Hybrid Signal-and-Link-Parametric Speech Quality Measurement for VoIP Communications

Tiago H. Falk, *Student Member, IEEE*, and Wai-Yip Chan

***Abstract*—A hybrid signal-and-link-parametric approach to speech quality measurement for voice-over-Internet protocol (VoIP) communications is described. Connection parameters are used to determine a base quality representative of the transmission link. Degradation factors, computed from perceptual features extracted from the decoded speech signal, are used to quantify distortions not captured by the connection parameters. The algorithm is tested on speech degraded by acoustic noise, temporal clippings, and noise suppression artifacts, thus simulating degradations present in wireless-VoIP tandem connections. Hybrid measurement is shown to overcome the limitations of pure link parametric and pure signal-based measurement methods, resulting in better measurement accuracy for modern VoIP communications. In addition, the proposed algorithm incurs modest computational overhead relative to pure link parametric measurement and attains up to 88% reduction in processing time relative to the ITU-T standard P.563 signal-based algorithm.**

***Index Terms*—Live call monitoring, quality impairments, speech distortions, speech quality measurement, voice-over-Internet protocol (VoIP).**

I. Introduction

VOICE-over-Internet protocol (VoIP) has increased in popularity over the past few years, mainly due to its low cost and capability of integrating data and real-time voice traffic on existing Internet protocol (IP) network infrastructures. IP networks are optimized for data communications, where variable losses and delay are not critical since retransmission can be performed. With voice communications, however, retransmission of missing packets is not a viable option and packet losses have become a major source of perceptual quality degradation. Commonly, packet loss concealment (PLC) strategies are used to replace lost packets and to improve speech quality. Nonetheless, VoIP links need to be constantly monitored such that necessary actions can be taken in order to maintain acceptable quality of service (QoS) [1].

For the purpose of real-time VoIP quality monitoring and control, automated objective speech quality measurement is required. Objective measurement replaces expensive and time-consuming subjective quality tests. Objective methods can be classified as either signal based or link parametric. Signal-based methods use perceptual features computed from the speech signal to estimate subjective quality. Link parametric methods, on the other hand, use connection parameters to estimate quality. For IP networks, connection parameters can include codec and PLC type, packet loss pattern (random or bursty), packet loss rate, jitter, and delay. Commonly, connection parameters are obtained from the real-time transport protocol (RTP) header [2], real-time transport control protocol (RTCP) [3], and RTCP extended reports (RTCP-XR) [4].

While signal-based algorithms perform well for traditional telephony applications, recent research has shown that algorithm performance may decrease when applied to VoIP communications. Several independent studies have shown that signal based schemes can be sensitive to connection parameters [5]–[8], as well as attain high "per-call" quality estimation errors [8]–[10]; such limitations restrict the usefulness of signal-based methods for online quality monitoring and control applications. Additionally, high computational complexity also acts as a major limiting factor for the widespread use of standard signal-based measures for online QoS control. As a consequence, link parametric methods have gained wide popularity in recent years. Studies have shown that signal based methods can be up to 1000 times more computationally complex (in terms of millions of instructions per second) than link parametric methods [10]. Representative link parametric models include ITU-T Recommendation G.107 (E-model) [11] and proprietary algorithms, such as Telchemy's VQmon [12].

Link parametric performance, however, can be severely affected by distortions that are not captured by connection parameters. Sources of such distortions can include acoustic noise, temporal clippings, and tandem connections with links that do not convey upstream equipment and signal conditions downstream (e.g., [13]). According to [14], [15], the number of wireless-VoIP tandem connections has grown substantially in the last few years, and it is just a matter of time before such connections become ubiquitous. With wireless-VoIP tandem conditions, speech signals are corrupted by varying levels and types of background noise prior to packetization. Moreover, advanced wireless communications standards such as the selectable mode vocoder (SMV) [16] are equipped with noise suppression capabilities. It is a known fact that noise suppression algorithms can introduce unwanted perceptual artifacts, such as the so-called "musical noise" phenomenon [17]. As will be shown here, background noise and artifacts introduced by noise suppression algorithms pose a serious threat to the performance of link parametric algorithms.

In this paper, a method that overcomes the limitations of both pure signal based and pure link parametric quality measurement is presented. A hybrid signal-and-link-parametric

Manuscript received September 10, 2007; revised July 19, 2008. Current version published October 17, 2008. This work was supported in part by the Natural Science and Engineering Research Council of Canada and Nortel. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yariv Ephraim.

The authors are with the Department of Electrical and Computer Engineering, Queen's University, Kingston, ON K7L-3N6, Canada (e-mail: tiago.falk@ece.queensu.ca; geoffrey.chan@queensu.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2008.2004524

approach to single-ended quality measurement of packet speech is proposed, based on extending the work described in [8]. The method makes use of IP connection parameters such as codec and PLC type, packet size, and packet loss pattern to determine a "base quality" representative of specific transmission links. Degradation factors are then computed from perceptual features extracted from the speech signal and are used to adjust the base quality accordingly. Our experiments focus on distortions that are not captured by connection parameters, such as those that occur in modern wireless-VoIP tandem connections. Distortions include VoIP speech codec impairments in combination with degradations caused by packet losses, temporal clippings, background acoustic noise, or noise suppression artifacts. The proposed method is shown to incur modest computational overhead relative to pure link parametric approaches. Combined with improved measurement accuracy, hybrid measurement is shown to be better suited for online monitoring of modern VoIP communications.

The remainder of this paper is organized as follows. In Section II, an overview of subjective and objective speech quality measurement is presented; previous hybrid approaches proposed in the literature are also discussed. Section III describes experiments which highlight the limitations of pure signal-based and link parametric approaches, thus serving as motivation for hybrid signal-and-link-parametric quality measurement. In particular, the sensitivity of P.563 to different VoIP network parameters is discussed. Section IV describes the architecture of the proposed algorithm, and experimental results are detailed in Section V. Algorithm computational complexity is discussed in Section VI, and conclusions are presented in Section VII.

II. Speech Quality Measurement

In this section, a brief overview of subjective speech quality measurement is given in Section II-A; objective speech quality measurement is discussed in Section II-B.

A. Subjective Measurement

Subjective speech quality measurement plays a key role in characterizing the quality of telecommunications products and services. It is known, for example, that the perceived quality of a speech signal processed by a novel speech coding algorithm, or transmitted over a novel network architecture, significantly influences the end user's experience with the system. Subjective speech quality assessment attempts to quantify this user experience. Moreover, the results of subjective evaluations can be used to define performance targets, to ensure appropriate product behavior, to define national and international standards, as well as to benchmark objective quality measures [18].

The majority of subjective measurement tests can be grouped into two classes: listening and conversational tests. Listening tests, as the name suggests, involve listeners "passively" rate (on a prespecified scale) the quality of the short-duration speech signal they have just heard. Conversational tests, on the other hand, are interactive and listeners are asked to rate the quality of a call based on the listening quality *and* on their ability to converse during the call. Conversational tests account for additional factors such as echoes and delays.

Listening quality tests are widely used by the telecommunications industry. The International Telecommunication Union (ITU-T) has published several recommendations describing guidelines for conducting subjective evaluations of listening quality in order to obtain reliable test results [19]–[21]. The most widely used subjective listening test method makes use of a listener panel to measure speech quality on an integer absolute category rating (ACR) scale ranging from 1 to 5, with 1 corresponding to bad speech quality and 5 corresponding to excellent speech quality. The average of the listener scores is termed the subjective mean opinion score (MOS). Current MOS terminology recommends the use of abbreviations MOS-LQS and MOS-CQS to distinguish between listening quality and conversational quality subjective MOS, respectively [21].

B. Objective Measurement

Despite being the most valid quality measurement method, subjective tests are expensive and time-consuming. Hence, for the purpose of real-time VoIP quality monitoring and control, objective speech quality measurement is required. Objective methods aim to deliver estimated quality scores that are highly correlated with the quality scores obtained from subjective listening experiments. Objective measurement methods can be classified as either signal based or link parametric. Signal-based approaches can be further classified as double-ended (also known as end-to-end) or single-ended, as described in the remainder of this section. Current MOS terminology recommends the use of abbreviations MOS-LQO and MOS-LQE to distinguish between listening quality MOS obtained from an objective model and E-model planning MOS estimates, respectively [21]. Similarly, abbreviations MOS-CQO and MOS-CQE are used for conversational quality. The focus of this paper will be on objective *listening* quality measurement.

1) Double-Ended Signal-Based Measurement: Double-ended measurement systems [Fig. 1(a)] are "comparison-based" and depend on some form of distance metric between the input (clean) and output (degraded) speech signals to estimate subjective quality. Representative algorithms include perceptual speech quality measure (PSQM) [22], measuring normalizing block (MNB) [23], [24], and statistical data mining quality assessment [25]. ITU-T Recommendation P.862, also known as perceptual evaluation of speech quality (PESQ), represents the current state-of-art double-ended algorithm for traditional telephony applications [26]. Recent research, however, has suggested decreased PESQ performance for VoIP communications and sensitivity to connection parameters such as speech codec and PLC type, packet size, packet loss rate, and packet loss pattern (e.g., [5], [6], [9]).

2) Single-Ended Signal-Based Measurement: Single-ended measurement systems [Fig. 1(b)] do not require access to a clean reference signal and commonly rely on models of normative speech behavior. In [27], vector quantizer (VQ) codebook representations of perceptual features of clean speech are used. In [28], VQ codebooks are replaced by Gaussian mixture probability models to improve quality measurement performance. Other proposed schemes have made use of vocal tract models [29] and spectro–temporal representations of normative speech

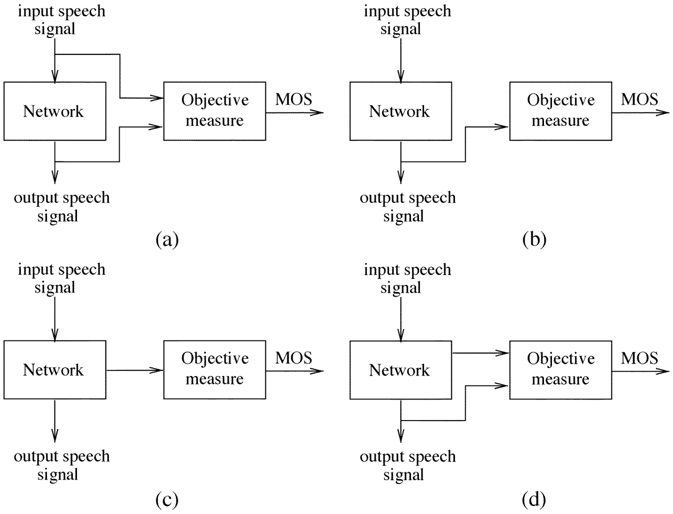


Fig. 1. Block diagram of (a) double-ended and (b) single-ended signal-based objective measurement, (c) link parametric measurement, and (d) hybrid signal-and-link parametric measurement.

behavior [30] for single-ended quality measurement. ITU-T Recommendation P.563 represents the current state-of-art single-ended algorithm for traditional telephony applications [31]. Recent research, however, has suggested that P.563 performance is also compromised for VoIP applications [6]–[8]. In addition, the experiments described in Section III also suggest algorithm sensitivity to different VoIP connection parameters and high per-call estimation errors.

3) Link Parametric Measurement: Link parametric models [Fig. 1(c)] make use of network parameters to estimate listening and/or conversational subjective quality. The E-model is a widely used transmission planning tool that describes several parametric models of specific network impairments and their interaction with subjective quality [11]. The basic assumption 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

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

where I_s, I_d, and $I_{e-\mathrm{eff}}$ represent speech transmission impairment factors (e.g., impairments due to quantization distortion), delay impairment factors (e.g., impairments due to echoes), and effective equipment impairment factors (e.g., impairments due to packet loss for different codec types), respectively. R_0 describes a base factor representative of the signal-to-noise ratio (SNR) and A an advantage factor. The R rating ranges from 0 (bad) to 100 (excellent) and can be mapped to MOS-CQE (if delay impairment factors are considered) or MOS-LQE using equations described in ITU-T Recommendation G.107 Annex B [11].

Over the years, an extensive list of equipment impairment factors has been derived [11], [32], [33]. In addition, ITU-T Recommendations P.833 [34] and P.834 [35] have been proposed to describe methodologies used to obtain equipment impairment factor values from subjective tests and instrumental models such as PESQ, respectively. More recently, new methodologies have been proposed to compute equipment impairment factors for wideband speech codecs [36]. 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. It is known, for example, that the simplifying assumption that impairments are additive in the perceptual domain does not hold true for high levels of "orthogonal" (unrelated) impairments. Proprietary algorithms, such as VQmon, use nonlinear impairment combination models that are shown to be more accurate when high levels of dissimilar impairments are present [37].

As will be shown here, extended E-model implementations provide accurate estimates for many VoIP scenarios, as well as attain low per-call quality estimation errors. Experiments described in Section III, however, suggest that link parametric performance can be severely affected by distortions that are not captured by connection parameters. Examples of such distortions include varying levels of acoustic noise, noise suppression artifacts, and temporal clippings. These shortcomings motivate the need for a hybrid signal-and-link-parametric methodology. Previously proposed hybrid architectures are described next; our proposed method is described in detail in Section IV.

4) Hybrid Measurement—Previous Investigations: Hybrid signal-and-link parametric measurement [Fig. 1(d)] uses link parameters in addition to the voice payload to estimate subjective quality. A few hybrid approaches have been proposed previously. In [38] and [39], PESQ is used to estimate the quality of the received speech signal and the estimated MOS-LQO is converted into an equipment impairment factor which, along with transmission delay estimates, is input to the E-model. While such approaches are useful to quickly obtain nontabulated equipment impairment factors, the high computational complexity, the need for a clean reference signal, and the sensitivity of PESQ to connection parameters make them impractical for online QoS control. Moreover, the use of PESQ for systems equipped with noise suppression algorithms is not recommended [40], thus limiting its usability in modern wireless-VoIP tandem conditions.

More recently, the work described in [41] proposes a hybrid methodology where temporal clippings and SNR are estimated from the degraded speech signal in a single-ended manner using complex signal-processing techniques. RTP and RTCP analysis is used to obtain the packet loss rate. Different impairment models are computed and combined with the E-model for a final quality rating. The algorithm is shown to correlate well with PESQ quality scores; however, due to the aforementioned PESQ limitations, it is not obvious if the method accurately predicts subjective quality. Moreover, the performance and complexity of the hybrid scheme is not compared to benchmark algorithms such as the E-model and P.563; thus, its improvement over existing algorithms is still unknown.

III. Motivation for Hybrid Signal-and-Link Parametric Speech Quality Measurement

In this section, experiments are described which motivate the need for hybrid signal-and-link parametric quality measurement. Experiments are carried out with a subjectively scored database, as described in Section III-A. The sensitivity of P.563 to different VoIP network parameters is investigated in Section III-B. Limitations of pure link parametric approaches, here characterized by the performance of an extended E-model implementation, are discussed in Section III-C.

A. Database Description

A bilingual (English and French) subjectively scored speech database is used in our experiments. The speech database contains a wide range of typically encountered VoIP scenarios. In particular, it comprises speech processed by G.711, G.729 and Adaptive Multi-Rate (AMR) codecs, with the latter operating at full rate (12.2 kb/s). The PLC used for G.711 is described in [42]; G.729 and AMR have their own built-in PLCs. Speech signals processed by G.711 with a simple silence insertion PLC scheme are also included. Packet durations of 10, 20, and 30 ms are used, except for AMR where only 20-ms packets are available. Random and bursty losses are simulated at 1, 2, 4, 7, and 10% with the ITU-T G.191 software package [43]; the Bellcore model is used for bursty losses. Losses are applied to speech packets, thus simulating a transmission network with voice activity detection (VAD).

To investigate the limitations of pure link parametric measurement methods, several signal-based distortions are generated in combination with codec distortions (with and without packet losses). Signal-based distortions include temporal clippings, acoustic background noise, and noise suppression artifacts. In order to maintain the simplifying E-model assumption that impairments are additive in the perceptual domain, low levels of noise and packet loss rates are used. Temporal clipping distortions are either manually generated by replacing the beginning of a talkspurt with a copy of the noise floor, or simulated by forcing VAD false negatives in the G.711 and G.729 codecs. Acoustic noise distortions are generated by corrupting clean speech with three noise types (hoth, babble, and car) at two SNR levels (10 and 20 dB). Noisy speech is then processed by the three aforementioned speech codecs (singly or in tandem) with and without packet losses. In the former scenario, random and bursty packet losses are simulated at 2% and 4%.

Noise suppression artifacts in combination with codec distortion are used to further simulate impairments introduced by wireless-VoIP tandem connections. Here, two noise suppression algorithms are tested. The first is the spectral subtraction algorithm available in the Adobe Audition software; a suppression factor of 75% is used. The second is the state-of-art noise suppression algorithm available as a preprocessing module in the SMV codec. Speech is corrupted by four noise types (Hoth, car, street, and babble) at three SNR levels (0, 10, and 20 dB). Noisy speech is processed by the noise suppression algorithms and the noise-suppressed signal is input to the G.711, G.729, or AMR speech codec.

The raw speech files were recorded in an anechoic chamber by four native Canadian French talkers and four native English talkers (half male and half female). Reference speech signals were filtered using the modified intermediate reference system (MIRS) send filter according to ITU-T Recommendation P.830 Annex D [20]. 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 [44], each speech file comprises two sentences separated by an approximately 650-ms pause.

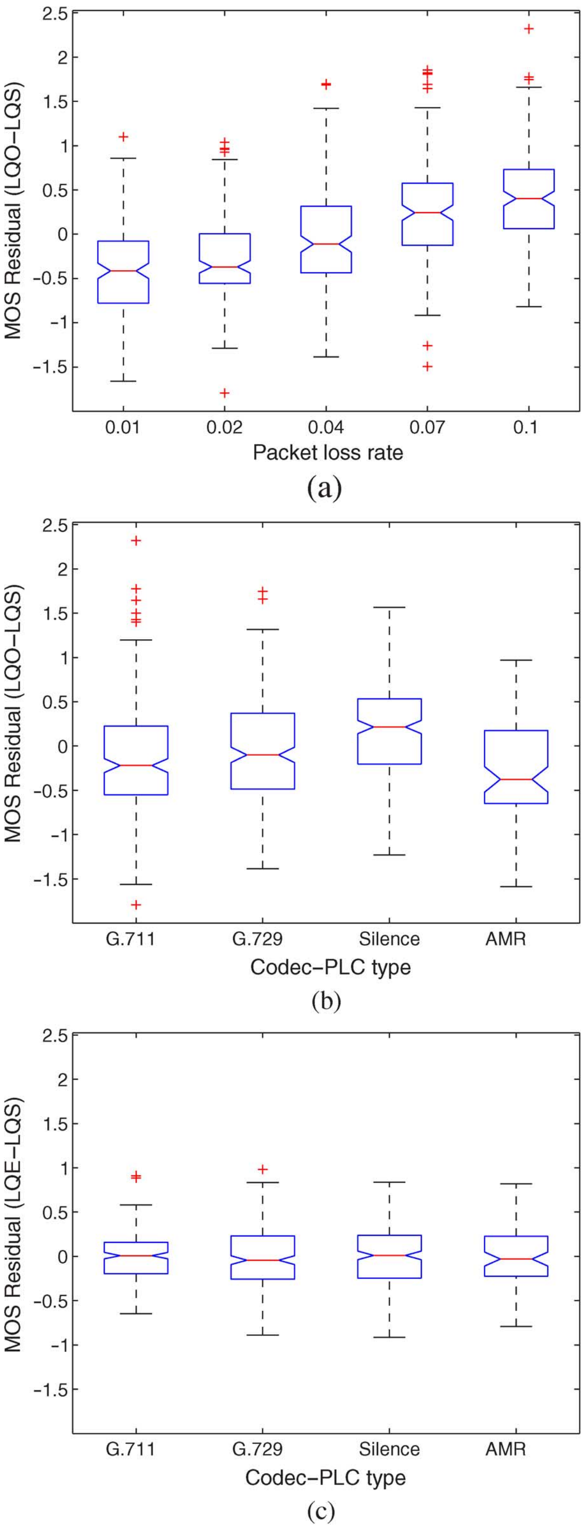
The subjective MOS test was conducted in 2006 following the requirements defined in [19] and [20]. Sixty listeners (native in each language; roughly half male and half female) participated in each listening quality test. The headphones used were Beyerdynamic DT 770 and the ambient noise level in the listening room was kept at around 27–28 dBA. A total of 300 degradation conditions are available per language. Of the 300 impairment conditions, 146 are due to packet losses, 21 to temporal clipping, 31 to acoustic noise *and* codec distortion, 54 to acoustic noise, codec distortion *and* packet losses, and 48 are due to noise suppression *and* codec distortion. The tests described in Section III-B make use of the 146 packet loss degradation conditions (total of 1168 speech files), and the tests in Section III-C make use of the 85 noisy speech conditions (total of 672 speech files).

B. Limitations of Pure Signal Based Measurement

In this section, statistical analysis is used to assess the relationship between P.563 performance and four connection parameters: codec-PLC type, packet size, packet loss pattern (random or bursty), and packet loss rate. For real-time quality monitoring and control applications, objective quality measures are required to attain low per-call estimation errors. Hence, we use per-call MOS residual as the performance criterion. MOS residual is given by MOS-LQO minus MOS-LQS (or MOS-LQE minus MOS-LQS) and is abbreviated as "LQO-LQS" (or "LQE-LQS") in Figs. 2–4. Analysis of variance suggests that two parameters have significant main impacts on P.563 accuracy, as described in Section III-B1. Two significant two-way interaction effects on P.563 accuracy are described in Section III-B2.

1) One-Way Interactions: One-way interaction analysis suggests that codec-PLC type and packet loss rate incur significant main effects on P.563 accuracy ($p < 0.0001$); the box and whisker plots depicted in Fig. 2(a) and (b) illustrate this behavior, respectively. The boxes have lines at the lower quartile, median, and upper quartile values; the whiskers extend to 1.5 times the interquartile range. Outliers (data with values beyond the ends of the whiskers) are represented by the symbol "+." From Fig. 2(a), it can be seen that P.563 performance is strongly dependent on packet loss rates. P.563 underestimates MOS-LQS for low loss rates and overestimates MOS-LQS for higher loss rates; MOS residuals greater than 2 MOS points are obtained at a 10% loss rate.

Fig. 2(b) suggests that P.563 attains high per-call estimation errors, in particular for the G.711 PLC scheme. According to [31], P.563 has only been validated for PLC schemes in CELP (codebook-excited linear prediction) codecs (e.g., 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 1.5 MOS point; on a five-point MOS scale, this can be the difference between having acceptable and unacceptable quality [9]. Moreover, the smallest median MOS residual

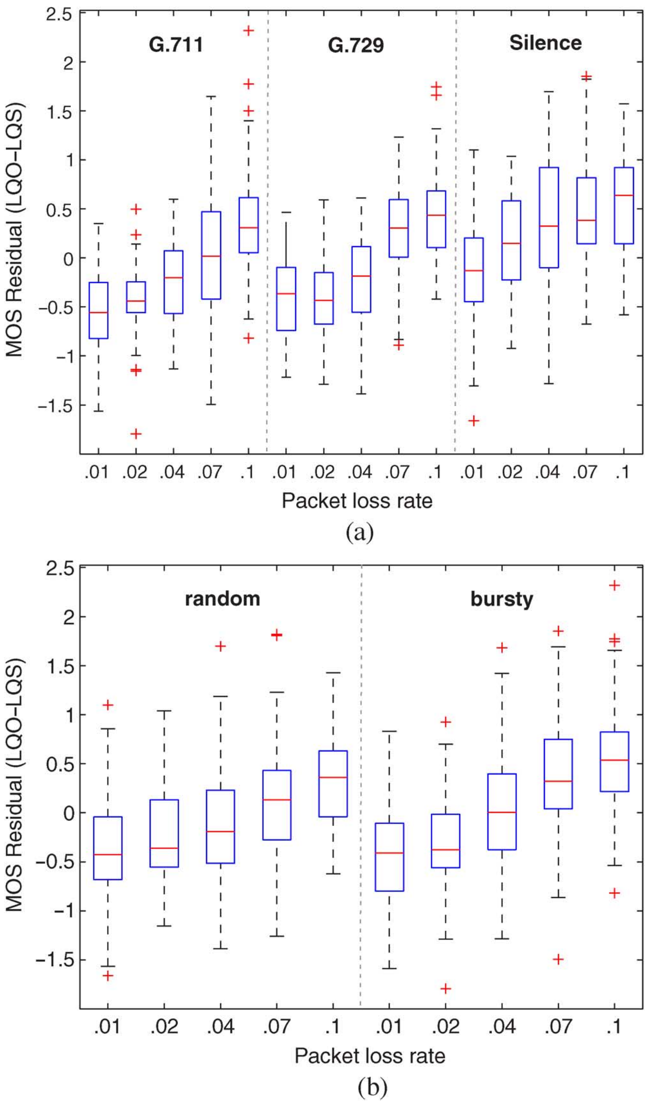


(a)

(b)

(c)

Fig. 2. Significant one-way effects of (a) packet loss rates and (b) codec-PLC type on P.563 accuracy. For comparison purposes, (c) depicts nonsignificant effects of codec-PLC type on extended E-model accuracy.



(a)

(b)

Fig. 3. Significant two-way interactions of (a) codec-PLC type and loss rate, and (b) loss rate and loss pattern on P.563 accuracy.

occurs with the simple G.711 silence insertion loss concealment scheme; this can be explained by the fact that P.563 is equipped with a temporal clipping detection module. As will be shown in Section III-B2; however, this does not hold true for high packet loss rates. For comparison purposes, Fig. 2(c) depicts the *nonsignificant* effects of speech codec-PLC type on extended E-model (described in Section III-C) performance. The substantially smaller residual errors validate the accuracy of the extended E-model implementation and corroborate the popularity of link parametric measurement for VoIP online quality monitoring.

2) Two-Way Interactions: Two-way statistical analysis has suggested two significant two-way interaction effects on P.563 performance: codec-PLC type and packet loss rate $(p < 0.007)$, and loss pattern and packet loss rate $(p < 0.003)$. Box and whisker plots depicted in Fig. 3(a) and (b) illustrate this behavior, respectively. From Fig. 3(a), it can be seen that P.563 underestimates MOS-LQS for low packet loss rates for both G.711 and G.729 codecs. The simple silence insertion scheme attains median residual values closer to zero, except for high packet loss rates (10%) where it attains the highest residual median value. The largest residual errors (outliers) occur for the G.711 codec under a 10% packet loss rate.

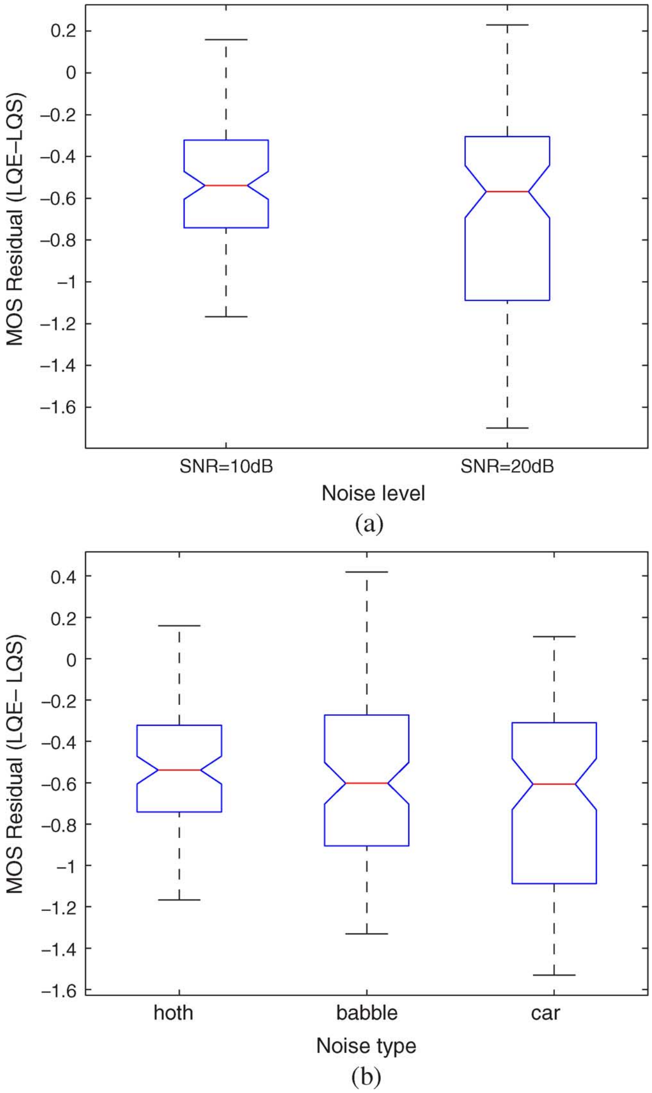


Fig. 4. Significant one-way effects of (a) noise level and (b) noise type on extended E-model accuracy.

Moreover, Fig. 3(b) suggests that P.563 accuracy varies less for random losses than for bursty losses. For low packet loss rates, median residual MOS values are similar for both random and bursty packet losses. For higher loss rates, bursty losses attain median residual MOS values almost one-quarter of a MOS point higher than random losses. Relative to link parametric measurement, P.563 is shown to be more sensitive to connection parameters and to attain higher per-call estimation errors.

C. *Limitations of Pure Link Parametric Measurement*

Parameters used in the E-model represent terminal, network, and environmental quality factors which are assumed to be known *a priori*. Extended E-model implementations propose to estimate E-model parameters (e.g., SNR) in real-time [41]. In this experiment, an extended E-model implementation is used. Nontabulated equipment impairment factors are obtained from subjectively scored speech data and the noise level is computed using the clean reference speech signals. Note that link parametric measurement is favored with this unrealistic assumption that true noise level information is available online. Commonly, only estimated noise levels are available, as in the experiment described in Section V. Here, statistical analysis is used to investigate the effects of noise level, noise type, and noise *and* packet loss on extended E-model measurement performance. The analysis suggests significant one-way interaction effects of noise level $(p < 0.006)$ and noise type $(p < 0.04)$; the box and whisker plots depicted in Fig. 4(a) and (b) illustrate this behavior, respectively.

From the plots, it can be observed that the extended E-model underestimates MOS-LQS and has a higher residual MOS variance for lower noise levels $(\text{SNR} = 20\text{ dB})$ and for babble and car noise. On the other hand, we observe that noise level does *not* show significant effects on P.563 accuracy. 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 (2500–3500 Hz) spectral flatness. It is observed, however, that P.563 performance is lower for babble and car noise, both of which have "low-pass" characteristics.

The experiments described in Sections III-B and III-C serve as motivation for a hybrid methodology which combines the strengths of both link parametric and signal-based approaches. In our hybrid approach, IP connection parameters are used to obtain a base quality representative of network connections characterized by each specific set of parameter values. Signal distortions not captured by the connection parameters, such as acoustic noise, are captured by perceptual features extracted from the speech signal. Next, a description of the proposed algorithm is given. It will be shown that, by using statistical models of normative speech codec behavior, a simple scheme can be devised that overcomes the limitations of pure signal-based and pure link parametric measurement methods.

IV. Architecture of Proposed Algorithm

The overall architecture of the proposed algorithm is depicted within the dotted lines in Fig. 5. Offline, E-model ratings and subjective listening tests are used to determine the base quality of several VoIP communications scenarios. As an embodiment of the proposed approach, we obtain base quality values for commonly used codec-PLC types with different packet sizes, under different packet loss patterns and packet loss rates. Base quality values are stored in a lookup table for fast online operation. Statistical models, in particular Gaussian mixture (GM) models, are designed using perceptual features extracted from speech signals processed by the various speech codecs operating under clean reference conditions. Reference GM model parameters, $\boldsymbol{\lambda}$ as defined in Section IV-B, are also stored in a lookup table for each codec. The speech codecs used in our experiments are described in Section III-A.

Online, IP header-extracted parameters are used to obtain the base quality and reference GM model parameters from lookup tables. Once packets are decoded and PLC is performed, the speech signal is level-normalized and filtered. Perceptual features are then extracted from the preprocessed test signal. The VAD labels the feature vector of each frame as either active or inactive. The extracted features are compared to stored models of normative codec operation behavior via a simple consistency

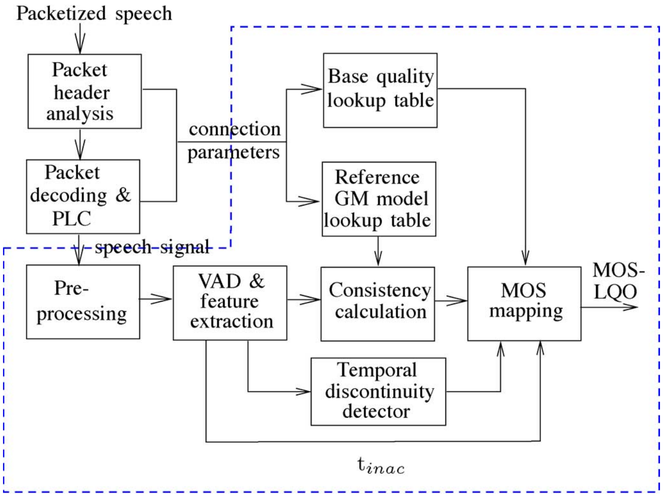


Fig. 5. Architecture of the proposed hybrid signal-and-link-parametric quality measurement algorithm.

measure. Temporal discontinuity detection is used to detect temporal clippings, an impairment which occurs commonly in VoIP communications [45]. Lastly, a MOS-mapping module is used to map the base quality, computed consistency measures, noise spectrum tilt, and detected temporal discontinuities to a final MOS-LQO. A more detailed description of each signal-based processing block is provided in the remainder of this section.

A. Preprocessing, VAD, and Feature Extraction

The preprocessing module performs level normalization and IRS filtering. The level of the speech signal is normalized to -26 dBov using the P.56 voltmeter [46] and the MIRS filter is applied to emulate the handsets used in listening tests. Voice activity detection is employed to label speech frames as active or inactive; the VAD from the G.729 codec [47] is used.

Perceptual linear prediction (PLP) cepstral coefficients [48] are extracted from the speech signal and serve as primary features. The coefficients are obtained from an "auditory spectrum," constructed to exploit three essential psychoacoustic properties: critical band analysis, equal-loudness preemphasis, and cubic-root compression. The auditory spectrum is approximated by an all-pole autoregressive model, whose coefficients are transformed to pth-order PLP cepstral coefficients $\mathbf{c} = \{c_i\}_{i=1}^{p}$. The zeroth cepstral coefficient c_0 is employed as an energy measure [49], and $p = 5$ is chosen from previous experiments [50]. When describing the PLP vector for a given frame m, the notation $\mathbf{c}_m = \{c_{i,m}\}_{i=1}^{p}$ and $c_{0,m}$ will be used.

Differential PLP cepstral coefficients are also used as a measure of signal spectral dynamics. In particular, delta and double-delta cepstral coefficients are used. Delta coefficients represent the local time derivatives (slope) of the cepstral sequence and are computed as [51]

$$\mathbf{\Delta c}_m = f(\mathbf{c}_m, L) = \sum_{l=-L}^{L} l\mathbf{c}_{m+l} \qquad (2)$$

where the normalization factor $\sum_{l=-L}^{L} l^2$ is omitted as it does not affect the simulation results. Delta coefficients indicate the rate of change (speed) of spectral components; in our simulations $L = 5$ is used. Double-delta coefficients are the second-order local time derivatives of the cepstral sequence and are computed using (2) as $\mathbf{\Delta}^2\mathbf{c}_m = f(\mathbf{\Delta c}_m, L = 3)$. Double-delta coefficients indicate the acceleration of the spectral components.

Lastly, motivated by the results described in Section III-C, noise-related features are extracted. Pilot experiments are carried out with noise spectral flatness and noise spectrum tilt; the latter (henceforth referred to as $\mathrm{t}_{\mathrm{inac}}$) resulted in superior performance and is used throughout the remainder of this paper. As in [52], $\mathrm{t}_{\mathrm{inac}}$ is approximated by the first-order linear prediction coefficient averaged over inactive speech frames.

B. Gaussian Mixture Reference Models

Reference models of normative codec behavior are designed for commonly used speech codecs. For active speech frames, GM models are trained for PLP cepstral coefficients appended with delta and double-delta coefficients, i.e., $\mathbf{x}_{act,m} = [c_{0,m}, \mathbf{c}_m, \mathbf{\Delta c}_m, \mathbf{\Delta}^2\mathbf{c}_m]$. For inactive speech frames, GM models are obtained from PLP cepstral coefficients, i.e., $\mathbf{x}_{\mathrm{inac},m} = [c_{0,m}, \mathbf{c}_m]$. A Gaussian mixture model is a weighted sum of M component densities

$$p(\mathbf{u}|\boldsymbol{\lambda}) = \sum_{i=1}^{M} \alpha_i b_i(\mathbf{u}) \qquad (3)$$

where $\alpha_i \geq 0$, $i = 1, \ldots, M$ are the mixture weights, with $\sum_{i=1}^{M} \alpha_i = 1$, and $b_i(\mathbf{u})$ are K-variate Gaussian densities with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\mathbf{\Sigma}_i$. The parameter list, $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_M\}$, defines a particular Gaussian mixture density, where $\boldsymbol{\lambda}_i = \{\boldsymbol{\mu}_i, \mathbf{\Sigma}_i, \alpha_i\}$. In our experiments, diagonal covariance matrix Gaussian components are used.

The expectation–maximization (EM) algorithm [53] is commonly used to obtain $\boldsymbol{\lambda}$ from training data. The disadvantage of the EM approach is that the number of Gaussian components M has to be determined *a priori*. Large M may result in a model that overfits the training data, whereas small M may result in models that are not accurate. Commonly, the approach taken is to try several values of M and use the one that results in superior performance on an unseen test set [8]. Here, an alternate approach is taken and a recursive (greedy) EM algorithm is used. Greedy EM implementations estimate model parameters and the number of components simultaneously.

We experiment with a pruning approach which starts with a large number of components and uses a stochastic approximation recursive learning algorithm to prune irrelevant components [54], and a progressive approach which starts with a single component and adds components sequentially [55]. In order to guard against overfitting, the largest admissible M is chosen such that the training ratio (ratio between number of scalar parameters that need to be estimated during training and the number of training samples available) is maintained above an empirically set value of 100. In our experiments, it is observed that both implementations result in similar performance and $M = 32$ is chosen for active frames and $M = 4$ for inactive frames.

C. Consistency Calculation

In [8], signal-based distortions were measured by means of the Kullback–Leibler distance between offline-trained reference models and GM models obtained online from the test speech signal. While the approach was shown to be effective, real-time operation becomes questionable as speech packets need to be stored in order for accurate online GM models to be obtained. Here, for the benefit of low computational complexity and real-time operation, a reference-model consistency measure is used. For a given speech signal, the consistency $\mathcal{C}_{\text{class}}$ between the observation and the reference GM models is defined as the normalized (log-)likelihood

$$\mathcal{C}_{\text{class}}(\mathbf{X}_{\text{class}}) = \frac{1}{N_{\text{class}}} \sum_{j=1}^{N_{\text{class}}} \log \left(p_{\text{class}}(\mathbf{x}_{\text{class},j}|\boldsymbol{\lambda})\right) \quad (4)$$

where $\mathbf{X}_{\text{class}} = \{\mathbf{x}_{\text{class},i}\}_{i=1}^{N_{\text{class}}}$ denotes the set of all N_{class} feature vectors that have been classified as belonging to a given speech class (active or inactive). Normalization is required as N_{class} varies for different test signals.

D. Temporal Discontinuity Detection

Temporal discontinuities (also known as "clippings") are a known source of quality degradation in VoIP systems [45]. Front-end, midspeech (short mutes), and back-end clippings may occur due to erroneous VAD decisions, erroneous line echo cancelation decisions, or simple silence insertion PLC schemes. Here, a simple energy-thresholding scheme is proposed and temporal discontinuities are detected by evaluating abrupt changes in c_0.

From our experiments, abrupt stops (back-end clippings) can be accurately detected if

$$\mathcal{T}_i = \frac{c_{0,i+1}}{c_{0,i}} < 0.35.$$

Abrupt starts (front-end clippings) are detected if $\mathcal{T}_i > 2.02$. Experiments on our databases show that with this simple energy-thresholding scheme, approximately 98% of front-end clippings are correctly classified. On the other hand, approximately 10% of "normal" abrupt starts, such as those experienced with certain plosive consonants (e.g.,/d/), are misclassified as clippings. To improve classification performance, more complex machine learning methods can be used [50]. Since abrupt starts have, intuitively, less significant impact on perceived speech quality [30], [56], such classification errors are shown not to be detrimental to overall speech quality measurement. For the sake of reduced computational complexity, the simple energy-thresholding scheme is used in the experiments described in Section V. Lastly, midspeech clippings are detected when an abrupt stop is followed by an abrupt start during speech activity. The mute length is estimated from the number of consecutive frames for which $\mathcal{T} \simeq 1$.

Previous studies suggest that frequency of occurrence and midspeech clipping duration are two major factors affecting subjective quality [56]. We define frequency of occurrence as the ratio of the number of detected discontinuities over active speech duration (clips/second); frequency of occurrence is computed for front- and back-end clippings (f_f and f_b, respectively). For midspeech clippings, subjective tests suggest that similar quality is attained for high occurrence of short mutes and low occurrence of long mutes [56]. As a consequence, frequency of occurrence is computed for midspeech clippings of short duration ($\text{f}_{\text{mid}-s}$) and long duration ($\text{f}_{\text{mid}-l}$). Mutes between 10–70 ms are classified as short duration and mutes between 70–260 ms as long duration.

E. MOS Mapping

Machine learning tools are used to devise an accurate mapping between the base quality (MOS_0), computed consistency measures, noise spectrum tilt, and clipping frequency-of-occurrence to a final MOS-LQO. Here, a support vector (SV) regressor [57], trained on subjectively scored data, is used. The input to the MOS mapping module is the eight-dimensional feature vector consisting of

$$\mathbf{z} = [\text{MOS}_0, \mathcal{C}_{\text{act}}, \mathcal{C}_{\text{inac}}, \text{t}_{\text{inac}}, \text{f}_f, \text{f}_b, \text{f}_{\text{mid}-s}, \text{f}_{\text{mid}-l}].$$

A subset of the ITU-T Supplement 23 (experiment 3) database [44], along with material from three other proprietary databases, are used to train the MOS mapping module. The Supplement 23 subset includes speech processed by the G.729 codec (singly or in tandem conditions) with random and bursty losses at 3% and 5%. Clean and noisy conditions (street, hoth and vehicle noise at an SNR = 20 dB) are included. Proprietary databases include temporal-clipped speech material, speech processed by the G.711 codec with 3% random packet losses, and noisy speech. The latter includes speech degraded by car and street noise ($\text{SNR} = 15$ dB) and processed by the SMV codec, operating at full and half rate (8.5 and 4 kb/s, respectively), with 1% random losses. A total of 2672 speech samples are used to train the MOS mapping function.

V. Experimental Results

The proposed hybrid signal-and-link parametric measurement algorithm is tested on a subset (1232 speech samples) of the corpus described in Section III-A. The subset includes 154 impairment conditions covering temporal clipping, noise *and* codec distortion, noise *and* packet losses, and noise suppression *and* codec distortion. Hence, the test set covers distortions which are not captured by connection parameters, such as those present in modern wireless-VoIP tandem connections. We emphasize that degradation conditions and speech files available in the test set are distinct from those in the training set, thus are unseen to the proposed algorithm. Comparisons are carried out with P.563 and the extended E-model.

To the best of our knowledge, equipment impairment factor values for noise suppression algorithms are not available. In fact, artifacts introduced by such enhancement schemes are dependent on noise type and noise levels. As a consequence, pure link parametric measurement is performed in a manner similar to that of [41], where the SNR is estimated online and used in the extended E-model rating. Here, the SNR is estimated with the P.563 "noise analysis" module described in [31]. While such an approach is useful to quantify noise artifacts that remain after enhancement, it does not account for distortions that arise during speech activity. The proposed hybrid scheme overcomes this

TABLE I
PER-CONDITION PERFORMANCE OF HYBRID, PURE LINK PARAMETRIC (EXTENDED E-MODEL), AND PURE SIGNAL-BASED (P.563) MEASUREMENT. RESULTS ARE REPORTED BEFORE AND AFTER THIRD-ORDER POLYNOMIAL REGRESSION

	Hybrid (Proposed)		Link (Extended E-model)				Signal (P.563)			
	ρ	ϵ	ρ	$\%\rho_{link}$	ϵ	$\%\epsilon_{link}$	ρ	$\%\rho_{signal}$	ϵ	$\%\epsilon_{signal}$
before	0.813	0.329	0.696	38.5	0.665	50.5	0.701	37.5	0.494	33.4
after	0.821	0.301	0.712	37.8	0.352	14.5	0.720	36.1	0.407	26.0

limitation by computing consistency measures for both active and inactive speech segments.

Correlation (ρ) and root-mean-square error (ϵ) are used as algorithm figures of merit. Correlation between MOS-LQS and MOS-LQO (or MOS-LQE) is obtained via Pearson's formula

$$\rho = \frac{\sum_{i=1}^{N}(w_i - \bar{w})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(w_i - \bar{w})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{5}$$

where $\bar{w}$ is the average of w_i (which represents the MOS-LQS), and $\bar{y}$ is the average of y_i (which represents MOS-LQO or MOS-LQE). The error ϵ is given by

$$\epsilon = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(w_i - y_i)^2}. \tag{6}$$

We report improvement in correlation incurred by the proposed scheme over pure signal-based or link parametric measurement by the so-called "ρ-improvement" measure ($\%\rho_{\text{pure}}$)

$$\%\rho_{\text{pure}} = \frac{\rho_{\text{hybrid}} - \rho_{\text{pure}}}{1 - \rho_{\text{pure}}} \times 100\% \tag{7}$$

where the subscripts "hybrid" and "pure" represent the proposed hybrid scheme and either pure signal-based or pure link parametric measurement algorithms, respectively. The ρ-improvement measure indicates the percentage reduction of the performance gap to perfect correlation. Improvement in ϵ attained by using the proposed algorithm is reported by means of the conventional percentage reduction in root-mean-square error ($\%\epsilon_{\text{pure}}$).

As recommended in [31], results in Table I are reported on a per-condition basis where condition-averaged MOS-LQS and condition-averaged MOS-LQO (or MOS-LQE) are used to estimate ρ and ϵ. Results are reported before and after third-order monotonic polynomial regression. As can be seen, the proposed method improves on pure link parametric measurement by approximately 38% and 15% in terms of ρ and ϵ, respectively (post third-order mapping). Improvements of approximately 36% and 26% (ρ and ϵ, respectively) are attained relative to pure signal-based measurement. For comparison purposes, PESQ attains $\rho = 0.831$ and $\epsilon = 0.458$ with the mapping described in [58]. Thus, the hybrid single-ended scheme offers somewhat lower ϵ than the state-of-art double-ended standard algorithm. Note, however, that the usage of PESQ is not recommended for systems that include a noise suppression algorithm. Furthermore, statistical analysis shows that noise type and noise level have *insignificant* effects on the performance of the proposed hybrid scheme.

TABLE II
PER-CALL ROOT-MEAN-SQUARE ERROR ϵ OF HYBRID, PURE LINK PARAMETRIC AND PURE SIGNAL-BASED MEASUREMENT

	Hybrid	Link	Signal
ϵ	0.513	0.689	0.587
$\%\epsilon$	–	25.6	12.6

As mentioned previously, for online quality monitoring applications, per-call residual MOS error is an important performance metric. Per-call ϵ is reported in Table II for extended E-model, P.563, and the proposed hybrid scheme. As can be seen, the proposed algorithm attains reductions in per-call ϵ of approximately 13% and 26% relative to pure signal-based measurement and pure link parametric measurement, respectively. From Table II, it can also be observed that under noisy and wireless-VoIP tandem conditions, P.563 attains smaller per-call residual errors than the extended E-model, thus corroborating the fact that pure link parametric measurement is compromised for degradations not captured by connection parameters.

Furthermore, careful analysis of the quality estimates obtained by the proposed scheme has suggested that per-call estimation errors are, on average, 23% larger for noise-suppressed speech relative to noisy and temporal-clipped speech. This behavior is expected as quantifying perceptual distortions of noise-suppressed speech is a difficult task and has been the focus of current objective [59] and subjective quality measurement research [60]. With the advance of signal-based measurement for noise-suppressed speech, improved hybrid measurement performance is expected.

VI. ALGORITHM COMPUTATIONAL COMPLEXITY

In this section, algorithm computational complexity is discussed in Section VI-A. To further reduce complexity, a VAD-integrated processing scheme is proposed and described in Section VI-B.

A. Computational Complexity

As mentioned previously, pure link parametric measurement has gained widespread use due to its low computational complexity. As a consequence, it is important to measure the computational overhead incurred by the signal-based branch of the proposed algorithm. We use algorithm processing time as a measure of computational complexity. Processing time is defined as the time it takes to process ten speech files randomly selected

TABLE III
Algorithm Processing Time for P.563 and the Proposed Hybrid Scheme, With and Without VAD-Integrated Processing

Algorithm	Proc. time (s)	%↓
P.563	26.25	–
Proposed	12.47	52.5
Proposed (VAD-integrated)	3.06	88.3

from the test set described in Section V. With the exception of the VAD algorithm (taken from the ANSI-C G.729 reference implementation), the proposed algorithm is implemented using Matlab version 7.2 Release 2006a. Simulations are run on a PC with a 2.8-GHz Pentium 4 processor and 2 GB of RAM. The ten files combined have a total length of 62.57 s. Algorithm processing times for the Matlab implementation of the proposed algorithm and the ANSI-C reference implementation of P.563 are reported in Table III. The column labeled "% ↓" describes the percentage reduction in processing time obtained by the proposed scheme relative to P.563. As can be seen, a reduction in processing time of approximately 53% is attained; note that a complete C implementation of the proposed algorithm would further increase the speedup.

B. VAD-Integrated Processing

An in-depth analysis of the computational complexity of each operational module depicted in Fig. 5 shows that over 75% of the algorithm processing time is attributable to voice activity detection. To further reduce processing time, the proposed algorithm can take advantage of the fact that in most VoIP codec implementations, VAD decisions are transmitted by the encoder and are readily available at the decoder. Moreover, in the event of a lost packet, VAD decisions are predicted by the decoder based on previously received packets. Hence, the hybrid scheme can reuse these VAD decisions in lieu of recomputing them. To investigate the gains obtained with such "VAD-integrated" processing, we use the Matlab implementation of the G.723.1 speech codec described in [61] where inactive frames are detected as "null frames" in the G.723.1 bitstream. Table III also exhibits the gains obtained with the VAD-integrated hybrid quality measurement scheme. As can be seen, an overall reduction in processing time of approximately 88% can be attained relative to P.563.

VII. Conclusion

A hybrid signal-and-link-parametric quality measurement algorithm for packet speech is proposed. Experiments described herein serve to demonstrate the gains obtained by combining the strengths of pure signal-based and pure link parametric quality measurement paradigms to devise a more comprehensive quality measurement scheme. The proposed hybrid methodology improves on pure link parametric approaches by measuring distortions that are not captured by connection parameters. Furthermore, lower per-call estimation errors are attained relative to pure signal based measurement. Additionally, the proposed scheme is shown to have modest computational overhead relative to pure link parametric measurement, and when operated in an integrated manner, can attain computational complexity that is 88% lower than the ITU-T standard P.563 algorithm. Moderate computational complexity, low per-call estimation errors, and the ability to account for distortions not captured by connection parameters are valuable attributes for online speech quality monitoring, in particular for modern VoIP communications.

Acknowledgment

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

References

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

[2] "RTP: A transport protocol for real-time applications," IETC RFC 3550, Jul. 2003.

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

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

[5] S. Pennock, "Accuracy of the perceptual evaluation of speech quality (PESQ) algorithm," in *Proc. Int. Conf. Meas. Speech Audio Quality Netw.*, Jan. 2002.

[6] M. Varela, I. Marsh, and B. Gronvall, "A systematic study of PESQ's behavior (from a networking perspective)," in *Proc. Int. Conf. Meas. Speech Audio Quality Netw.*, May 2006.

[7] L. Ding, A. Radwan, M. El-Hennaway, and R. Goubran, "Performance study of objective voice quality measures in VoIP," in *Proc. Symp. Comput. Commun.*, Jul. 2007, pp. 197–202.

[8] T. H. Falk, H. Yuan, and W.-Y. Chan, "A hybrid signal-and-link-parametric approach to single-ended quality measurement of packetized speech," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2007, vol. 4, pp. 841–844.

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

[10] A. Clark, "VoIP performance management," in *Proc. Internet Telephony Conf.*, 2005.

[11] "The E-Model, a computational model for use in transmission planning," Int. Telecom. Union, ITU-T Rec. G.107, 2005.

[12] A. Clark, "Modeling the effects of burst packet loss and recency on subjective voice quality," in *Proc. IP Telephony Workshop*, 2001.

[13] S. Dimolitsas, F. Corcoran, and C. Ravishankar, "Impact of network routing on the end-to-end transmission quality of cellular and PCS connections," *IEEE Trans. Veh. Technol.*, vol. 47, no. 4, pp. 1227–1233, Nov. 1998.

[14] J. Gibson and B. Wei, "Tandem voice communications: Digital cellular, VoIP, and voice over Wi-Fi," in *Proc. Global Telecomm. Conf.*, 2004, pp. 617–621.

[15] P. Perala and M. Varela, "Some experiences with VoIP over converging networks," in *Proc. Int. Conf. Meas. Speech Audio Quality Netw.*, Jun. 2007.

[16] "Selectable mode vocoder (SMV) service option for wideband spread spectrum communication systems," 3GPP2 C.S0030-0, Jan. 2004.

[17] Y. Ephraim, D. Malah, and B.-H. Juang, "On the application of hidden Markov models for enhancing noisy speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 12, pp. 1846–1856, Dec. 1989.

[18] L. Thorpe, "Subjective evaluation of speech compression codes and other non-linear voice-path devices for telephony applications," *Int. J. Speech Technol.*, vol. 2, pp. 273–288, 1999.

[19] "Methods for subjective determination of transmission quality," Int. Telecom. Union, ITU-T P.800, 1996.

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

[21] "Mean opinion score (MOS) terminology," Int. Telecom. Union, ITU-T P.800.1, 2003.

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

[23] 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, Jul. 1999.

[24] S. Voran, "Objective estimation of perceived speech quality—Part II: Evaluation of the measuring normalizing block technique," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 4, pp. 383–390, Jul. 1999.

[25] W. Zha and W.-Y. Chan, "Objective speech quality measurement using statistical data mining," *EURASIP J. Appl. Signal Process.*, vol. 2005, no. 9, pp. 1410–1424, Jun. 2005.

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

[27] J. Liang and R. Kubichek, "Output-based objective speech quality," in *Proc. IEEE Veh. Technol. Conf.*, Jun. 1994, vol. 3, pp. 1719–1723.

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

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

[30] 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.

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

[32] "Transmission impairments due to speech processing," Int. Telecom. Union, ITU-T Rec. G.113, 2001.

[33] "Provisional Planning Values for the Equipment Impairment Factor IE and Packet Loss Robustness Factor BPL," Int. Telecom. Union, ITU-T Rec. G.113—Appendix I, 2002.

[34] "Methodology for derivation of equipment impairment factors from subjective listening-only tests," Int. Telecom. Union, ITU-T P.833, 2001.

[35] "Methodology for the Derivation of Equipment Impairment Factors From Instrumental Models," Int. Telecom. Union, ITU-T P.834, 2002.

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

[37] "Voice quality estimation in wireless and TDM environments," Telchemy, Application Note. Series: Understanding VoIP Performance, Apr. 2006.

[38] C. Hoene, S. Wietholter, and A. Wolisz, "Predicting the perceptual service quality using a trace of VoIP packets," in *Proc. Int. Workshop Quality of Future Internet Services*, 2004, pp. 21–30.

[39] L. Sun and E. Ifeachor, "New methods for voice quality evaluation for IP networks," in *Proc. Int. Teletraffic Congr.*, Sep. 2003, pp. 1201–1210.

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

[41] L. Ding, Z. Lin, A. Radwan, M. El-Hennaway, and R. Goubran, "Non-intrusive single-ended speech quality assessment in VoIP," *Speech Commun.*, vol. 49, no. 6, pp. 477–489, Jun. 2007.

[42] "A High Quality Low-Complexity Algorithm for Packet Loss Concealment With G.711," Int. Telecom. Union, ITU-T Rec. G.711-Annex I, 1996.

[43] "Software tools for speech and audio coding standardization," Int. Telecom. Union, ITU-T Rec. G.191, 2005, , .

[44] "ITU-T Coded-Speech Database," Int. Telecom. Union, ITU-T Rec. P. Supplement 23, 1998.

[45] S. Voran, "Perception of temporal discontinuity impairments in coded speech—Proposal for objective estimators and some subjective test results," in *Proc. Int. Conf. Meas. Speech Audio Quality Netw.*, May 2003.

[46] "Objective measurement of active speech level," Int. Telecom. Union, ITU-T P.56, 1993.

[47] "A Silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70," Int. Telecom. Union, ITU-T Rec. G.729–Annex B, 1996.

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

[49] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing*. Upper Saddle River: Prentice-Hall, 2001.

[50] 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.

[51] J. Picone, "Signal modeling techniques in speech recognition," *Proc. IEEE*, vol. 81, no. 9, pp. 1215–1247, Sep. 1993.

[52] V. Grancharov, J. Samuelsson, and W. B. Kleijn, "Noise-dependent postfiltering," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Jan. 2004, vol. 1, pp. 457–460.

[53] A. Dempster, N. Lair, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc.*, vol. 39, pp. 1–38, 1977.

[54] Z. Zivkovic and F. van der Heijden, "Recursive unsupervised learning of finite mixture model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 5, pp. 651–656, May 2004.

[55] N. Vlassis and A. Likas, "A greedy EM algorithm for Gaussian mixture learning," *Neural Process. Lett.*, vol. 15, pp. 77–87, 2002.

[56] J. Gruber and L. Strawczynski, "Subjective effects of variable delay and speech clipping in dynamically managed voice systems," *IEEE Trans. Commun.*, vol. COM-33, no. 8, pp. 801–808, Aug. 1985.

[57] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer, 1995.

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

[59] Y. Hu and P. Loizou, "Evaluation of objective measures for speech enhancement," in *Proc. Int. Conf. Spoken Lang. Process.*, 2006, pp. 1447–1450.

[60] "Subjective Test Methodology for Evaluating Speech Communication Systems That Include Noise Suppression Algorithms," Int. Telecom. Union, ITU-T P.835, 2003.

[61] P. Kabal, "ITU-T G.723.1 Speech Coder: A Matlab Implementation," McGill Univ., Montreal, QC, Canada, Aug. 2004, Tech. Rep.

Tiago H. Falk (S'00) was born in Recife, Brazil, in September 1979. He received the B.Sc. degree from the Federal University of Pernambuco, Recife, in 2002, and the M.Sc. (Eng.) degree from Queen's University, Kingston, ON, Canada, in 2005, all in electrical engineering. He is currently pursuing the Ph.D. degree at Queen's University.

His research interests include multimedia quality measurement and enhancement, multimedia coding and communications, biomedical signal processing, pattern recognition, and communication theory.

Mr. Falk is recipient of several research excellence awards, including the IEEE Kingston Section Ph.D. Research Excellence Award (2008), the Best Student Paper Awards at the International Conference on Acoustics, Speech, and Signal Processing (2005) and the International Workshop for Acoustic Echo and Noise Control (2008), and the Prof. Newton Maia Young Scientist Award (2001). Mr. Falk has also received several prestigious scholarships, most notably the NSERC Canada Graduate Scholarship (2006) and the Harvard–LASPAU Organization of the American States Graduate Scholarship (2003). Mr. Falk is also a student member of the International Speech Communication Association where he has served as President of the Student Advisory Committee (2007–2008).

Wai-Yip Chan received the B.Eng. and M.Eng. degrees from Carleton University, Ottawa, ON, Canada, and the Ph.D. degree from University of California, Santa Barbara, all in electrical engineering.

He is currently with the Department of Electrical and Computer Engineering, Queen's University, Kingston, ON, Canada. He has held positions with the Communications Research Centre Canada, Bell Northern Research (Nortel), McGill University, and Illinois Institute of Technology. His current research interests are speech quality measurement and enhancement, and multimedia coding and communications. He is an Associate Editor of *EURASIP Journal on Audio, Speech, and Music Processing*. He has helped organizing IEEE sponsored conferences in speech coding, image processing, and communications.

Dr. Chan received the CAREER Award from the U.S. National Science Foundation.