Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2009, Article ID 104382, 11 pages
doi:10.1155/2009/104382

Research Article

Performance Study of Objective Speech Quality Measurement for Modern Wireless-VoIP Communications

Tiago H. Falk[1] and Wai-Yip Chan[2]

[1] *Bloorview Research Institute, University of Toronto, Toronto, ON, Canada M5S 1A1*
[2] *Department of Electrical and Computer Engineering, Queen's University, Kingston, ON, Canada K7L 3N6*

Correspondence should be addressed to Tiago H. Falk, tiago.falk@ieee.org

Received 7 April 2009; Revised 10 June 2009; Accepted 8 July 2009

Recommended by James Kates

Wireless-VoIP communications introduce perceptual degradations that are not present with traditional VoIP communications. This paper investigates the effects of such degradations on the performance of three state-of-the-art standard objective quality measurement algorithms—PESQ, P.563, and an "extended" E-model. The comparative study suggests that measurement performance is significantly affected by acoustic background noise type and level as well as speech codec and packet loss concealment strategy. On our data, PESQ attains superior overall performance and P.563 and E-model attain comparable performance figures.

Copyright © 2009 T. H. Falk and W.-Y. Chan. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. Introduction

Due to the "best-effort" nature of current Internet Protocol (IP) connections, real-time speech quality monitoring is needed in order to maintain acceptable quality of service for voice over IP (VoIP) communications [1]. Traditionally, subjective quality assessment tests, such as the mean opinion score (MOS) test [2, 3], are used to quantify perceived speech quality. Subjective tests, however, are expensive and time-consuming and, for the purpose of real-time quality monitoring, have been replaced by objective speech quality measurement methods.

For VoIP communications, objective methods can be classified as either signal or parameter based. Signal-based methods use perceptual features extracted from the speech signal to estimate quality. Parameter-based methods, on the other hand, use VoIP connection parameters, such as codec, packet loss pattern, loss rate, jitter, and delay, to compute impairment factors which are then used to estimate speech quality. Such parameters are commonly obtained from the real-time transport protocol (RTP) header [4], real-time transport control protocol (RTCP) [5], and RTCP extended reports (RTCP-XRs) [6].

Current state-of-the-art signal based quality estimation algorithms perform well for traditional telephony applications but recent studies have found large "per-call" quality estimation errors and error variance [7–9]. Large estimation errors limit the use of signal-based methods for online quality monitoring and control purposes [10]. Parameter-based methods, on the other hand, can provide lower per-call quality estimation errors [11, 12] and have been widely deployed in VoIP communication services. The major disadvantage of parameter-based measurement is that distortions that are not captured by the connection parameters are not accounted for. Examples of such distortions include acoustic noise type, temporal clippings, noise suppression artifacts, as well as distortions in tandem connections caused by unidentified equipment and signal conditions.

Today, with the emergence of advanced technologies such as wireless local and wide area networks, the number of wireless-VoIP connections has grown substantially [13, 14]. Recent research by consulting firm ON World has suggested that by 2011 the number of wireless-VoIP users around the world will rise to 100 million from 7 million in 2007 [15]. Wireless-VoIP inter-networking results in tandeming of heterogeneous links which can produce new impairment

combinations that are not addressed by current standard quality measurement algorithms. Representative combinations of distortions can include (i) acoustic background noise (with varying levels and types) combined with packet loss concealment artifacts, (ii) acoustic background noise combined with temporal clipping artifacts (resultant from voice activity detection errors), or (iii) noise suppression artifacts (and residual noise) combined with speech codec distortions. In this paper, the effects of such "modern" degradation combinations on the performance of three International Telecommunications Union ITU-T standard algorithms are investigated. Focus is placed on *listening* quality, hence, factors such as jitter and delay, which affect *conversational* quality [16], are not considered.

The remainder of this paper is organized as follows. First, a brief overview of subjective and objective quality measurement is given in Section 2. Simulated wireless-VoIP impairments and details about the subjective listening test are presented in Section 3. Analysis of variance tests and quantitative algorithm performance comparisons are described in Section 4 and the quantification of overall performance loss is presented in Section 5. Lastly, conclusions are drawn in Section 6.

2. Speech Quality Measurement

In this section, a brief overview of subjective and objective speech quality measurement methods is given.

2.1. Subjective Measurement. Speech quality is the result of a subjective perception-and-judgment process, during which a listener compares the perceptual event (speech signal heard) to an internal reference of what is judged to be good quality. Subjective assessment plays a key role in characterizing the quality of emerging telecommunications products and services, as it attempts to quantify the end user's experience with the system under test. Commonly, the mean opinion score (MOS) test is used wherein listeners are asked to rate the quality of a speech signal on a 5-point scale, with 1 corresponding to unsatisfactory speech quality and 5 corresponding to excellent speech quality [2, 3]. The average of the listener scores is termed the subjective listening MOS, or as suggested by ITU-T Recommendation P.800.1 [17], MOS-LQS (listening quality subjective). Formal subjective tests, however, are expensive and time consuming, thus unsuitable for "on-the-fly" applications.

2.2. Objective Measurement. Objective speech quality measurement replaces the listener panel with a computational algorithm, thus facilitating automated real-time quality measurement. Indeed, for the purpose of real-time quality monitoring and control on a network-wide scale, objective speech quality measurement is the only viable option. Objective measurement methods aim to deliver quality estimates that are highly correlated with those obtained from subjective listening experiments. As mentioned previously, for VoIP communications objective quality measurement can be classified as either signal based or parameter based. Such measurement methods are described in the subsections to follow.

2.2.1. Signal-Based Measures. Signal based methods can be further classified as double-input (Figure 1(a)) or single-input (Figure 1(b)) depending on whether a clean reference signal is required or not, respectively. Such schemes are commonly referred to as double-ended or single-ended, respectively. Research into double-ended signal based quality measurement dates back to the early 1980s [18]. ITU-T Recommendation P.862 [19] (better known as perceptual evaluation of speech quality, PESQ) is the current state-of-the-art double-ended standard measurement algorithm. An in-depth description of the PESQ algorithm is available in [20, 21].

Single-ended measurement, on the other hand, is a more recent research field, and only recently (late 2004) has an algorithm been standardized. The ITU-T Recommendation P.563 [22] represents the current state-of-the-art single-ended standard algorithm for traditional telephony applications. A detailed description of the P.563 signal processing steps is available in [23]. Throughout the remainder of this paper, listening quality MOS obtained from an objective model will be referred to as MOS-LQO [17].

2.2.2. Parameter-Based Measures. Parameter based measurement, as depicted in Figure 1(c), was first proposed in the early 1990s by the European Telecommunications Standards Institute (ETSI). The ETSI computation model (so-called E-model) was developed as a network planning tool and describes several parametric models of specific network impairments and their interaction with subjective quality [24]. In the late 1990s, the E-model was standardized by the ITU-T as Recommendation G.107 [25]. The basic assumption of the E-model is that transmission impairments can be transformed into psychological impairment factors, which in turn, are additive in the psychoacoustic domain.

A transmission rating factor R is obtained from the impairment factors by

$$R = R_0 - I_s - I_d - I_{e-eff} + A, \tag{1}$$

where I_s, I_d, and I_{e-eff} represent impairment factors due to transmission (e.g., quantization distortion), delay, and effective equipments (e.g., codec impairments at different packet loss scenarios), respectively. R_0 describes a base factor representative of the signal-to-noise ratio and A an advantage factor; the R rating ranges from 0 (bad) to 100 (excellent). If the delay impairment factor I_d is *not* considered, the R rating can be mapped to listening quality MOS by means of equations described in ITU-T Recommendation G.107 Annex B [25]. Throughout the remainder of this paper, listening quality MOS obtained from E-model planning estimates will be referred to as MOS-LQE [17].

Several improvements to Recommendation G.107 have been proposed or are under investigation [26] in order to incorporate more modern transmission scenarios. Impairment factors, obtained from subjective tests, are described

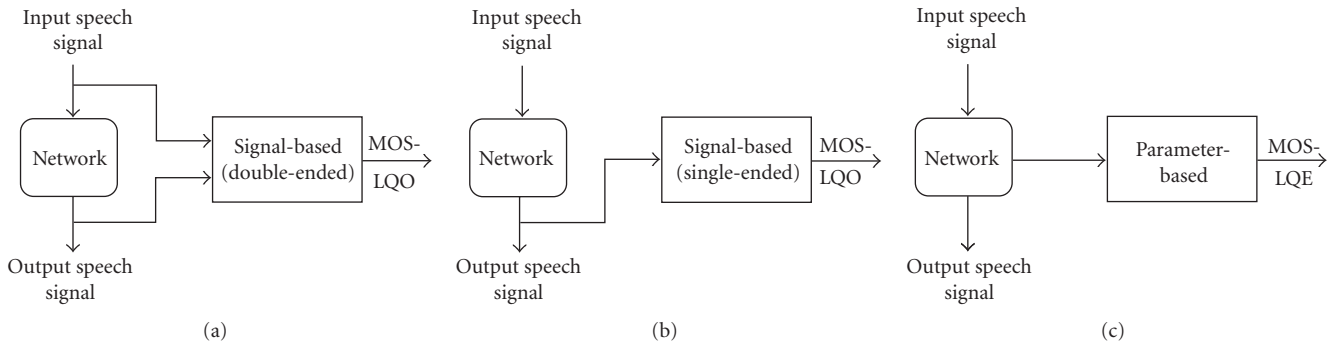


Figure 1: Block diagram of (a) double- and (b) single-ended single-based measurement, and (c) parameter-based measurement.

in [25, 27, 28] for several common network configurations. Impairment factors for alternate configurations can be obtained either from subjective MOS tests (according to ITU-T Recommendation P.833 [29]) or from objective methods (Recommendation P.834 [30]). As mentioned previously, the E-model is a transmission planning tool and is not recommended for online quality measurement. Hence, several extensions have been proposed to improve E-model performance for online monitoring. Representative extensions include nonlinear impairment combination models to compensate for high levels of "orthogonal" (unrelated) impairments [31], or online signal-to-noise ratio (SNR) estimation to account for varying background noise levels [32].

In this study, an "extended" E-model implementation is used. With the extended version, nontabulated equipment impairment factors (e.g., codecs described in Section 3 under 4% random and bursty packet losses) are obtained from subjectively scored speech data [12, 29]. Moreover, since degraded speech files have been artificially generated, the true noise level is used to compute MOS-LQE for noise-corrupted speech. Note that extended E-model performance may be favored with this unrealistic assumption that true noise level information is available online. In order to investigate a more realistic scenario, the noise level is measured in real time and is incorporated into the E-model in a manner similar to that described in [12, 32]. Here, the "noise analysis" module available in P.563 [22] is used to estimate noise levels online for the noisy and noise-suppressed speech signals. In controlled experiments, the noise level meter attained a correlation of 0.96 with the true noise level, computed both prior to and post-noise-suppression.

Moreover, equipment impairment factor values are currently not available for noise suppression algorithms and are the focus of ongoing research [26]. In fact, artifacts introduced by such enhancement schemes are dependent on the noise type and noise levels. In our experiments, the estimated noise level (post enhancement) is used in the computation of MOS-LQE for noise-suppressed speech. It is important to emphasize, however, that while using the estimated (or measured) noise level is convenient for quantifying noise artifacts that remain after enhancement, noise suppression artifacts that arise during speech activity are not accounted for. As emphasized in Section 5, this is a major shortcoming of parameter based measurement methods.

3. Experiment Setup

In this section, the degradation conditions available in the datasets—simulated wireless-VoIP, reference, and conventional VoIP—as well as the subjective listening tests are described.

3.1. Wireless-VoIP Degradation Conditions. The source speech signals used in our experiments are in English and French (four signals per language) and have been artificially corrupted to simulate distortions present in modern wireless-VoIP connections. Degradation sources that are commonly present in the wireless communications chain can include signal-based distortions such as acoustic background noise or noise suppression artifacts. These impairments are combined with distortions present in the VoIP chain, which may include codec distortions and packet loss concealment (PLC) artifacts.

To simulate the effects of acoustic background noise and codec distortions (including PLC artifacts), clean speech signals are corrupted by three additive noise sources (hoth, babble, car) at two SNR levels (10 dB and 20 dB). Noisy speech is then processed by three speech codecs: G.711, G.729, and Adaptive Multi rate (AMR). Random and bursty packet losses are simulated at 2% and 4% using the ITU-T G.191 software package [33]; the Bellcore model is used for bursty losses. Losses are applied to speech packets, thus simulating a transmission network with voice activity detection (VAD). The G.729 and AMR codecs are equipped with built-in PLC algorithms to compensate for lost packets. For the G.711 codec, two PLC strategies are investigated: the one described in [34] and a simple silence insertion scheme which is included to investigate the effects of acoustic noise combined with temporal clipping artifacts; the latter is referred to as "G.711*" throughout the remainder of this paper. Packet sizes are 10 milliseconds for the G.729 codec and 20 milliseconds for the remaining codecs. Moreover,

Table 1: Description of the 54 available noise-related degradation conditions.

Conditions	Tandem	Codec-PLC	Noise Type	SNR (dB)	Loss Type	Loss Rate (%)
1–48	None	G.711, G.711*, G.729, AMR	hoth	10,20	random, bursty	2
			car	10	random, bursty	2
			babble	10	random, bursty	2
			hoth	10,20	bursty	4
			car	10	bursty	4
			babble	10	bursty	4
49–54	Asynchronous	G.729 × G.729, AMR× AMR	car	10	random, bursty	2
			car	10	bursty	4

in the case of G.729 and AMR codecs, asynchronous codec tandem conditions are also considered (e.g., G.729 × G.729). A total of 432 speech signals (half English and half French) are available, covering 54 noise-related degradation conditions as detailed in Table 1.

Noise suppression artifacts combined with codec distortions are used to further simulate impairments introduced by wireless-VoIP connections. Here, the noise suppression algorithm available as a preprocessing module in the SMV codec is used [35]. The SMV codec per se is not used in our experiments as ITU-T P.563 has not been fully validated for such technologies [22]. Clean speech is corrupted by four noise types (hoth, car, street, and babble) at three SNR levels (0 dB, 10 dB, and 20 dB). Noisy speech is processed by the noise suppression algorithm and the noise-suppressed signal is input to the G.711, G.729 or AMR speech codec. As mentioned above, tandem conditions are also considered for the G.729 and AMR codecs. For noise suppression related impairments, a total of 192 speech signals (half English half French) are available, covering 24 degradation conditions as described in Table 2.

3.2. Reference Degradation Conditions. The multilingual datasets also include 128 reference-condition speech files which are commonly used in subjective listening tests to facilitate validation of test measurements and comparison with measurements from other tests. Reference conditions include modulated noise reference unit (MNRU) [36] at seven different signal-to-noise ratios (5–35 dB, 5 dB increments), as well as G.711, G.729, and AMR codecs operating in clean conditions either singly or in tandem. As described in Section 4.1, these reference conditions are used to map the datasets to a common MOS-LQS scale.

3.3. Conventional VoIP Degradation Conditions. Conventional VoIP degradation conditions are also included with the English and French datasets. With conventional VoIP conditions, *clean* speech, as opposed to noise-corrupted or noise-suppressed speech, is processed by the G.711, G.711*, G.729, and AMR codec-PLC schemes (singly or in tandem), under 2% and 4% random and bursty packet loss conditions. A total of 192 speech files (half English half French) covering 24 degradation conditions (no tandem: 4 codec-PLC types × 2 loss types × 2 loss rates; tandem: 2 codecs × 2 loss types × 2 loss rates) are available. Conventional VoIP data is used as a benchmark in Section 5 to quantify the decrease in quality measurement accuracy due to wireless-VoIP distortions.

3.4. Subjective Listening Tests. Source speech files were recorded in an anechoic chamber by four native Canadian French talkers and four native Canadian English talkers. Half of the talkers were male and the other half female. Clean speech signals were filtered using the modified intermediate reference system (MIRS) send filter according to ITU-T Recommendation P.830 Annex D [37]. Degraded speech signals were further filtered using the MIRS receive filter. In both instances, speech signals were level adjusted to −26 dBov (dB overload) and stored with 8 kHz sampling rate and 16-bit precision. Similar to the ITU-T Supp. 23 dataset [38], each speech file comprises two sentences separated by an approximately 650 milliseconds pause.

Two subjective MOS tests (one per language) were conducted in 2006 following the requirements defined in [2, 37]. Sixty listeners, native in each language, participated in each listening quality test and rated processed speech files described in Sections 3.1–3.3. Listener gender ratio was roughly one-to-one and listeners consisted of naïve adults (aged 18–50) with normal hearing recruited from the general population. Beyerdynamic DT 770 headphones were used and the listening room ambient noise level was kept below 28 dBA. Statistics for the subjective scores collected in the listening tests are listed in Table 3 for the wireless-VoIP, reference, and conventional VoIP degradation conditions.

4. Performance of ITU-T Standard Algorithms

In this section, two methods are used to assess the performance of PESQ, P.563, and the extended E-model for the wireless-VoIP distortion combinations described in Section 3.1. The first method, based on analysis of variance tests, investigates the performance sensitivity of current state-of-the-art algorithms to different wireless-VoIP degradation sources, such as noise type, noise level, packet losses, and codec-PLC type. The second method uses correlation and root-mean-square errors, computed between MOS-LQS and MOS-LQO (or MOS-LQE), to quantify the performance of

Table 2: Description of the 24 available noise suppression related degradation conditions.

Conditions	Tandem	Codec- PLC	Noise Type	SNR (dB)
1–18	None	G.711, G.729, AMR	hoth	0,10
			car	10
			street	10
			babble	10,20
19–24	Asynchronous	G.729 × G.729, AMR × AMR	hoth, car, babble	10

Table 3: Subjective score (MOS-LQS) statistics, separately for the English and French speech files, for the wireless-VoIP, reference, and conventional VoIP degradation conditions.

Statistic	Wireless-VoIP		Reference		VoIP	
	English	French	English	French	English	French
Minimum	1.05	1.02	1.30	1.11	2.52	2.09
Maximum	3.80	4.20	4.70	4.67	4.55	4.40
Average	2.43	2.42	3.77	3.63	3.67	3.36
Standard deviation	0.50	0.61	0.89	0.91	0.45	0.53

existing standard algorithms under modern wireless-VoIP communication scenarios.

4.1. Analysis of Variance. In this section, factorial analysis of variance (ANOVA) is used to assess the effects of different wireless-VoIP degradation on objective quality measurement performance. In particular, the effects of codec-cum-PLC type and acoustic background noise (type and level) are investigated using the noise-corrupted and noise-suppressed speech signals described in Section 3.1. For noise-corrupted speech, the effects of packet loss rates and packet loss patterns (random or bursty) are also investigated.

For the purpose of real-time quality monitoring and control, it is known that objective measures are required to provide low per-call estimation errors. Hence, we use per-sample MOS residual as the performance criterion; MOS residual is given by MOS-LQO minus MOS-LQS (or MOS-LQE minus MOS-LQS). In the analysis, raw MOS-LQO and MOS-LQE results (without mappings) are used. As shown in Section 4.2, mappings such as the one described in [39] can actually decrease algorithm performance.

In order to obtain a sufficiently large number of samples for variance analysis, a combined English-French speech dataset is used. It is important to emphasize that the English and French datasets were produced concurrently by the same organization, under identical conditions, with the only differences between them being the speakers, the spoken text, and the listener panels. Notwithstanding, in order to remove the differences between the English and French subjective scales, as suggested by the statistics in Table 3, a third-order monotonic mapping is trained between the English reference-condition MOS-LQS values and the French reference MOS-LQS values. The scatter plot in Figure 2 illustrates the French versus English reference

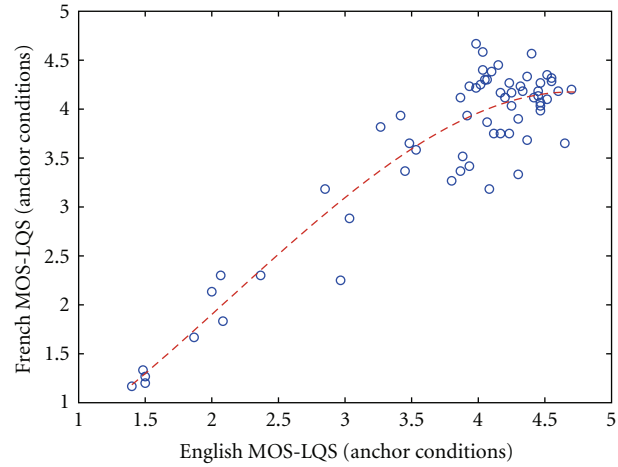


Figure 2: Scatter plot of English and French MOS-LQS values obtained from anchor conditions available in the datasets. The dotted line depicts the obtained third-order monotonic mapping used for dataset combination.

MOS-LQS values available in the datasets; the dotted curve represents the obtained polynomial. The combined dataset used for analysis comprises the French dataset described in Section 3.1 and the English dataset mapped to the French scale. This English-to-French scale mapping was chosen as it resulted in a larger reduction in mean absolute error between the two datasets of 0.13 MOS. After the mapping is applied, no significant differences are observed between the scales, as suggested by a t significance test ($P = .22$).

4.1.1. Main Effects. Table 4 shows F-statistics (F) and P-values (P) obtained from factorial ANOVA for noise-corrupted speech files. As can be seen, with a 95% confidence level, codec-PLC and noise type have significant main effects (i.e., $P < .05$) on the performance of all three objective measures. Noise level is shown to have significant main effects on E-model and PESQ performance and packet loss rate only on E-model performance. The box and whisker plots depicted in Figures 3(a)–3(c) assist in illustrating these behaviors, respectively; the plots illustrating the effects of packet losses on E-model performance are omitted for brevity.

The boxes have lines at the lower quartile, median, and upper quartile values; the whiskers extend to 1.5 times the interquartile range. Outliers are represented by the

Table 4: F-statistics (F) and P-values (P) of MOS residual errors (MOS-LQO/LQE minus MOS-LQS) obtained from factorial ANOVA with 95% confidence levels for noise-corrupted speech files.

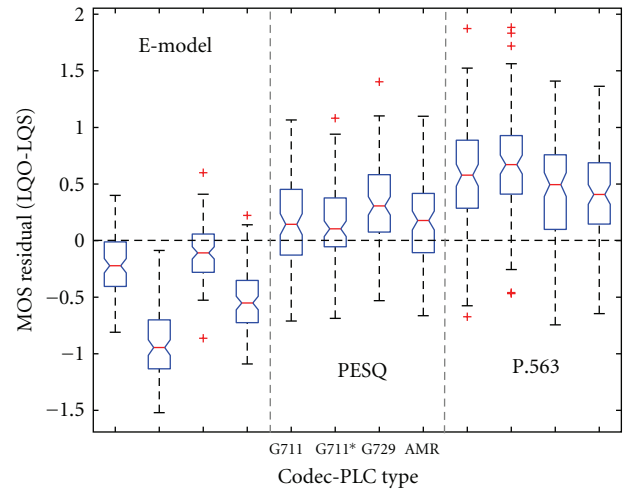
Impairment	E-model		PESQ		P.563	
	F	P	F	P	F	P
Codec-PLC	162.4	0	5.7	10^{-3}	8.2	10^{-5}
Packet loss rate	15.9	10^{-4}	.13	.7	3.1	.08
Noise type	3.1	.04	145.2	0	43.1	0
Noise level	7.7	.006	35.1	10^{-8}	.85	.36

symbol "+". The vertical width of the notches that cut into the boxes at the median line indicates the variability of the median between samples. When the notches of two boxes do not overlap, their medians are significantly different at the 95% confidence level [40]. In the plots, abscissa labels are omitted to avoid crowding; the missing labels can be obtained by periodically replicating the shown labels. Moreover, the abbreviation "LQO-LQS" is used for the ordinate labels to represent the MOS residual for all three measurement algorithms.

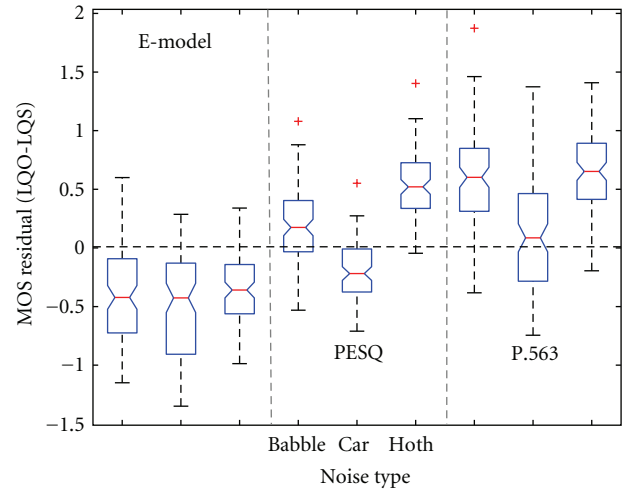
From Figure 3(a), it can be seen that larger E-model residual errors are attained for the silence insertion PLC scheme (represented by "G711*") followed by the AMR codec-cum-PLC. Furthermore, P.563 performance is lower for G.711-processed speech irrespective of the PLC strategy. According to [22], P.563 has only been validated for PLC schemes in CELP (codebook-excited linear prediction) codecs such as G.729; this can explain the poor performance obtained for G.711. Nonetheless, for the G.729 codec, P.563 attains residual errors that can be greater than one MOS point; on a five-point MOS scale, this can be the difference between having acceptable and unacceptable quality [9].

From Figures 3(b) and 3(c), it can be observed that E-Model and PESQ underestimate MOS-LQS for speech corrupted by car noise; E-model underestimates MOS-LQS for all noise types and levels. Figure 3(c) shows that P.563 performance is not significantly affected by noise level. This may be due to the fact that P.563 is equipped with a noise analysis module which not only estimates the SNR but also takes into account other spectrum-related measures such as high frequency spectral flatness. High frequency analysis, however, may be the cause of P.563 sensitivity to noise type since babble and car noise have low-pass characteristics. Ongoing research is seeking a better understanding of the limitations of signal [41, 42], and parameter-based [26] measurement of noisy speech.

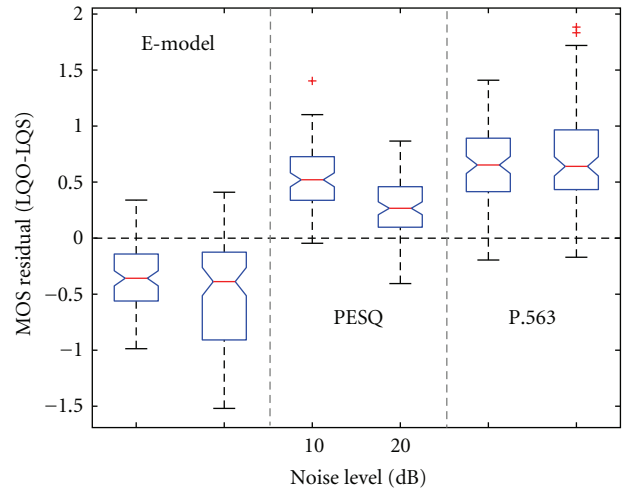
Table 5 shows F-statistics and P-values obtained from factorial ANOVA for noise-suppressed speech files. With a 95% confidence level, it can be seen that for noise-suppressed speech, codec-PLC type incurs significant main effects on E-model and PESQ performance. Noise type significantly affects PESQ and P.563 performance, and noise level (prior to noise suppression) incurs significant main effects on the performance of all three algorithms. The box and whisker plots depicted in Figures 4(a)–4(c) help illustrate these behaviors, respectively.



(a)



(b)



(c)

Figure 3: Significant main effects of (a) codec, (b) noise type, and (c) noise level on the accuracy of objective quality measurement of noise corrupted speech.

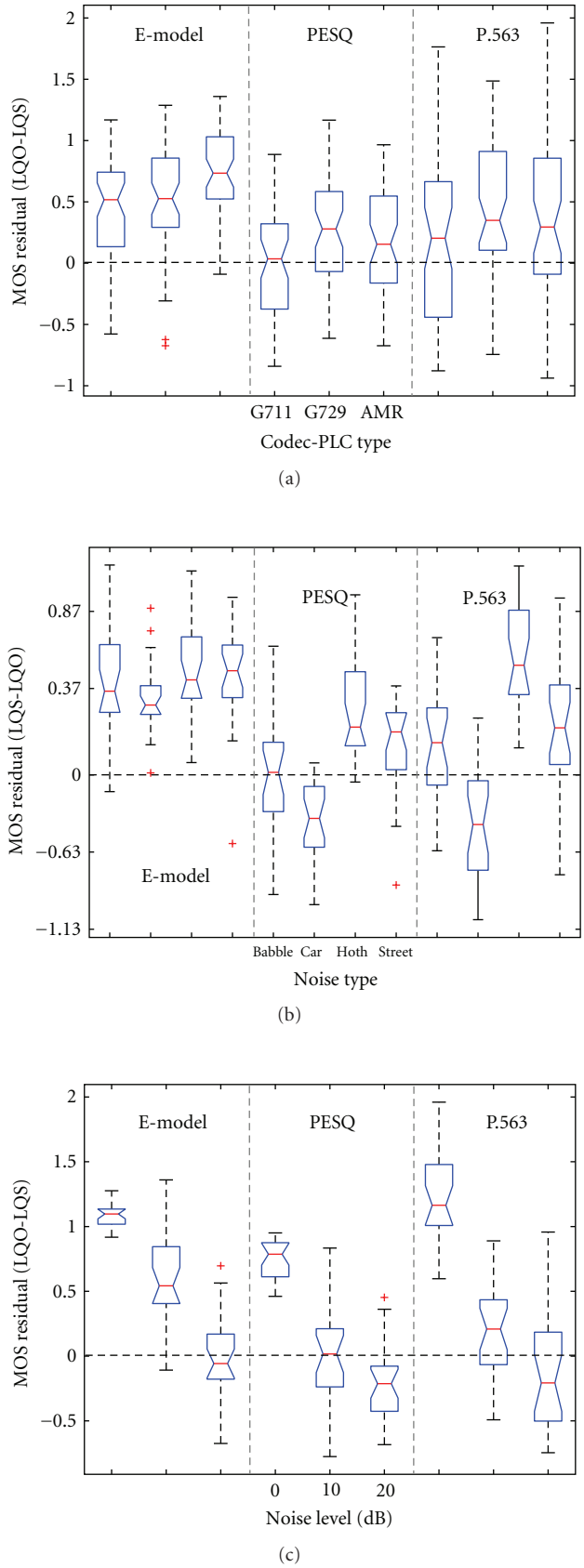


Figure 4: Significant main effects of (a) codec, (b) noise type, and (c) noise level on the accuracy of objective quality measurement of noise-suppressed speech.

Table 5: F-statistics (F) and p-values (P) of MOS residual errors (MOS-LQO/LQE minus MOS-LQS) obtained from factorial ANOVA with 95% confidence levels for noise-suppressed speech files.

Impairment	E-model		PESQ		P.563	
	F	P	F	P	F	P
Codec-PLC	5.5	.005	3.9	.02	1.23	.30
Noise type	1.6	.19	21.6	10^{-10}	33.5	10^{-14}
Noise level	94.2	0	70.3	0	80.5	0

As seen from the plots, E-model performance is inferior for AMR-processed speech. Both PESQ and P.563 underestimate MOS-LQS for car noise and overestimate MOS-LQS for hoth and street noise. Moreover, similar effects of noise level on estimation accuracy are observed for all three algorithms, with superior performance attained for speech corrupted by noise at higher SNR levels (10 dB and 20 dB). At low SNR (0 dB prior to noise suppression), all three algorithms overestimate MOS-LQS and PESQ attains superior performance.

Table 6 summarizes all significant main effects of wireless-VoIP impairments on PESQ, P.563, and E-model performance for both noisy and noise-suppressed degradation conditions. For noise-suppressed speech, packet loss rate effects were not included in the datasets hence are represented by the term "NI" in the table. As observed, PESQ and E-model performance are most sensitive to wireless-VoIP distortions.

4.1.2. Two-Way Interactions. Factorial ANOVA with a 95% confidence level has suggested four significant two-way interaction effects on noise-corrupted speech files:

(i) codec-PLC and packet loss rate (E-model, $F = 10.3$, $P = 0$),

(ii) codec-PLC and loss pattern (PESQ, $F = 3.4$, $P = 0.02$),

(iii) codec-PLC and noise type (E-model, $F = 9.1$, $P = 0$), and

(iv) codec-PLC and noise level (E-model, $F = 19.5$, $P = 0$).

Significant interaction effects were not observed for noise-suppressed speech. Figures 5(a) and 5(b) depict box and whisker plots that help illustrate the two-way interaction effects of codec-PLC and noise type as well as codec-PLC and noise level on E-model performance. Plots illustrating the remaining two-way interactions are omitted for brevity. As can be seen from the plots, inferior performance is attained for babble and car noise (and for SNR = 20 dB) with G.711-processed speech with the silence insertion packet loss concealment scheme (G711*). In such scenarios, perceptual artifacts are introduced due to the sudden changes in signal energy (i.e., temporal clippings); such artifacts are not accounted for by E-model quality estimates if the speech sample is additionally corrupted by noise. Other algorithms such as G.729 and AMR are equipped with comfort noise

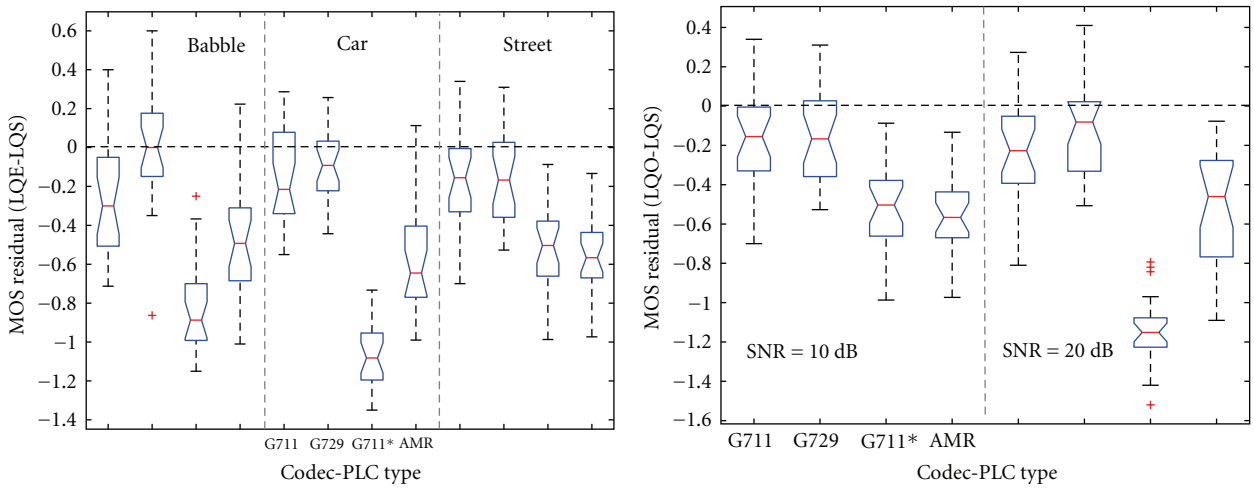


Figure 5: Significant two-way interactions of (a) codec and noise type and (b) codec and noise level on E-model accuracy.

Table 6: Significant main effects of wireless-VoIP degradation sources on PESQ, P.563, and E-model performance. A check mark (√) indicates significant deviations from subjective test data; "NI" stands for "not included" in the datasets.

Impairment	Noise-corrupted			Noise-suppressed		
	PESQ	P.563	E-model	PESQ	P.563	E-model
Codec-PLC	√	√	√	√	—	√
Packet loss rate	—	—	√	NI	NI	NI
Noise type	√	√	√	√	√	—
Noise level	√	—	√	√	√	√

generation capabilities which may be used to reduce the perceptual annoyance resultant from temporal clippings [43].

4.2. Analysis of Variance. In this section, we investigate the accuracy of the three algorithms with speech degraded under wireless-VoIP conditions by means of correlation (R) and root-mean-square error (RMSE) measures. The correlation between N MOS-LQS (y_i) and MOS-LQO (w_i) samples is computed using Pearson's formula

$$R = \frac{\sum_{i=1}^{N}(w_i - \overline{w})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(w_i - \overline{w})^2 \sum_{i=1}^{N}(y_i - \overline{y})^2}}, \quad (2)$$

where $\overline{w}$ is the average of w_i, and $\overline{y}$ is the average of y_i. The *RMSE*, in turn, is computed using

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(w_i - y_i)^2}{N}}. \quad (3)$$

Results in Table 7 are reported on a per-condition basis where MOS-LQS and MOS-LQO sample values are averaged over each degradation condition prior to computation of R and RMSE. For comparison, performance figures are reported before and after 3rd-order monotonic polynomial regression for P.563 and E-model. Moreover, as suggested by [44], PESQ performance is reported before and after the mapping described in [39]. Mappings are obtained for each dataset separately and the post mapping performance figures are represented by R^* and RMSE* in Table 7.

Using Fisher's z-test, PESQ performance is shown to be significantly different (with a 95% confidence level) from E-model and P.563 performance for the English dataset for both R and R^*. On the French dataset, however, PESQ performance is shown to be significantly different from E-model and P.563 only for R. Similarly, E-model and P.563 performances are only significantly different for R^*. Additionally, using Levene's test (here we assume MOS-LQO/MOS-LQE estimates are unbiased, thus RMSE values are treated as sample variances), it is observed that RMSE values are significantly different (95% confidence level) between E-model and PESQ, and between E-model and P.563 for both the English and the French datasets. For the English dataset, RMSE values between PESQ and P.563 are also shown to be significantly different. In terms of RMSE*, significant differences were only observed on the French dataset between P.563 and PESQ.

Overall, PESQ attains superior performance and P.563 and E-model attain comparable performance. In all cases, performance is substantially lower than that reported for traditional telephony applications (e.g., see [20, 21, 23]). The plots in Figures 6(a)–6(c) depict the overall per-condition

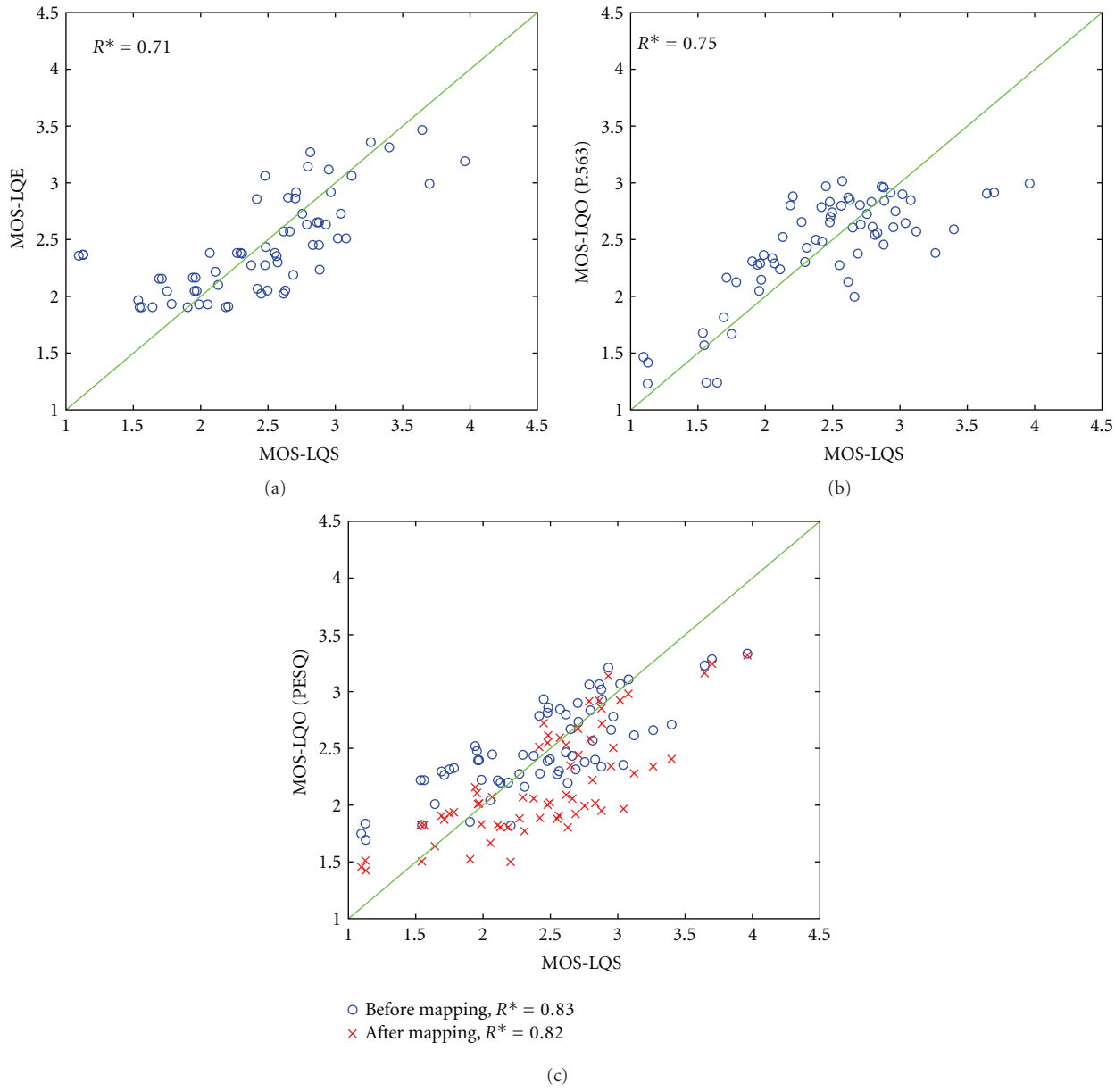


Figure 6: Per-condition MOS-LQO/LQE versus MOS-LQS for the overall dataset after 3rd-order polynomial mapping for (a) the E-model and (b) P.563, and (c) PESQ before ("∘") and after ("×") the mapping described in [39].

Table 7: Per-condition performance of E-model, PESQ, and P.563 on wireless-VoIP degradation conditions available in the English and French datasets. Post mapping performance is represented by R^* and RMSE*.

	E-model				PESQ				P.563			
	R	RMSE	R^*	RMSE*	R	RMSE	R^*	RMSE*	R	RMSE	R^*	RMSE*
English	0.69	0.70	0.71	0.38	0.83	0.29	0.82	0.47	0.72	0.52	0.75	0.41
French	0.71	0.65	0.73	0.42	0.78	0.38	0.77	0.46	0.74	0.50	0.78	0.37

MOS-LQO versus MOS-LQS for the English dataset obtained with the E-model, P.563, and PESQ, respectively. Plots (a) and (b) are after 3rd-order polynomial mapping and plot (c) depicts PESQ MOS-LQO before ("∘") and after ("×") the mapping described in [39]. As can be seen from the plots and from Table 7, PESQ performance decreases once the mapping is applied. This suggests that an alternate mapping function needs to be investigated for modern degradation conditions such as those present in wireless-VoIP communications.

5. Quantification of Overall Performance Loss

The comparisons described above suggest that the performance of three standard objective quality measurement algorithms is compromised for degradation conditions present in wireless-VoIP communications. To quantify the decrease in measurement accuracy, the conventional VoIP degraded speech data described in Section 3.3 is used as a benchmark. With the conventional VoIP impairment scenarios, standard algorithms are shown to perform reliably (e.g., see [23, 45]). For the benchmark data, it is observed that MOS-LQE estimates attain an average RMSE* of 0.21; that is, 48% lower than the average RMSE* reported in Table 7. PESQ and P.563 MOS-LQO estimates, in turn, attain average RMSE* values of 0.29 and 0.26, respectively; that is, approximately 35% lower than the average values reported in Table 7.

As observed, E-model performance is affected more severely by wireless-VoIP distortions. Such behavior is expected as the E-model is a parameter based measurement method and, as such, overlooks signal-based distortions that are not captured by the link parameters. As a consequence, improved performance is expected from hybrid signal-and-parameter based measurement schemes where signal based distortions are estimated from the speech signal and used to improve parameter based quality estimates. Hybrid measurement has been the focus of more recent quality measurement research (e.g., see [12, 32]).

6. Conclusions

We have investigated the effects of wireless-VoIP degradation on the performance of three state-of-the-art quality measurement algorithms: ITU-T PESQ, P.563 and E-model. Factorial analysis of variance tests has suggested that the performance of the aforementioned algorithms is sensitive to several degradation sources including background noise level, noise type, and codec-PLC type. Factorial analysis has also suggested several significant two-way interaction effects, in particular on E-model performance (e.g., codec and noise type or codec and noise level). Additionally, quantitative analysis has suggested that algorithm performance can be severely compromised and root-mean-square errors can increase by approximately 50% relative to conventional VoIP communications.

Acknowledgments

The authors would like to thank Drs. M. El-Hennawey, L. Thorpe, L. Ding, and R. Lefebvre for their vital support and the anonymous reviewers for their insightful comments.

References

[1] A. Takahashi, H. Yoshino, and N. Kitawaki, "Perceptual QoS assessment technologies for VoIP," *IEEE Communications Magazine*, vol. 42, no. 7, pp. 28–34, 2004.

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

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

[4] IETC RFC 3550, "RTP: a transport protocol for real-time applications," July 2003.

[5] IETC RFC 3551, "RTP profile for for audio and video conferences with minimal control," July 2003.

[6] IETC RFC 3611, "RTP control protocol extended reports (RTCP XR)," November 2003.

[7] S. Pennock, "Accuracy of the perceptual evaluation of speech quality (PESQ) algorithm," in *Proceedings of the International Conference on Measurement of Speech and Audio Quality in Networks (MESAQIN '02)*, January 2002.

[8] M. Varela, I. Marsh, and B. Gronvall, "A systematic study of PESQ's behavior (from a networking perspective)," in *Proceedings of the International Conference on Measurement of Speech and Audio Quality in Networks (MESAQIN '06)*, May 2006.

[9] S. R. Broom, "VoIP quality assessment: taking account of the edge-device," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 1977–1983, 2006.

[10] A. Clark, "VoIP performance management," in *Proceedings of the Internet Telephony Conference*, 2005.

[11] C. Nicolas, V. Gautier-Turbin, and S. Möller, "Influence of loudness level on the overall quality of transmitted speech," in *Proceedings of the 123rd Audio Engineering Society Convention (AES '07)*, December 2007.

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

[13] J. D. Gibson and B. Wei, "Tandem voice communications: digital cellular, VoIP, and voice over Wi-Fi," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '04)*, vol. 2, pp. 617–621, 2004.

[14] P. Perala and M. Varela, "Some experiences with VoIP over converging networks," in *Proceedings of the International Conference on Measurement of Speech and Audio Quality in Networks (MESAQIN '07)*, June 2007.

[15] M. Hatler, D. Phaneuf, and M. Ritter, "Muni wireless broadband: service oriented mesh-ups," ON World Report, July 2007.

[16] ITU-T Rec. P.805, "Subjective evaluation of conversational quality," April 2007.

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

[18] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*, Prentice-Hall, New York, NY, USA, 1988.

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

[20] A. W. Rix, M. P. Hollier, A. P. Hekstra, and J. G. Beerends, "PESQ, the new ITU standard for objective measurement of perceived speech quality—part 1: time alignment," *Journal of the Audio Engineering Society*, vol. 50, pp. 755–764, 2002.

[21] J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, "PESQ, the new ITU standard for objective measurement of perceived speech quality—part II: perceptual model," *Journal of the Audio Engineering Society*, vol. 50, pp. 765–778, 2002.

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

[23] L. Malfait, J. Berger, and M. Kastner, "P.563-The ITU-T standard for single-ended speech quality assessment," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 1924–1934, 2006.
[24] N. Johannesson, "The ETSI computation model: a tool for transmission planning of telephone networks," *IEEE Communications Magazine*, vol. 35, no. 1, pp. 70–79, 1997.
[25] ITU-T Rec. G.107, "The E-model, a computational model for use in transmission planning," 2005.
[26] ITU-T Study Group 12, "Question 8: E-model extension towards wideband transmission and future telecommunication and application scenarios," Study period: 2009–2012.
[27] ITU-T Rec. G.113, "Transmission impairments due to speech processing," 2001.
[28] ITU-T Rec. G.113 - Appendix I, "Provisional planning values for the equipment impairment factor Ie and packet loss robustness factor Bpl," 2002.
[29] ITU-T P.833, "Methodology for derivation of equipment impairment factors from subjective listening-only tests," 2001.
[30] ITU-T P.834, "Methodology for the derivation of equipment impairment factors from instrumental models," 2002.
[31] Telchemy, "Voice quality estimation in wireless and TDM environments," Application Note. Series: Understanding VoIP Performance, April 2006.
[32] L. Ding, Z. Lin, A. Radwan, M. S. El-Hennawey, and R. A. Goubran, "Non-intrusive single-ended speech quality assessment in VoIP," *Speech Communication*, vol. 49, no. 6, pp. 477–489, 2007.
[33] ITU-T Rec. G.191, "Software tools for speech and audio coding standardization," 2005.
[34] ITU-T Rec. G.711-Annex I, "A high quality low-complexity algorithm for packet loss concealment with G.711," 1996.
[35] 3GPP2 C.S0030-0, "Selectable mode vocoder (SMV) service option for wideband spread spectrum communication systems," January 2004.
[36] ITU-T P.810, "Modulated noise reference unit—MNRU," 1996.
[37] ITU-T P.830, "Subjective performance assessment of telephone-band and wideband digital codecs," 1996.
[38] ITU-T Rec. P. Supplement 23, "ITU-T coded-speech database," February 1998.
[39] ITU-T P.862.1, "Mapping function for transforming P.862 raw result scores to MOS-LQO," 2003.
[40] R. McGill, J. W. Tukey, and W. A. Larsen, "Variations of box plots," *The American Statistician*, vol. 32, no. 1, pp. 12–16, 1978.
[41] Ditech Networks, "Limitations of PESQ for measuring voice quality in mobile and VoIP networks—a white paper," December 2007.
[42] ITU-T Study Group 12 Temporary Document TD-42, "Requirements for a new model for objective speech quality assessment P.OLQA," June 2006.
[43] L. Ding, A. Radwan, M. S. El-Hennawey, and R. A. Goubran, "Measurement of the effects of temporal clipping on speech quality," *IEEE Transactions on Instrumentation and Measurement*, vol. 55, no. 4, pp. 1197–1203, 2006.
[44] ITU-T P.862.3, "Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2," 2005.
[45] A. W. Rix, M. P. Hollier, J. G. Beerends, and A. P. Hekstra, "PESQ—the new ITU standard for end-to-end speech quality assessment," in *Proceedings of the 109th Audio Engineering Society Convention (AES '00)*, pp. 1–18, Los Angeles, Calif, USA, September 2000.