination based on perceptual measurement techniques," in *Applications of Digital Signal Processing to Audio and Acoustics*, M. Kahrs and K. Brandenburg, Eds. Norwell, USA: Kluwer, 1998, ch. 1, pp. 1–38.

[3] ITU-T, *Recommendation P.800.1, Mean Opinion Score (MOS) Terminology*. International Telecommunication Union - Telecommunication Standardization Sector, 2006.

[4] ——, *Recommendation P.862, Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*. International Telecommunication Union - Telecommunication Standardization Sector, 2001.

[5] ——, *Recommendation P.862.2, Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs*. International Telecommunication Union - Telecommunication Standardization Sector, 2005.

[6] A. Rix, J. G. Beerends, D.-S. Kim, P. Kroon, and O. Ghitza, "Objective assessment of speech and audio quality—technology and applications," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1890–1901, November 2006.

[7] ITU-T, *Recommendation P.563, Single-Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony Applications*. International Telecommunication Union - Telecommunication Standardization Sector, 2004.

[8] ITU-R, *Recommendation BS.1116-1, Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multivhannel Sound Systems*. International Telecommunication Unions - Radiocommunication Sector, 1997.

[9] ——, *Recommendation BS.1534-1, Methods for the Subjective Assessment of Intermediate Quality Level of Coding Systems*. International Telecommunication Union - Radiocommunication Sector, 2003.

[10] ——, *Recommendation BS.1387-1, Method for Objective Measurements of Perceived Audio Quality*. International Telecommunication Union - Radiocommunication Sector, 2001.

[11] V. Grancharov, D. Y. Zhao, J. Lindblom, and W. B. Kleijn, "Low-complexity, nonintrusive speech quality assessment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1948–1956, November 2006.

[12] C.-T. Tan, B. C. J. Moore, N. Zacharov, and V.-V. Mattila, "Predicting the perceived quality of nonlinearly distorted music and speech signals," *Journal of the Audio Engineering Society*, vol. 52, no. 7/8, pp. 699–711, July/August 2004.

[13] R. Huber and B. Kollmeier, "PEMO-Q - a new method for objective audio quality assessment using a model of auditory perception," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1902–1911, November 2006.

[14] R. Tu and J. Zhao, "Digital watermarking and quality evaluation of audio and speech," in *Advances in Audio and Speech Signal Processing: Technologies and Applications*, H. Perez-Meana, Ed. Hershey, USA: Idea Group, 2007, ch. VI, pp. 161–188.

[15] ITU-T, *Recommendation P.Sup23, ITU-T Coded-Speech Database*. International Telecommunication Union - Telecommunication Standardization Sector, 1998.