Single-Ended Speech Quality Measurement Using Machine Learning Methods

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

***Abstract*—We describe a novel single-ended algorithm constructed from models of speech signals, including clean and degraded speech, and speech corrupted by multiplicative noise and temporal discontinuities. Machine learning methods are used to design the models, including Gaussian mixture models, support vector machines, and random forest classifiers. Estimates of the subjective mean opinion score (MOS) generated by the models are combined using hard or soft decisions generated by a classifier which has learned to match the input signal with the models. Test results show the algorithm outperforming ITU-T P.563, the current "state-of-art" standard single-ended algorithm. Employed in a distributed double-ended measurement configuration, the proposed algorithm is found to be more effective than P.563 in assessing the quality of noise reduction systems and can provide a functionality not available with P.862 PESQ, the current double-ended standard algorithm.**

***Index Terms*—Mean opinion score (MOS), objective quality measurement, quality model, single-ended measurement, speech communication, speech distortions, speech enhancement, speech quality, subjective quality.**

I. Introduction

THE TELECOMMUNICATIONS industry is going through a phase of rapid development. New services and technologies emerge continuously. The plain old telephone system is being replaced by wireless and voice-over-internet protocol (VoIP) networks. Service providers are faced with offering speech services over increasingly heterogenous network connections. Identifying the root cause of speech quality problems has become a challenging task. The evaluation of speech quality, consequently, has become critically important, serving as an instrument for network design, development, and monitoring, and also for improvement of quality of service. Despite all of the advances in modern telecommunication networks, speech quality measurement has remained costly and labor intensive.

Speech quality is a subjective opinion, based on the user's reaction to the speech signal heard. A common subjective test method makes use of a listener panel to measure speech quality on an integer absolute category rating (ACR) scale ranging from one to five, with one corresponding to bad speech quality and five corresponding to excellent speech quality. The average of the listener scores is the subjective mean opinion score (MOS) [1]. This has been the most reliable method of speech quality assessment but it is very expensive and time consuming, making it unsuitable for "on-the-fly" applications. Objective measurement methods, which replace the listener panel with a computational algorithm, have been the focus of more recent quality measurement research. Objective methods can be implemented by either software or hardware and can be embedded into network nodes for real-time monitoring and control. Objective methods aim to deliver estimated MOSs that are highly correlated with the MOSs obtained from subjective listening experiments. Current MOS terminology recommends the use of the abbreviations MOS-LQS and MOS-LQO to distinguish between "listening quality" subjective MOS and "listening quality" objective MOS, respectively [2].

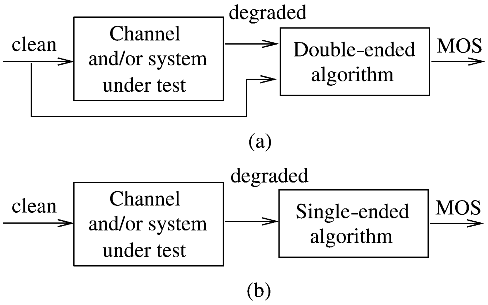


Fig. 1. Block diagram of (a) double-ended and (b) single-ended speech quality measurement.

Algorithms for objective quality measurement can be classified as double- or single-ended [Fig. 1(a) and (b), respectively]. Double-ended algorithms depend on some form of distance metric between the input (clean) and output (degraded) speech signals to estimate MOS-LQS. Double-ended schemes often have two underlying requirements: 1) that the input signal be of high quality, i.e., clean, and 2) that the output signal be of quality no better than the input. These requirements prohibit the use of double-ended algorithms in some scenarios where the input is degraded and the system being tested is equipped with a speech enhancement algorithm. On the other hand, single-ended algorithms do not depend on a reference signal and are invaluable tools for monitoring speech quality of in-service systems and networks. Moreover, multiple single-ended "probes" can be distributed throughout the network to pinpoint locations where quality degradations occur.

Double-ended quality measurement has been studied since the early 1980s [3]. Earlier methods were implemented to assess the quality of waveform-preserving speech coders;

Manuscript received February 1, 2006; revised July 17, 2006. This work was supported in part by Natural Science and Engineering Research Council of Canada, Ontario Centres of Excellence, and Nortel. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Doh-Suk Kim.

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 Figs. 3–6 are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2006.883253

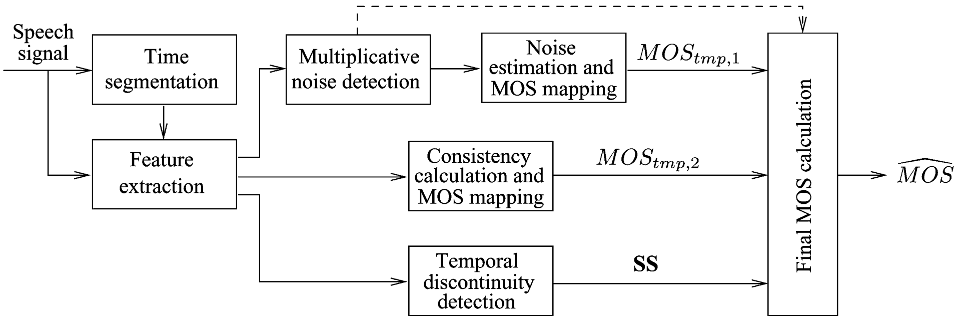

Speech signal
Time segmentation
Feature extraction
Multiplicative noise detection
Noise estimation and MOS mapping
$MOS_{tmp,1}$
Consistency calculation and MOS mapping
$MOS_{tmp,2}$
Temporal discontinuity detection
SS
Final MOS calculation
$\widehat{MOS}$

Fig. 2. Architecture of the proposed single-ended measurement algorithm.

representative measures include signal-to-noise ratio (SNR) and segmental SNR [4]. More sophisticated measures (e.g., [5]) were proposed once low-bit-rate speech coders, which may not preserve the original signal waveform, were introduced. More recently, quality measurement research has focused on algorithms that exploit models of human auditory perception. Representative algorithms include Bark spectral distortion (BSD) [6], perceptual speech quality measure (PSQM) [7], measuring normalizing block (MNB) [8], [9], and statistical data mining quality assessment [10]. The International Telecommunications Union ITU-T P.862 standard, also known as perceptual evaluation of speech quality (PESQ), represents the current "state-of-art" double-ended algorithm [11].

On the contrary, single-ended measurement is a more recent research field. In [12], comparisons between features of the received speech signal and vector quantizer codebook representations of the features of clean speech are used to estimate speech quality. In [13], the degree of consistency between features of the received speech signal and Gaussian mixture probability models (GMMs) of normative clean speech behavior are used as indicators of speech quality. The works described in [14] and [15] make use of vocal tract models and modulation-spectral features derived from the temporal envelope of speech, respectively, as quality cues for single-ended quality measurement. The ITU-T standard P.563 represents the current "state-of-art" single-ended algorithm [16].

This paper presents two novel approaches to speech quality measurement. First, a GMM-based single-ended algorithm is proposed. The algorithm exploits innovative techniques to detect and measure the amount of multiplicative noise and to detect temporal discontinuities in the test signal. Second, the proposed single-ended algorithm is applied to form a more flexible double-ended measurement architecture. The proposed scheme, as opposed to current double-ended algorithms, can be applied to systems with noisy inputs and/or speech enhancement systems.

The remainder of this paper is organized as follows. In Section II, a detailed description of the single-ended algorithm is given. Algorithm design considerations are covered in Section III, and algorithm performance is evaluated in Section IV. The proposed double-ended measurement architecture is detailed in Section V, followed by the conclusion in Section VI.

II. Algorithm Description

A. Overview

In the proposed method, single-ended measurement algorithms are designed based on the architecture depicted in Fig. 2. Perceptual features are first extracted from the test speech signal every 10 ms. The time segmentation module labels the feature vector of each frame as belonging to one of three possible classes: active-voiced, active-unvoiced, or inactive (background noise). Signals are then processed by a multiplicative noise detector. During design, the detector is optimized in conjunction with the "noise estimation and MOS mapping" and the "consistency calculation and MOS mapping" modules. A preliminary quality score, namely $\mathrm{MOS}_{\mathrm{tmp},1}$, is computed from the estimated amount of multiplicative noise present in the signal. A second preliminary score, $\mathrm{MOS}_{\mathrm{tmp},2}$, is computed from six consistency measures, which in turn, are calculated relative to reference models of speech behavior. We note that $\mathrm{MOS}_{\mathrm{tmp},1}$ is shown to provide more accurate speech quality estimates, relative to $\mathrm{MOS}_{\mathrm{tmp},2}$, for certain degradation conditions. The objective of the multiplicative noise detector is, thus, to distinguish which conditions can be better represented by $\mathrm{MOS}_{\mathrm{tmp},1}$. Lastly, temporal discontinuities ($\mathbf{SS}$) are detected and a final quality rating ($\widehat{\mathrm{MOS}}$) is computed. The final rating is a linear combination of the preliminary scores adjusted by the negative effects temporal discontinuities have on perceived quality. A detailed description of each block is provided in the remainder of this section. Experimental optimization of algorithm parameters is presented in Section III-B

B. Time Segmentation and Feature Extraction

Time segmentation is employed to separate the speech frames into different classes. It has been shown that each class exerts different influence on the overall speech quality [13]. Time segmentation is performed using a voice activity detector (VAD) and a voicing detector. The VAD identifies each 10-ms speech frame as being active or inactive (background noise). The voicing detector further labels active frames as voiced or unvoiced. Here, the VAD from the adaptive multirate (AMR) speech codec [17] (VAD option 1) and the voicing determination algorithm described in [18] are used.

Perceptual linear prediction (PLP) cepstral coefficients [19] serve as primary features and are extracted from the speech signal every 10 ms. The coefficients are obtained from an "auditory spectrum," constructed to exploit three essential psychoacoustic precepts. First, the spectrum of the original signal is warped into the Bark frequency scale and a critical band masking curve is convolved with the signal. The signal is then pre-emphasized by a simulated equal-loudness curve to match the frequency magnitude response of the ear. Lastly, the amplitude is compressed by the cubic-root to match the nonlinear relation between intensity of sound and perceived loudness. The auditory spectrum is then approximated by an all-pole autoregressive model, whose coefficients are transformed to pth order PLP cepstral coefficients $\mathbf{x} = \{x_i\}_{i=0}^{p}$. The zeroth cepstral coefficient x_0 is employed as an energy measure [20] in our scheme. When describing the PLP vector for a given frame m, the notation $\mathbf{x}_m = \{x_{i,m}\}_{i=0}^{p}$ is used. Moreover, the PLP vector averaged over N_f frames $(\bar{\mathbf{x}})$ is given by

$$\bar{\mathbf{x}} = \frac{1}{N_f}\sum_{m=1}^{N_f}\mathbf{x}_m. \tag{1}$$

The order of the autoregressive model determines the amount of detail in the auditory spectrum preserved by the model. Higher order models tend to preserve more speaker-dependent information and are more complex to calculate. We experiment with fifth- and tenth-order PLP coefficients. On our databases, both models incur similar quality estimation performance; thus, for the benefit of lower computational complexity, fifth-order PLP coefficients are chosen. Fifth-order models have been successfully used in [12] and are shown in [19] to serve well as speaker-independent speech spectral parameters. Moreover, dynamic features in the form of delta and double-delta coefficients [20] have been shown to indicate the rate of change (speed) and the acceleration of the spectral components, respectively [21]. As will be shown in Section II-E, the delta information for the zeroth cepstral coefficient can be used to detect temporal discontinuities.

Lastly, the mean cepstral deviation $(\bar{\sigma})$ of a test signal is computed. In Section II-C, it will be shown that $\bar{\sigma}$ can be used to detect and estimate the amount of multiplicative noise. The mean cepstral deviation is the average of all "per-frame" deviations (σ_m) of the PLP coefficients (excluding the zeroth coefficient). The per-frame deviation is defined as

$$\sigma_m = \sqrt{\frac{1}{p-1}\sum_{i=1}^{p}\left(x_{i,m} - \left(\frac{1}{p}\sum_{j=1}^{p}x_{j,m}\right)\right)^2} \tag{2}$$

and $p = 5$.

C. Detecting and Estimating Multiplicative Noise

It is known that multiplicative noise (also known as speech-correlated noise) can be introduced by logarithmically companded PCM (e.g., G.711) or ADPCM (e.g., G.726) systems as well as by other waveform speech coders [22]. In fact, modulated noise reference unit (MNRU) [23] was originally devised to reproduce the perceptual distortion of log-PCM waveform coding techniques. MNRU systems produce speech that is corrupted by controlled speech-amplitude-correlated noise. The speech plus multiplicative noise output $y(n)$ of an MNRU system is given by

$$y(n) = s(n) + s(n)10^{-Q/20}N(n) \tag{3}$$

where $s(n)$ is the clean speech signal, and $N(n)$ is white Gaussian noise (unit variance). The amount of multiplicative noise $s(n)10^{-Q/20}N(n)$ is controlled by the parameter Q, which represents the ratio of input speech power to multiplicative noise power, and is expressed in decibels (dB). This parameter is often termed the "Q value."

Measuring multiplicative noise of the form (3), when *both* the clean signal and the degraded speech signals are available, is fairly straightforward. The task becomes more challenging when the original clean signal is unavailable. In such instances, Q must be estimated. To the best of our knowledge, the scheme presented in [16] is the only published method of estimating multiplicative noise using only the degraded speech signal. The process entails an evaluation of the spectral statistics of the signal during active speech periods.

Today, MNRU degradations and reference waveform codecs such as G.711 and G.726 are used extensively as "anchor" conditions in testing and standardization of emerging codec technologies and in network planning. Current speech quality measurement algorithms should handle such degradation conditions efficiently. In previous work [24], estimating multiplicative noise is shown to be beneficial for GMM-based speech quality measurement. A multiplicative noise estimator, similar to the one described in [16], was deployed and performance improvement was reported for MNRU degradations. This improvement in performance substantiates the need for an efficient method of estimating multiplicative noise. Here, an innovative and simple technique is employed.

The technique is based on PLP coefficients and their mean cepstral deviations. As discussed in [25], the multiplicative noise term in (3) introduces a fairly flat noise floor in regions of the spectrum of $y(n)$, where the power of $s(n)$ is small. On the other hand, in regions where the power of the input signal is sufficiently large, the spectrum of $s(n)$ is almost perfectly preserved (see examples in [25]). The amount of multiplicative noise is controlled by the parameter Q. As a result, as Q approaches 0 dB (i.e., power of multiplicative noise equals power of input speech), the flat spectral characteristic of the multiplicative noise starts to dominate the spectrum of $y(n)$. In such instances, information about the spectral envelope of the signal is lost, deteriorating the quality and intelligibility of the signal. To illustrate this behavior, Fig. 3(a)–(c) shows the spectrum of a speech frame prior to processing and after MNRU degradation with $Q = 25$ dB, and $Q = 5$ dB, respectively. As can be clearly seen, the spectrum of $y(n)$ becomes flatter as the amount of multiplicative noise increases.

The use of mean cepstral deviation as a measure of the amount of multiplicative noise present in a signal is inspired by the definition of cepstrum—the inverse Fourier transform of the log-spectrum of a signal [20]. Tests on our databases show that the cepstral deviation for MNRU speech correlates well with the flatness of the log-spectrum, i.e., with amount of multiplicative noise. As an example, a correlation of -0.93 is attained between

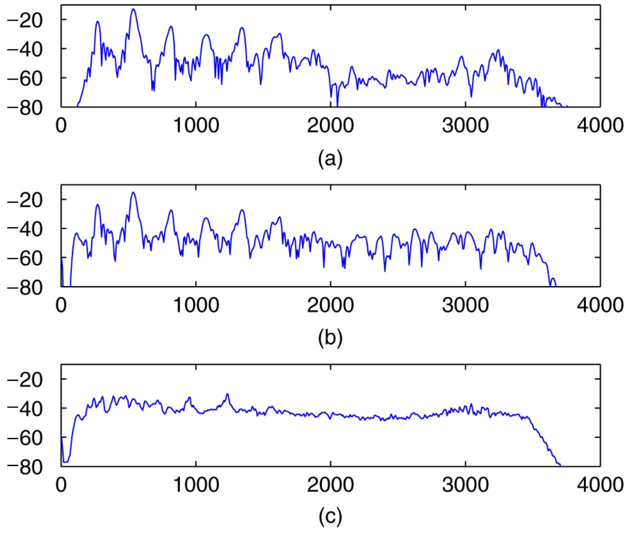


Fig. 3. Spectrum of a speech frame (a) before processing, and after (b) 25-dB MNRU and (c) 5-dB MNRU processing. The x-axis represents frequencies in hertz and the y-axis amplitudes in decibels.

the mean cepstral deviation of active speech frames ($\bar{\sigma}_{\text{active}}$) and Q values for MNRU-degraded speech files on our speech databases. Negative correlation is expected since lower Q values result in flatter spectra. In turn, spectrum and cepstrum are related via a Fourier transformation, thus a flat spectrum translates into a nonflat cepstrum, i.e., a high $\bar{\sigma}_{\text{active}}$. Once Q is estimated, $\text{MOS}_{\text{tmp},1}$ can be computed via simple regression. In fact, a polynomial mapping can be employed directly between $\bar{\sigma}_{\text{active}}$ and $\text{MOS}_{\text{tmp},1}$. As will be shown in Section III-B, $\text{MOS}_{\text{tmp},1}$ provides accurate estimates of perceived subjective quality for various different degradation conditions, in addition to corruption by MNRU multiplicative noise.

In this paper, the detection of the presence of high levels of multiplicative noise is treated as a supervised classification problem. In fact, the detector is trained to detect not only multiplicative noise, but also all other degradation conditions where $\text{MOS}_{\text{tmp},1}$ is better than $\text{MOS}_{\text{tmp},2}$ as an estimator of MOS-LQS (some example conditions are given in Section III-B-3). Detection is performed on a "per-signal" basis and depends on a 14-dimensional input consisting of the PLP vector averaged over active frames ($\bar{\mathbf{x}}_{\text{active}}$) and over inactive frames ($\bar{\mathbf{x}}_{\text{inactive}}$), and the mean cepstral deviation for active frames ($\bar{\sigma}_{\text{active}}$) and for inactive frames ($\bar{\sigma}_{\text{inactive}}$). Inactive frames are used as they provide cues for discriminating additive background noise from speech-correlated noise. Experiments are carried out with support vector classifiers (SVCs) [26], classification and regression trees (CARTs) [27], and random forests (RFs) [28] as candidate detectors. Training of the detectors will be described in more detail in Section III-B-3.

D. Consistency Calculation and MOS Mapping

Gaussian mixture models are used to model the PLP cepstral coefficients of each of the three classes of speech frames—voiced, unvoiced, and inactive. A Gaussian mixture density is a weighted sum of M component densities $p(\mathbf{u}|\boldsymbol{\lambda}) = \sum_{i=1}^{M} \alpha_i b_i(\mathbf{u})$, 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 μ_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 $\boldsymbol{\lambda}_i = \{\boldsymbol{\mu}_i, \mathbf{\Sigma}_i, \alpha_i\}$.

Here, a modification to the GMM-based architecture described in [13] is proposed. It is found that accuracy can be enhanced if the algorithm is also equipped with information regarding the behavior of speech degraded by different transmission and coding schemes [24]. To this end, clean speech signals are used to train three different Gaussian mixture densities, $p_{\text{clean,class}}(\mathbf{u}|\lambda)$. The subscript "class" represents either voiced, unvoiced, or inactive frames. For the degradation model, $p_{\text{degraded,class}}(\mathbf{u}|\boldsymbol{\lambda})$ are trained.

For the benefit of low computational complexity, we make a simplifying assumption that vectors between frames are independent. This assumption has been shown in [13] to provide accurate speech quality estimates. Nonetheless, improved performance is expected from more sophisticated models, such as hidden Markov models, where statistical dependencies between frames can be considered. This investigation, however, is left for future study. Thus, for a given speech signal, the consistency between the observation and the models is defined as the normalized (log-)likelihood

$$c_{\text{model,class}}(\mathbf{X}_{\text{class}}) = \frac{1}{N_{\text{class}}} \times \sum_{j=1}^{N_{\text{class}}} \log(p_{\text{model,class}}(\mathbf{x}_{\text{class},j}|\boldsymbol{\lambda})) \quad (4)$$

where $\mathbf{X}_{\text{class}} = \{\mathbf{x}_{\text{class},i}\}_{i=1}^{N_{\text{class}}}$ denotes the set of all N_{class} PLP vectors that have been classified as belonging to a given speech *class*. The subscript "model" represents either the clean or the degradation reference model. Normalization is required as N_{class} varies for different test signals.

In total, six consistency measures are calculated per test signal. For each class, the product of the consistency measure (4) and the fraction of frames of that class in the speech signal is computed; this product is referred to as a "feature." In the rare case when the fraction of frames of a specific class is zero (e.g., only voiced speech is detected), a constant $c_{\text{model,class}} = c = -15$ is used as the feature. Lastly, the six features are mapped to $\text{MOS}_{\text{tmp},2}$. We experiment with multivariate polynomial regression and multivariate adaptive regression spline (MARS) [29] as candidate mapping functions. With MARS, the mapping is constructed as a weighted sum of truncated linear functions (see [10] for more detail). On our databases, MARS is shown to provide superior performance. MARS models are designed based on the MOS-LQS of degraded speech. Simulation results show that a simple MARS function composed of a linear combination of 18 truncated linear functions provides accurate quality estimation performance. The experimental results presented in Section IV make use of a MARS model to map the six-dimensional consistency feature vector, calculated on a per-signal basis, into $\text{MOS}_{\text{tmp},2}$.

E. Temporal Discontinuity Detection

Motivated by the results reported in [30] and by first- and second-order methods used for edge detection in images (e.g.,

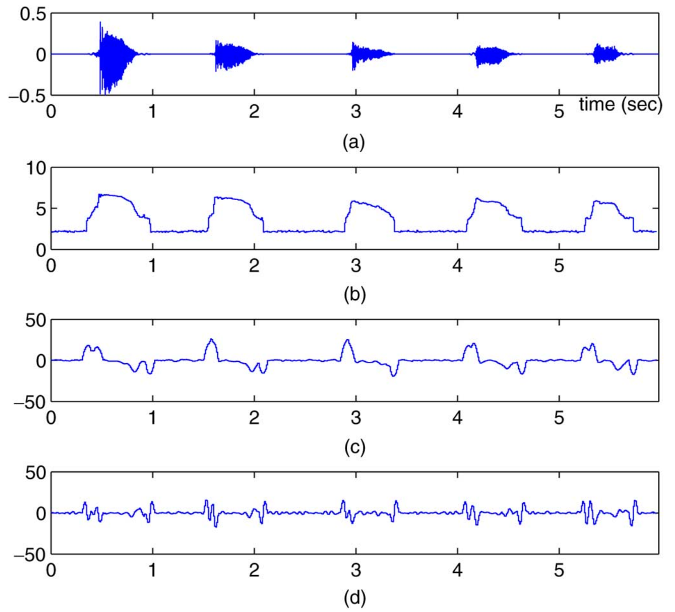


Fig. 4. Analysis of a signal's (a) waveform, (b) x_0, (c) Δ_0, and (d) Δ_0^2. Signal consists of five vowels uttered in a noisy office environment.

[31]), we employ delta and double-delta coefficients for temporal discontinuity detection. Delta coefficients represent the local time derivatives (slope) of the cepstral sequence and are computed according to

$$\Delta \mathbf{x}_m = \sum_{l=-L}^{L} l\, \mathbf{x}_{m-l}. \quad (5)$$

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 according to

$$\Delta^2 \mathbf{x}_m = \sum_{n=-N}^{N} n \Delta \mathbf{x}_{m-n}. \quad (6)$$

Double-delta coefficients indicate the acceleration of the spectral components; in our simulations, $N = 3$ is used.

As mentioned in Section II-B, the zeroth cepstral coefficient is used as an energy term. The delta and double-delta features, calculated from x_0, provide insight into the dynamics of the signal energy. The main assumption used here is that for natural speech, abrupt changes in signal energy do not occur. The two main temporal impairments that should be detected are abrupt starts and abrupt stops [15]. In abrupt starts, the signal energy, its rate of change, and acceleration increase abruptly. The opposite occurs with abrupt stops. This behavior is illustrated with Figs. 4 and 5. In Fig. 4(a)–(d), the waveform of a speech signal, the energy, and energy rate of change (Δ_0) and acceleration (Δ_0^2) are depicted, respectively. The signal consists of five vowels uttered by a male speaker in a noisy office environment. Vowels are chosen as their extremities are often erroneously detected as abrupt starts or stops. Notice the subtle spikes in Δ_0 and Δ_0^2 at each vowel extremity. In Fig. 5, temporal discontinuities, or "clippings," have been introduced at the beginning or at the end of each vowel. The abrupt starts and stops are indicated with arrows. Notice that the unnatural changes cause abnormal spikes in Δ_0 and Δ_0^2.

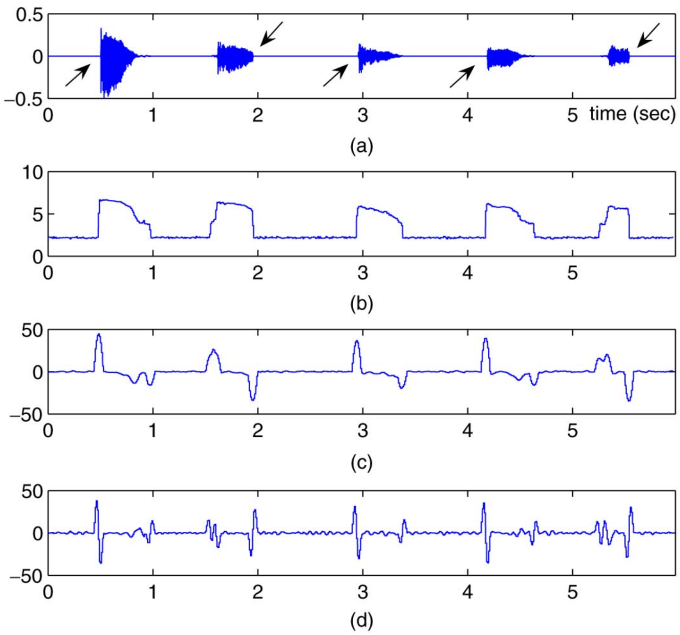


Fig. 5. Analysis of a "clipped" signal's (a) waveform, (b) x_0, (c) Δ_0, and (d) Δ_0^2. Abrupt starts and stops are indicated with arrows.

To detect abrupt starts or stops, two steps are required. First, the energy of frame at time t_c is compared to the energy of frame $t_c + \tau$. If the energy increase (or decrease) surpasses a certain threshold T, then a candidate abrupt start (or stop) is detected. The parameters T and τ are optimized on our training database, as described in Section III-B-4. Once a candidate discontinuity is detected, a support vector classifier is used to decide whether in fact a temporal discontinuity has occurred. The SVC is only invoked at candidate discontinuities in order to reduce the computational complexity of the algorithm. Here, two SVCs are used: one tests for abrupt starts (given a sudden increase in x_0) and the other for abrupt stops (given a sudden decrease in x_0). Input features to the SVC are Δ_0 and Δ_0^2 for the $\tau + F$ frames preceding t_c and the $\tau + F$ frames succeeding t_c. The parameter F is empirically set to 2, resulting in a ten-dimensional input feature vector. The output of each classifier is one of two possible classes, namely, "discontinuity" or "nondiscontinuity." We experiment with linear, polynomial and radial basis function (RBF) support vector classifiers; on our databases, an RBF SVC attained superior performance.

The output **SS** of the temporal discontinuity detection block, as depicted in Fig. 2, is a $(n_b + n_s + 2)$-dimensional vector comprised of the number of detected abrupt starts (n_b) and abrupt stops (n_s) and the approximate time at which each discontinuity occurs. As an example, suppose for a given speech file three abrupt starts are detected at times $\mathbf{t}_b = \{t_{b1}, t_{b2}, t_{b3}\}$ and two abrupt stops at times $\mathbf{t}_s = \{t_{s1}, t_{s2}\}$.The resulting parameter **SS** is represented by $\mathbf{SS} = \{n_b, \mathbf{t}_b, n_s, \mathbf{t}_s\}$.

F. Final MOS Calculation

The final MOS-LQO calculation is based on a linear combination of the intermediate MOSs, adjusted by the negative effects temporal discontinuities have on perceived quality, i.e.,

$$\widehat{\text{MOS}} = p_m \text{MOS}_{\text{tmp},1} + (1 - p_m)\text{MOS}_{\text{tmp},2} - C(\mathbf{SS}). \quad (7)$$

Here, p_m is the probability that $\text{MOS}_{\text{tmp},1}$ is better than $\text{MOS}_{\text{tmp},2}$ as an estimator of MOS-LQS. This statistic is calculated by the detector on a "per-signal" basis. More detail regarding the computation of p_m is given in Section III-B-5.

The term $C(\mathbf{SS})$ resembles the effects temporal discontinuities have on perceived quality. Experiments, such as [32], suggest that humans can perform continuous assessment of time-varying speech quality. It is also noted that the location of a discontinuity within a signal can affect the listener's perception of quality; this short-term memory effect is termed "the recency effect." Impairments detected at the end of the signal have more negative effect on the perceived quality than impairments detected at the beginning. In [15], a decay model is used to emulate the recency effect. More recently, however, experiments carried out in [33] suggest that the recency effect is harder to observe in speech signals of short time duration. Instead, a "subconscious integration" is performed where unconsciously, multiple degradations are combined and reported as a single level of speech quality.

Since the files in our databases are of short time durations (an average 6 s), we do not consider the recency effect and model $C(\mathbf{SS})$ as

$$C(\mathbf{SS}) = C(n_b, n_s) = n_b K_b + n_s K_s \quad (8)$$

where K_b and K_s are penalty terms for the detected abrupt starts and stops, respectively. These constants are optimized on the training databases, as will be discussed in Section III-B5. In this paper, since the recency effect is not considered, $\mathbf{t}_s$ and $\mathbf{t}_b$ are not computed. Nevertheless, for longer speech files, (8) can be modified to incorporate such temporal information; in particular, a decay model can be employed.

III. Algorithm Design Considerations

A. Database Description

In total, 20 MOS-LQS-labeled databases are used in our experiments. The speech databases are described in Table I. We separate 14 databases for training (databases 1–14) and the remaining six are used for testing (databases 15–20). In addition, during training several algorithm parameters need to be optimized. To this end, 20% of the training set is randomly chosen to be used for parameter validation; henceforth, this subset will be referred to as the "validation set." Parameter calibration is discussed in further detail in Section III-B. The content of each database is described next.

Databases 1–7 are the ITU-T P-series Supplement 23 (Experiments 1 and 3) multilingual databases [34]. The three databases in Experiment 1 have speech processed by various codecs (G.726, G.728, G.729, GSM-FR, IS-54 and JDC-HR), singly or in different cross tandem configurations (e.g., G.729–G.728–GSM-FR). The four databases in Experiment 3 contain single- and multiple-encoded G.729 speech under various channel error conditions (BER 0–10%; random and burst FER 0–5%) and input noise conditions (clean, vehicle, street, and hoth noises).

Databases 8 and 9 are two wireless databases with speech processed, respectively, by the IS-96A and IS-127 EVRC (Enhanced Variable Rate Codec) codecs under various channel error conditions (forward and reverse 3% FER) with or without the G.728 codec in tandem. Database 10 is a mixed wireless–wireline database with speech under a wide range of degradation conditions—tandemings, channel errors, temporal clippings, and amplitude variations. A more detailed description of the conditions in database 10 can be found in [35]. Databases 11–13 comprise speech coded using the G.711, G.726 and the G.728 speech coders, alone and in various different tandem configurations. Database 14 has speech from standard speech coders (G.711, G.726, G.728, G.729, and G.723.1), under various channel degradation conditions (clean, 0.01% BER, 1–3% FER).

Databases 15–17 comprise speech coded with the 3GPP2 Selectable Mode Vocoder (SMV) under different tandeming,

TABLE I
Properties of Speech Databases Used in our Experiments

Database	Language	No. of Files	No. of Conditions	Training	Testing
1	French	176	44	✓	
2	Japanese	176	44	✓	
3	English	176	44	✓	
4	French	200	50	✓	
5	Italian	200	50	✓	
6	Japanese	200	50	✓	
7	English	200	50	✓	
8	English	96	24	✓	
9	English	96	24	✓	
10	English	240	60	✓	
11	Italian	2440	20	✓	
12	Japanese	2440	20	✓	
13	English	2440	20	✓	
14	English	2088	46	✓	
15	English	3072	48		✓
16	English	3072	48		✓
17	English	3072	48		✓
18	English	3328	52		✓
19	English	96	24		✓
20	English	448	28		✓

channel impairments, and environment noise degradation conditions. Database 18 has speech from standard speech coders (G.711, G.726, G.728, G.729E, and GSM-EFR) and speech processed by a cable VoIP speech coder, under various channel degradation conditions. Lastly, databases 19 and 20 have speech recorded from an actual telephone connection in the San Francisco area and live network speech samples collected from AMPS, TDMA, CDMA, and IS-136 forward and reverse links. In all databases described above, speech degraded by different levels of MNRU are also included.

Speech files from databases 15–20 are used solely for testing and are unseen to the algorithm. Database 15–18 are kept for testing as they provide speech files coded using newer codecs than the codecs represented in the training datasets. Evaluation using these databases demonstrates the applicability of the proposed algorithm to emerging codec technologies. Database 19 has speech files that are composed of two spoken utterances, one by a male speaker and the other by a female speaker, and thus are regarded as being composite male–female signals. Although this is not common in listening tests, we are interested in seeing how robust the proposed algorithm is to speaker and gender changes. Furthermore, database 20 is composed of speech files that have been processed by older wireless codecs. Many of the files in this database are of poor speech quality (MOS-LQS $<$ 2) and comprise degradation conditions not represented in the training datasets.

B. Algorithm Parameter Calibration

In order to optimize algorithm parameters, preliminary "calibration" experiments are carried out. In the sequel, we describe the steps taken to calibrate each of the processing blocks depicted in Fig. 2.

1) Multiplicative Noise Estimation and MOS Mapping: For optimization of the multiplicative noise estimator, MNRU-degraded training files are used. Experiments are carried out with second- and third-order polynomial mappings between $\bar{\sigma}_{\text{active}}$ and the Q value. On the validation set, the latter presented better performance. The estimated amount of multiplicative noise achieved a 0.92 correlation with the true Q value. The multiplicative noise estimator described in [16] resulted in a correlation of 0.66. For the noise-to-MOS mapping, it is found that a simple linear regression between the estimated amount of multiplicative noise and $\text{MOS}_{\text{tmp},1}$ suffices. The two mappings are replaced by one single third-order polynomial mapping between $\bar{\sigma}_{\text{active}}$ and $\text{MOS}_{\text{tmp},1}$. A 0.95 correlation between $\text{MOS}_{\text{tmp},1}$ and the true MOS-LQS is attained for MNRU validation files.

2) Consistency Calculation: To calibrate the consistency calculation block, an effective combination of GMM configuration parameters (M and covariance matrix type) needs to be found. For voiced and unvoiced frames, we experiment with diagonal matrices and $M = 8$, 16, or 32, and $M = 2$, 3, or 5 for full covariance matrices. For inactive frames, we only experiment with diagonal matrices and $M = 2$, 3, or 6. The calibration experiment suggests the use of three full GMM components for voiced frames and 32 diagonal components for unvoiced frames, for both the clean and the degradation model. For inactive frames, six diagonal components are needed for the degradation model and three for the clean model. This is consistent with the fact that for clean speech, inactive frames have virtually no signal energy and fewer Gaussian components are required.

The consistency-to-MOS mapping is designed using a MARS regression function with parameters optimized using degraded MOS-LQS labeled training files. The function maps the six consistency measures into $\text{MOS}_{\text{tmp},2}$. As mentioned previously, the designed MARS regression function is composed of a simple weighted sum of 18 truncated linear functions. The mapping is performed once per speech signal and incurs negligible computational complexity (approximately 18 scalar multiplications and 54 scalar additions). For files in the validation set, a 0.82 correlation is attained between $\text{MOS}_{\text{tmp},2}$ and the actual MOS-LQS; if MNRU degraded files are removed, the correlation increases to 0.86. This result suggests that a combination of $\text{MOS}_{\text{tmp},1}$ and $\text{MOS}_{\text{tmp},2}$ may lead to better performance when compared to using $\text{MOS}_{\text{tmp},2}$ alone.

3) Multiplicative Noise Detection: The multiplicative noise detector is optimized to select the best preliminary quality score, $\text{MOS}_{\text{tmp},1}$ or $\text{MOS}_{\text{tmp},2}$, for a given test signal. To gain a sense of which conditions are best represented by each preliminary score, tests are performed on the training set where the true MOS-LQS is known. As expected, of 288 files processed by the G.711 and G.726 codecs, 252 are better represented by $\text{MOS}_{\text{tmp},1}$. Similarly, of 252 MNRU-degraded files with 0 dB $< Q <$ 35 dB, 209 are better represented by $\text{MOS}_{\text{tmp},1}$. If only files with $Q <$ 20 dB are considered, 103 (out of 108) are better estimated by $\text{MOS}_{\text{tmp},1}$. The primary objective of the detector, thus, is to detect signals corrupted by high levels of multiplicative noise.

Nonetheless, for some degradation conditions other than multiplicative noise conditions, $\text{MOS}_{\text{tmp},1}$ is also shown to be a better estimator of MOS-LQS than $\text{MOS}_{\text{tmp},2}$. Some examples include speech signals processed by low bit-rate vocoders (e.g., G.723.1 at 5.3 kbit/s), where the quality of five (out of 32) of the signals are better represented by $\text{MOS}_{\text{tmp},1}$. Moreover, of 112 samples processed by medium bitrate codecs (e.g., G.729E at 11.8 kbit/s), the quality of 22 signals are better estimated by $\text{MOS}_{\text{tmp},1}$ than by $\text{MOS}_{\text{tmp},2}$. As a consequence, in instances where high levels of multiplicative noise is not detected, the classifier learns which temporary score results in best estimation performance.

To calibrate the detector, first, all training samples are processed by the top *and* middle branches depicted in the block diagram in Fig. 2. The estimated preliminary MOSs are compared to the true MOS-LQS and all samples in which the top branch achieved smallest estimation error receive a label "TOP"; otherwise, a label "MID" is assigned. This new labeled training set discriminates which preliminary score best estimates the true MOS-LQS for a given speech signal and is used to train the detector. The detector can be designed to operate in two different modes: hard-decision or soft-decision. In hard-decision mode, the detector selects the single best preliminary quality score and $p_m \in \{0, 1\}$ is used in (7). With this mode, only one preliminary score needs to be computed. On the contrary, soft-decision detection requires that both preliminary scores be estimated, and a "weight" is assigned to each score. The weight $(0 \leq p_m \leq 1)$ is computed by the detector on a "per-signal" basis and reflects the probability of $\text{MOS}_{\text{tmp},1}$ more accurately predicting MOS-LQS than $\mathrm{MOS}_{\mathrm{tmp},2}$. The term p_m resembles the likelihood of the presence of high levels of multiplicative noise in the signal. After detector optimization, signals in the validation set with high levels of multiplicative noise have p_m that approach unity.

We experiment with three different candidate classifiers: CART, SVC, and RF. The classifiers are trained using the aforementioned labeled training set. An RF classifier is an ensemble of unpruned decision trees induced from bootstrap samples of the training data. The final class decision is based on a majority vote from all individual trees (see [28] for more details regarding random forest classifiers). On our validation set, an RF classifier with 500 trees (and three inputs considered per tree node) achieved the best classification performance; all files with high levels of multiplicative noise (e.g., MNRU with $Q < 12$ dB) were correctly detected.

4) Temporal Discontinuity Detection: Calibrating the temporal discontinuity detector encompasses the determination of parameters T and τ, and training of the support vector classifiers. On our data it was found that if the values of x_0 doubled (or halved) within 20–50 ms, a candidate discontinuity could be detected. With these possible values of τ, the SVCs correctly identified all abrupt stops and starts on the validation dataset. In an attempt to reduce the number of times the SVCs are executed, a more stringent threshold, $\tau = 2$ (equivalent to 2 ms), is used.

5) Final MOS Calculation: Lastly, the parameters in (7) are optimized. Initially, $C(\mathbf{SS}) = 0$ is assumed and we experiment with hard-decision detection and soft-decision detection. On the validation set, soft-decision detection resulted in superior performance. With soft-decision detection, p_m is computed by the RF classifier and represents the fraction of the 500 individual decision trees that have selected $\mathrm{MOS}_{\mathrm{tmp},1}$ as the best estimator of subjective quality. Once the soft-decision mode is set, the parameters K_b and K_s in (8) are estimated by minimizing the squared error between (7) and the true MOS-LQS for "clipped" training signals. On our data, $K_b = 0.09$ and $K_s = 0.13$ were found. These parameters are consistent with [15], where it is argued that the abrupt stops have, intuitively, a more significant impact on perceived speech quality relative to abrupt starts.

IV. Test Results

In this section, we compare the proposed algorithm to P.563 using the test databases described in Section III-A. The performance of the algorithms is assessed by the correlation (R) between the N MOS-LQS (w_i) and MOS-LQO (y_i) samples, using Pearson's formula

$$R = \frac{\sum_{i=1}^{N}(w_i - \bar{w})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(w_i - \bar{w})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{9}$$

where $\bar{w}$ is the average of w_i, and $\bar{y}$ is the average of y_i. MOS measurement accuracy is assessed using the root-mean-square error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(w_i - y_i)^2}{N}}. \tag{10}$$

TABLE II
Performance Comparison on Unseen Test Datasets. Results Are Per condition After Third-Order Polynomial Regression

Test Database	P.563		Proposed			
	R	$RMSE$	R	%↑	$RMSE$	%↓
15	0.863	0.253	0.908	32.8	0.206	18.6
16	0.835	0.274	0.864	17.6	0.249	9.1
17	0.748	0.273	0.868	47.6	0.212	22.3
18	0.916	0.218	0.939	27.4	0.187	14.2
19	0.421	0.456	0.868	77.2	0.455	0.2
20	0.758	0.569	0.909	62.4	0.362	36.4
Average	–	–	–	**44.2**	–	**16.8**

Table II presents "per-condition" R and RMSE between condition-averaged MOS-LQS and condition-averaged MOS-LQO, for each of the test datasets. The results are obtained after an individual third-order monotonic polynomial regression for each dataset, as recommended in [16]. The column labeled "% ↑" lists the percentage "R-improvement" obtained by using the proposed GMM-based method over P.563. The R-improvement is given by

$$\% \uparrow = \frac{R_{\mathrm{GMM}} - R_{P.563}}{1 - R_{P.563}} \times 100\% \tag{11}$$

and indicates percentage reduction of P.563's performance gap to perfect correlation. The column labeled "% ↓" lists percentage reduction in RMSE, relative to P.563, by using the proposed scheme. As can be seen, the proposed algorithm outperforms P.563 on all test databases. An average R-improvement of 44% and an average reduction in RMSE of 17% is attained.

An interesting result is obtained with database 19. Recall that this database had MOS-LQS-labeled speech signals composed of two utterances, one spoken by a male speaker and the other by a female speaker. On this database, P.563 achieves a poor correlation of 0.421. In fact, before applying the third-order monotonic polynomial mapping, P.563 achieves a very poor $R = 0.121$. This may be due to the fact that P.563 depends on vocal tract analysis to test for unnaturalness of speech. By rating the unnaturalness of speech separately for male and female voices, P.563 is compromised for composite male–female signals. As a sanity check, we test the performance of PESQ (with the mapping described in [36]) and an $R = 0.974$ and $\mathrm{RMSE} = 0.422$ is attained.

The plots in Fig. 6 show MOS-LQO versus MOS-LQS for the proposed algorithm and for P.563. Each data point represents one of the 332 different degradation conditions available in the test databases. In these plots and in the performance figures described below, the composite male–female quality estimates are left out. Plots (a) and (b) illustrate the relationship between GMM MOS-LQO and MOS-LQS, before and after third-order monotonic polynomial regression (optimized on each test dataset), respectively. Prior to polynomial mapping, an overall

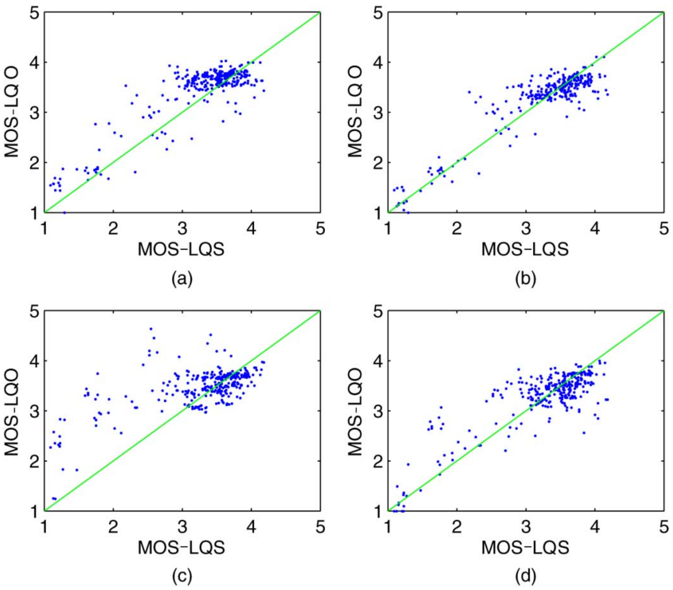


Fig. 6. Per-condition MOS-LQO versus MOS-LQS for (a) proposed algorithm prior to and (b) after third-order monotonic polynomial mapping, and for (c) P.563 before and (d) after the polynomial mapping.

$R = 0.874$ and RMSE $= 0.321$ is attained; after the mapping, $R = 0.918$ and RMSE $= 0.221$. Similarly, plots (c) and (d) illustrate the relationship between P.563 MOS-LQO and MOS-LQS, before and after the monotonic mapping. An overall $R = 0.7281$ and RMSE $= 0.391$ is attained prior to regression; after regression $R = 0.853$ and RMSE $= 0.292$.

The third-order monotonic polynomial regression is suggested in [16] in order to map the objective score onto the subjective scale. This mapping is used to compensate for variations of the MOS-LQS scale across different subjective tests, variations due to different voter groups, languages, contexts, among other factors. Monotonic mappings perform scale adjustments but do not alter the ranking of the objective scores. Ultimately, the goal in objective quality estimation is to design algorithms whose quality scores rank similarly to subjective quality scores. This is due to the fact that objective scores that offer good ranking performance produce accurate MOS-LQS estimates, given a suitable monotonic mapping is used for scale adjustment. To this end, we use rank-order correlations as an additional figure of merit of algorithm performance. Rank-order correlations are calculated using (9), with the original data values replaced by the *ranks* of the data values; this measure is often termed Spearman's rank correlation coefficient (R_S). For P.563, a "per-condition" $R_S = 0.705$ is attained on the test data. The proposed algorithm achieves $R_S = 0.793$, a 30% R-improvement. The results presented above, for all three performance measures, suggest that the proposed algorithm provides more accurate estimates of subjective quality relative to the current "state-of-art" P.563 algorithm.

In the sequel, a novel and more flexible architecture for double-ended measurement is presented. The scheme makes use of the proposed single-ended algorithm. It will be shown that the scheme is applicable to noise suppression systems and outperforms current state-of-art double-ended algorithms.

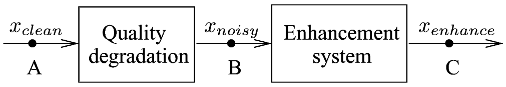


Fig. 7. Block diagram of a speech enhancement system.

V. Versatile Double-Ended Measurement Architecture

A. Background

Current single- and double-ended algorithms are only capable of estimating the quality of the received signal *per se*. If we are interested in analyzing the quality of a transmission system, assumptions on the input signal are needed. As mentioned previously, double-ended algorithms presuppose that the input is undistorted. Moreover, it is assumed that the output is of quality no better than the input. Current double-ended algorithms would fail if any of these assumptions were to fail. A scenario where *both* assumptions are not met can be seen in Fig. 7. Here, a clean signal x_{clean} suffers impairments that degrade speech quality. Common impairments may include interference on an analog access network, environment noise, noise introduced by equipment within the network, and lost packets in a VoIP network. The noisy signal x_{noisy} is then input to a speech enhancement system and the enhanced output x_{enhance} is of quality better than the input. Such system configuration commonly occurs when using a noise reduction algorithm to enhance speech. As will be shown next, performance of current double-ended schemes may be compromised when only x_{noisy} and x_{enhance} are made available to the algorithm.

B. Measurement Configuration

The objective here is to devise a measurement scheme that overcomes the above limitations. Our approach subsumes current single- and double-ended measurement architectures. The approach allows for double-ended measurement without the underlying assumptions mentioned above, i.e., the input signal does *not* need to be clean and the output *can* be of quality better than the input. With the proposed architecture, it is possible to analyze the quality of the system under test and both quality degradations and quality enhancements can be detected and handled. This section will give emphasis to quality enhancements.

The proposed architecture is depicted in Fig. 8. The conventional double-ended algorithm is replaced by two single-ended schemes, one at the input and another at the output of the system being tested, and a system diagnosis tool. This configuration requires information of both the input and the output signals, hence, is regarded as double-ended. In analogy to Fig. 7, if the input single-ended algorithm is placed at the point labeled "A" and the output single-ended algorithm is placed at point labeled "B," then quality degradations are handled. On the other hand, if the input single-ended algorithm is placed at the point labeled "B" and the output single-ended algorithm is placed at point labeled "C," then quality enhancements are handled. Here, we will focus on the latter scenario as it represents the case where the input signal is *not* clean and the output *is* of quality better

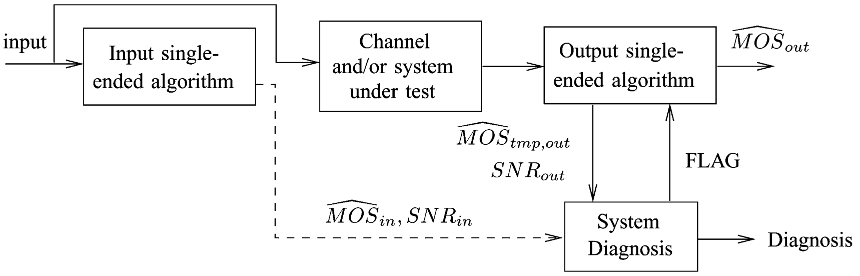

input
Input single-ended algorithm
Channel and/or system under test
Output single-ended algorithm
$\widehat{MOS}_{out}$
$\widehat{MOS}_{tmp,out}$
SNR_{out}
FLAG
$\widehat{MOS}_{in}, SNR_{in}$
System Diagnosis
Diagnosis

Fig. 8. Architecture of the proposed double-ended algorithm.

than the input. As mentioned previously, performance of current double-ended schemes may be compromised in this scenario.

To allow for accurate speech quality measurement of noise suppressed signals, the proposed GMM-based algorithm is updated to incorporate reference models of noise suppressed speech signals. Similar to the clean and the degradation models, the "noise-suppressed" model is designed for voiced, unvoiced, and inactive frames. If noise suppression is detected, consistency measures relative to all three reference models (clean, degraded, and noise suppressed) are computed; otherwise, consistency measures are computed only for the clean and degradation models. In the latter case, the MARS mapping described in Section III-B-2 is used; in the former, a separate MARS mapping is trained on a subjectively scored noise-suppressed speech database. Details regarding the database will be given in Section V-C. Noise-suppressed reference model and MARS mapping design considerations will be given in Section V-D.

With the proposed architecture, the transmission of input measurements (this is illustrated with the dashed arrow in Fig. 8) and a system diagnosis module are necessary in order to detect if noise suppression has occurred. We have investigated the effectiveness of transmitting the input SNR $(\mathrm{SNR}_{\mathrm{in}})$, computed by the VAD algorithm, and the input MOS-LQO $(\widehat{\mathrm{MOS}}_{\mathrm{in}})$, estimated based on the consistency measures calculated relative to the clean and the degradation reference models. The amount of side information is negligible; thus, this scheme is much more economical than existing double-ended schemes which require access to the input signal.

At the output end, $\mathrm{SNR}_{\mathrm{out}}$ is computed and a preliminary MOS-LQO $(\widehat{\mathrm{MOS}}_{\mathrm{tmp,out}})$ is estimated based on the clean and the degradation model-based consistency measures. These two measures are sent to the system diagnosis module. The diagnosis module, in turn, sends a flag back to the output single-ended algorithm indicating whether noise suppression has been detected. Detection occurs if $\mathrm{SNR}_{\mathrm{in}} < \mathrm{SNR}_{\mathrm{out}}$ and $\widehat{\mathrm{MOS}}_{\mathrm{in}} + \gamma < \widehat{\mathrm{MOS}}_{\mathrm{tmp,out}}$, where γ is the standard deviation of the estimated input MOS-LQO ($\gamma = 0.29$ on the noise suppressed database). With this detection rule, all of the noise suppressed speech files were correctly detected. If noise suppression is detected, the output single-ended algorithm calculates a final MOS-LQO $(\widehat{\mathrm{MOS}}_{\mathrm{out}})$ based on consistency measures calculated relative to the three reference models, otherwise $\widehat{\mathrm{MOS}}_{\mathrm{out}} = \widehat{\mathrm{MOS}}_{\mathrm{tmp,out}}$. Other diagnostic tests, such as measuring (in terms of MOS) the amount of quality degradation (or enhancement) imparted by the transmission system, or measuring SNR improvement, are also possible. Further characterization of the noise suppression algorithm may be aided with the transmission of other input measurements (e.g., see measures described in [37]).

C. Database Description

The proposed architecture is tested using the subjectively scored NOIZEUS database [38]. The database comprises speech corrupted by four types of noise (babble, car, street, and train) at two SNR levels (5 and 10 dB) and processed by 13 different noise suppression algorithms; a total of 1792 speech files are available. The noise suppression algorithms fall under four different classes: spectral subtractive, subspace, statistical-model based, and Wiener algorithms. A complete description of the algorithms can be found in [38], [39].

The subjective evaluation of the NOIZEUS database was performed according to ITU-T Recommendation P.835 [40]. With the P.835 methodology, listeners are instructed to successively attend to and rate three different signal components of the noise suppressed speech signal: 1) the speech signal alone using a five-point scale of signal distortion $[1 = \text{very distorted}, 5 = \text{not distorted}]$, 2) the background noise alone using a five-point scale of background intrusiveness $[1 = \text{very intrusive}, 5 = \text{not noticeable}]$, and 3) the overall effect using the five-point ACR scale $[1 = \text{bad}, 5 = \text{excellent}]$. Here, the average scores over all listeners are termed SIG-LQS, BCK-LQS, and OVRL-LQS, respectively. Note that OVRL is equivalent to the MOS described in [1].

D. Design Considerations

In order to train reference models of noise suppressed speech signals and to design the updated MARS mapping function, the NOIZEUS database has to be separated into a training and a test set. We perform this separation in three different ways to test the robustness of the proposed architecture to different unseen test conditions. First, speech files are separated according to noise levels; files with $\mathrm{SNR} = 10$ dB are used for training and files with $\mathrm{SNR} = 5$ dB are left for testing. Second, speech signals are separated according to noise sources. Signals corrupted by street and train noise are used for training and signals corrupted by babble and car noise are left for testing. Lastly, speech files are separated according to noise suppression algorithms. For training, noisy signals processed by spectral subtractive and subspace algorithms are used; noisy signals processed by statistical-model based and Wiener algorithms are left for

TABLE III
PERFORMANCE COMPARISON WITH PESQ AND P.563 ON THE THREE TEST SETS. CONFIGURATION 1 MAKES USE OF THE ORIGINAL CLEAN SIGNAL AND THE NOISE-SUPPRESSED SIGNAL, AND CONFIGURATION 2 OF THE NOISY SIGNAL AND THE NOISE-SUPPRESSED SIGNAL

Test No.	PESQ - Configuration 1		PESQ - Configuration 2		Proposed Architecture		P.563	
	R	$RMSE$	R	$RMSE$	R	$RMSE$	R	$RMSE$
1	0.886	0.178	0.610	0.305	0.861	0.196	0.587	0.311
2	0.864	0.233	0.581	0.377	0.827	0.261	0.563	0.384
3	0.922	0.181	0.731	0.321	0.817	0.270	0.637	0.361

testing. The number of conditions for each of the three test sets described above are 52, 52, and 64, respectively, out of a total of 104 degradation conditions for the entire database.

To design the reference models for noise suppressed speech signals we experiment with different combinations of GMM parameters. It is observed that for all three tests, 32 diagonal components for voiced and unvoiced frames and six diagonal components for inactive frames strike a balance between accuracy and complexity. Moreover, a separate MARS mapping function is designed for each of the three tests. Each MARS function maps a nine-dimensional feature vector into $\widehat{\mathrm{MOS}}_{\mathrm{out}}$.

E. Test Results

In this section, we compare the performance of the proposed architecture to that of PESQ. Two different PESQ configurations are tested: 1) a hypothetical configuration where the original clean signal is available, and 2) a more realistic scenario where only the noisy and the noise-suppressed signals are available. Configuration 1 makes use of the clean signal as reference input and although evaluation of noise reduction systems is not recommended in [41], the results to follow suggest accurate estimation performance. On the other hand, Configuration 2 exemplifies the case where the reference input signal is not clean, and the quality of the output is better than that of the input. As will be shown in the sequel, this configuration compromises PESQ performance. Moreover, swapping the input signals (i.e., noise-suppressed signal to reference input and noisy signal to degraded input) brought no improvement.

Table III presents "per-condition" R and RMSE between condition-averaged OVRL-LQS and condition-averaged OVRL-LQO, for the three test sets. Results are reported after third-order monotonic polynomial regression (for PESQ the mapping proposed in [36] is not used as it degrades performance substantially). As can be seen, when the original clean speech signal is available, PESQ achieves accurate estimation performance. However, when only the noisy signal is available as reference, substantial improvement is attained with the proposed architecture. For comparison purposes, Table III also shows the performance of P.563 on the three tests.

F. Component Quality Estimation

It is known that certain noise suppression algorithms can introduce unwanted artifacts such as "musical noise." With recommendation P.835, noise suppressed signals are rated based on the speech content alone (SIG), on the background noise alone (BCK), and on the speech plus noise content (OVRL). Currently, objective measurement algorithms (both single- and double-ended) can only attempt to estimate OVRL-LQS. However, it is unknown how humans integrate the individual contributions of speech and noise distortions when judging the overall quality of a noise-suppressed signal. To this end, devising an algorithm capable of also estimating SIG-LQS and BCK-LQS would be invaluable. The estimates can be used to test newer generations of noise reduction algorithms and to assess the algorithms' capability of maintaining speech signal naturalness while reducing background noise to nonintrusive levels. In [42], the NOIZEUS database is used to evaluate six double-ended objective estimates of SIG-LQS and BCK-LQS. The study makes use of the original clean signal as a reference and low correlations with subjective quality are reported $(R < 0.65)$.

TABLE IV
PERFORMANCE OF SIG-LQO AND BCK-LQO ESTIMATED BY THE PROPOSED ALGORITHM

Test No.	SIG-LQO		BCK-LQO	
	R	$RMSE$	R	$RMSE$
1	0.813	0.295	0.717	0.235
2	0.804	0.355	0.728	0.289
3	0.807	0.331	0.707	0.305

Due to the modular architecture of the proposed GMM-based algorithm, a simple extension can be implemented to allow for single-ended SIG-LQS and BCK-LQS estimation. In particular, two new MARS mapping functions are optimized on the training datasets. To estimate SIG-LQS, a six-dimensional MARS function is devised to map consistency measures of voiced and unvoiced frames (for all three reference models—clean, degraded, and noise suppressed) into SIG-LQO. To estimate BCK-LQS, a simple four-dimensional MARS function is designed to map consistency measures of inactive frames (for all three models) and the estimated SNR into BCK-LQO. Table IV presents "per-condition" R and RMSE between condition-averaged SIG-LQS (BCK-LQS) and condition-averaged SIG-LQO (BCK-LQO), for the three aforementioned test sets. Results are reported after third-order monotonic polynomial regression optimized on each test set. The results are encouraging given that the original clean signal is *not* available as a reference.

The results described in this section are "preliminary," as only one noise-suppressed speech dataset is used for training and testing. The results, however, are promising and suggest that the GMM-based single-ended algorithm, deployed in the proposed double-ended architecture, can be effectively used in scenarios where a noisy input is enhanced using a noise suppression algorithm. Moreover, the proposed architecture has the added benefit of estimating SIG-LQS and BCK-LQS, invaluable information for assessing the performance of noise suppression algorithms.

VI. Conclusion

The purpose of this paper has been two fold. First, a novel single-ended speech quality estimation algorithm employing speech signal models designed using machine learning methods is presented. Comparisons with the current state-of-art P.563 algorithm demonstrate the efficacy of the algorithm and its potential for providing more accurate measurements. Second, the proposed algorithm is extended and applied to a distributed double-ended measurement architecture. The results demonstrate that, besides offering the conventional function of measuring the quality of systems that degrade speech, the algorithm is capable of measuring the quality of speech enhancement systems. In this role, the proposed algorithm performs better than P.563 and provides a functionality not available with the current double-ended standard P.862 PESQ.

Acknowledgment

The authors would like to thank Prof. P. Loizou for providing the NOIZEUS database and the anonymous reviewers for their valuable comments and suggestions.

References

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

[2] "Mean opinion score (MOS) terminology," ITU, Geneva, Switzerland, 2003, ITU-T Rec. P.800.1.

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

[4] N. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

[5] R. Kubichek, D. Atkinson, and A. Webster, "Advances in objective voice quality assessment," in *Proc. IEEE GLOBECOM Conf.*, 1991, pp. 1765–1770.

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

[7] "Objective quality measurement of telephone-band (300–3400 Hz) speech codecs," ITU, Geneva, Switzerland, 1996, ITU-T Rec. P.861.

[8] 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.

[9] ——, "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.

[10] 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.

[11] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," ITU, Geneva, Switzerland, 2001, ITU-T Rec. P.862.

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

[13] 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.

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

[15] 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.

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

[17] "Adaptive multi-rate (AMR) speech codec: Voice activity detector (VAD), release 6," 2004, 3GPP2 TS 26.094.

[18] W. B. Kleijn and K. K. Paliwal, *Speech Coding and Synthesis*. Amsterdam, The Netherlands: Elsevier, 1995, ch. A Robust Algorithm for Pitch Tracking (RAPT), pp. 495–518.

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

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

[21] H. Hermansky, "Mel cepstrum, deltas, double-deltas-What else is new?," in *Proc. Robust Methods for Speech Recognition in Adverse Conditions*, Tampere, Finland, 1999.

[22] N. Jayant, "Digital coding of speech waveforms: PCM, DPCM, and DM quantizers," *Proc. IEEE*, vol. 62, no. 5, pp. 611–632, May 1974.

[23] "Modulated noise reference unit—MNRU," ITU, Geneva, Switzerland, 1996, ITU-T Rec. P.810.

[24] T. H. Falk and W.-Y Chan, "Enhanced non-intrusive speech quality measurement using degradation models," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, May 2006, vol. I, pp. 837–840.

[25] S. Voran, "Observations on the t-reference condition for speech coder evaluation," 1992, CCITT SG-12, Document Number SQ.13.92.

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

[27] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Monterey, CA: Wadsworth & Brooks, 1984.

[28] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[29] J. H. Friedman, "Multivariate adaptive regression splines," *Ann. Stat.*, vol. 19, no. 1, pp. 1–141, Mar. 1991.

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

[31] J. F. Canny, "Finding edges and lines in images," MIT—Artificial Intelligence Laboratory, 1983, Tech Rep. 720.

[32] M. Hansen and B. Kollmeier, "Continuous assessment of time-varying speech quality," *J. Acoust. Soc. Amer.*, vol. 106, no. 5, pp. 2888–2899, Nov. 1999.

[33] S. Voran, "A basic experiment on time-varying speech quality," in *Proc. Int. Conf. Measurement Speech Audio Quality Netw.*, Jun. 2005.

[34] "ITU-T coded-speech database," ITU, Geneva, Switzerland, 1998, ITU-T Rec. P.Suppl. 23.

[35] L. Thorpe and W. Yang, "Performance of current perceptual objective speech quality measures," in *Proc. IEEE Speech Coding Workshop*, 1999, pp. 144–146.

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

[37] E. Paajanen, B. Ayad, and V. Mattila, "New objective measures for characterisation of noise suppression algorithms," in *Proc. IEEE Speech Coding Workshop*, Sep. 2000, pp. 23–25.

[38] Y. Hu and P. Loizou, "Subjective comparison of speech enhancement algorithms," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, May 2006, vol. I, pp. 153–156.

[39] ——, "Subjective comparison and evaluation of speech enhancement algorithms," in *Speech Commun.*, 2006, submitted for publication.

[40] "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithms," ITU, Geneva, Switzerland, 2003, ITU-T Rec., P.835.

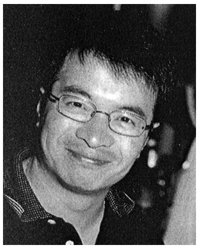
[41] "Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2," ITU, Geneva, Switzerland, 2005, ITU-T Rec. P.862.3.

[42] Y. Hu and P. Loizou, "Evaluation of objective measures for speech enhancement," in *Proc. Int. Conf. Spoken Language Process.*, 2006, to be published.

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 coding and communications, communication theory, and channel modeling.

Mr. Falk is a member of the International Speech Communication Association Student Advisory Committee and also a Student Member of the Brazilian Telecommunication Society. Mr. Falk was recipient of the Natural Sciences and Engineering Research Council (NSERC) Canada Graduate Scholarship, in 2006, the Best Student Paper Award (in the Speech Processing category) at the International Conference on Acoustics, Speech, and Signal Processing, in 2005, and the Prof. Newton Maia Young Scientist Award, in 2001.



Wai-Yip Chan, also known as Geoffrey Chan, received the B.Eng. and M.Eng. degrees from Carleton University, Ottawa, ON, Canada, and the Ph.D. degree from the 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. He has held positions in academia and industry: McGill University, Montreal, QC, Canada, Illinois Institute of Technology, Chicago, Bell Northern Research, Ottawa, and Communications Research Centre, Ottawa. His research interests are in multimedia coding and communications. He is an Associate Editor of the *EURASIP Journal on Audio, Speech, and Music Processing*

Dr. Chan received a CAREER Award from the National Science Foundation. He has helped organize IEEE-sponsored conferences in speech coding, image processing, and communications.