Single-Ended Quality Measurement of Noise Suppressed Speech Based on Kullback-Leibler Distances

Tiago H. Falk, Hua Yuan and Wai-Yip Chan
Department of Electrical and Computer Engineering
Queen's University, Kingston, Ontario, Canada
E-mail: {falkt, yuanh, chan}@ee.queensu.ca

***Abstract*— In this paper, a single-ended quality measurement algorithm for noise suppressed speech is described. The proposed algorithm computes fast approximations of Kullback-Leibler distances between Gaussian mixture (GM) reference models of clean, noise corrupted, and noise suppressed speech and a GM model trained online on the test speech signal. The distances, together with a spectral flatness measure, are mapped to an estimated quality score via a support vector regressor. Experimental results show that substantial improvement in performance and complexity can be attained, relative to the current state-of-art single-ended ITU-T P.563 algorithm. Due to its modular architecture, the proposed algorithm can be easily configured to also perform signal distortion and background intrusiveness measurement, a functionality not available with current standard algorithms.**

***Index Terms*— Single-ended measurement, speech quality, Gaussian mixture model, Kullback-Leibler distance, noise suppression.**

I. Introduction

With the advances in speech communication technologies, noise suppression has become essential for applications such as hearing aids, mobile phones, and voice-controlled systems. Today, algorithms such as the selectable mode vocoder (SMV) [1] are capable of performing noise suppression with minimal detrimental effects to the speech signal. The performance of noise suppression algorithms can be assessed via subjective listening tests ([2], [3]). Subjective testing, however, is very expensive, time consuming, and not suitable for real-time 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 classified as either double- or single-ended, given a clean reference signal is required or not, respectively. The block diagrams depicted in Fig. 1 illustrate the two objective measurement paradigms.

In the past, various double-ended measures have been proposed to characterize the performance of noise suppression algorithms (e.g., [4]). These measures, however, did not take into account human perceptual characteristics, thus, did not correlate well with subjective quality. Recently, widely used double-ended objective measures were tested as quality estimators of noise suppressed speech (e.g., [5], [6]). Included in the measures was the current state-of-art International Telecommunications Union ITU-T double-ended algorithm, P.862 (PESQ). Low correlations with subjective quality were reported for most measures. Single-ended measurement, on the other hand, is a more recent research field, thus not many measures have been proposed. We have experimented with the current state-of-art ITU-T P.563 algorithm and low correlations with subjective quality have also been found [7]. In summary, to date, a generally accepted evaluation metric for noise suppressed speech, be it double- or single-ended, is still *not* available.

In this paper, our recent advances in single-ended quality measurement are described. In particular, an algorithm that is suitable for noise suppressed speech is presented. The proposed algorithm is based on a fast approximation of the Kullback-Leibler distance (KLD) between Gaussian mixture (GM) reference models of clean, noise corrupted, and noise suppressed speech and a GM model trained online for the test speech signal. Experiments suggest that KLD can be employed as an effective measure of speech quality not only for noise suppressed speech but also for other commonly encountered degradation conditions.

The remainder of this paper is organized as follows. In Section II, a description of subjective and objective testing methodologies is given; Section II-C focuses on our current advances in single-ended quality measurement. In Section III, a detailed description of the proposed algorithm is given. Algorithm design considerations are covered in Section IV and algorithm performance is evaluated in Section V. Section VI describes applications of the proposed algorithm beyond the realm of speech quality estimation; in particular, degradation classification and characterization of background intrusiveness and signal distortion is investigated. Lastly, conclusions are given in Section VII.

II. Speech Quality Measurement

In this section, a brief overview of subjective and objective speech quality measurement is presented in Section II-A and Section II-B, respectively. Section II-C describes our recent advances in single-ended objective quality measurement.

A. Subjective Measurement

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

TABLE I.
SUBJECTIVE RATING SCALE FOR OVERALL QUALITY (ACR), SIGNAL DISTORTION (SIG) AND BACKGROUND INTRUSIVENESS (BCK)

Rating	ACR	SIG	BCK
5	Excellent	Not Distorted	Not Noticeable
4	Good	Slightly Distorted	Slightly Noticeable
3	Fair	Somewhat Distorted	Noticeable but Not Intrusive
2	Poor	Fairly Distorted	Somewhat Intrusive
1	Unsatisfactory	Very Distorted	Very Intrusive

measure speech quality on the integer absolute category rating (ACR) scale shown in Table I (column labeled ACR). The average of the listener scores is the subjective mean opinion score, MOS [8]. With the advances of noise suppression algorithms, unwanted artifacts such as “musical noise” arise. It is unknown how humans integrate the individual contributions of speech and noise distortions when judging the overall quality of a noise suppressed signal. As a result, the more recent ITU Recommendation P.835 instructs listeners to successively attend to and rate three different signal components of the noise suppressed speech signal. These three components are: (1) the speech content alone using the five-point scale of signal distortion shown in column labeled “SIG” of Table I, (2) the background content alone using the five-point scale of background intrusiveness shown in column labeled “BCK,” and (3) the overall speech-plus-noise content using the five-point ACR scale.

Throughout most of this manuscript, focus is given to overall speech quality estimation. In Section VI-B, an extension to the proposed algorithm is presented which allows for estimation of not only overall quality but also signal distortion (SIG) and background intrusiveness (BCK). In summary, subjective testing, despite being the most reliable method of measuring speech quality, is costly and not suitable for real-time applications. As a consequence objective measurement methods are often preferred in practice. Next, a brief overview of objective quality measurement is given.

B. Objective Measurement: Brief Overview

As mentioned previously, objective quality measurement can be classified as double or single-ended (Fig. 1 (a) and (b), respectively). Double-ended measures depend on some form of distance metric between the input (clean) and output (degraded) speech signals to estimate the subjective MOS. 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 situations where the input is noisy and the system being tested is equipped with a noise suppression algorithm. In fact, ITU-T Recommendation P.862.3 [9] states that “the use of PESQ with systems that include noise suppression algorithms is not recommended.”

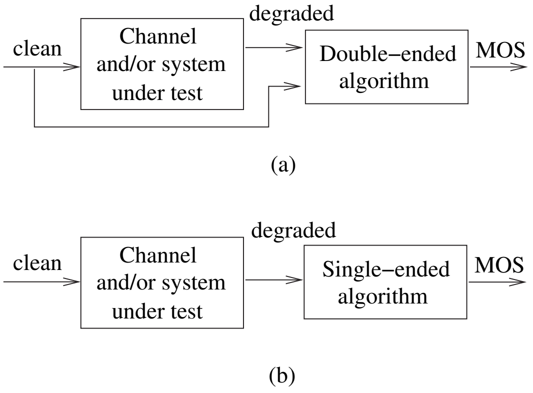

Figure 1. Block diagram of (a) double-ended and (b) single-ended speech quality measurement.

Moreover, double-ended objective measures have the disadvantage of requiring access to a clean reference signal. Often times, when noise suppression algorithms are used, only the noise corrupted signal is available, along with its enhanced counterpart. One of our recent experiments has shown that PESQ performance is compromised if the clean reference signal is unavailable and the noisy signal is used as reference instead. A decrease in performance of approximately 33% is attained if the noisy signal is used as reference, relative to using the clean signal [7].

Problems associated with the “validity” of the reference signal, such as those described above, can be avoided with the use of single-ended quality measurement algorithms. Currently, ITU-T standard P.563 is considered the state-of-art single-ended measurement algorithm [10]. The documentation describing P.563 states that it *is* suitable for transmission systems that include a noise suppression algorithm. Our previous experiments [7], however, and the results described herein, suggest otherwise. Next, we give a general overview of the technologies and methods we have been using to perform single-ended quality measurement. Our most recent scheme is shown to provide accurate quality estimation performance not only for noise suppressed speech but also for other commonly encountered degradation conditions. A detailed description of the proposed algorithm is given in Section III.

C. Objective Measurement: Recent Advances

Our research into single-ended quality measurement entails comparisons between the test speech signal and

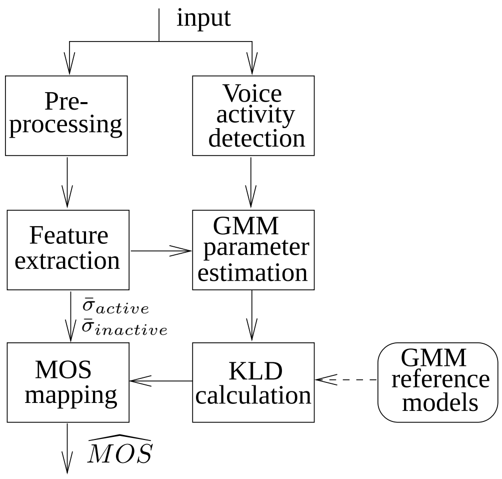

Figure 2. Architecture of the proposed algorithm.

normative behavior of speech signals. Statistical models, in particular, Gaussian mixture models (GMM), are used to generate artificial reference models of speech behavior. In [11], features of the received speech signal are compared to reference GMMs obtained from clean speech signals. A more robust system is achieved by equipping the algorithm with information regarding the behavior of speech degraded by different transmission and/or coding schemes as well as the behavior of clean speech ([12],[7]). In both schemes, the comparison between the test signal and the reference model is achieved by means of a consistency measure (normalized log-likelihood). A consistency value is computed with respect to each of the clean and degraded reference models and the values are then mapped to an objective MOS.

Here, the proposed algorithm is updated to incorporate reference models of noise suppressed speech signals. Moreover, we replace the consistency measure with a fast approximation of the Kullback-Leibler distance (KLD) [13] between the reference models and a model trained online from the test speech signal. Our experiments have suggested that the KLD serves to provide an effective indicator of quality for noise suppressed speech. A description of the proposed algorithm and the motivation behind the use of KLDs is described next.

III. Algorithm Description

The overall architecture of the proposed algorithm is depicted in Fig. 2. First, the level of the speech signal is normalized and the signal is filtered to simulate the handsets used in listening tests. Perceptual features are then extracted from the test speech signal every 10 milliseconds. The voice activity detector (VAD) labels the feature vector of each frame as either active or inactive (background noise). Offline, three reference models are created. High-quality undistorted speech signals, signals corrupted by additive noise at low signal-to-noise ratios (SNR), and noise suppressed speech signals are used to produce reference models of the behavior of clean, noisy, and noise suppressed speech features, respectively.

In all cases, the probability distribution of the features is modeled with a Gaussian mixture model; separate models are trained for active and for inactive frames. Online, the expectation-maximization algorithm is used to estimate a GM model for the features extracted from the test signal. To achieve low-complexity processing, an approximation of the KLD is used. KLDs are computed between the online estimated models and the three reference models. The calculated distances, together with a spectral flatness measure, serve as speech quality indicators and are mapped to an estimated mean opinion score, $\widehat{MOS}$ [3]. A detailed description of each block is provided in the remainder of this section.

A. Pre-processing and VAD

The pre-processing module performs level normalization and intermediate reference system (IRS) filtering. The level of the speech signal is normalized to -26 dBov using the P.56 speech voltmeter [14] and the modified IRS filter is applied to emulate the characteristic of the handset used in listening tests (see description in [15]). Voice activity detection (VAD) is employed to label speech frames as active or inactive. In our previous research, a voicing detector was also used to further label active frames as "voiced" or "unvoiced". Voicing decision is not carried out here as, in our experiments with noise suppressed speech, this extra processing did not garner substantial improvement in estimation performance. The VAD from the adaptive multi-rate (AMR) speech codec is used [16].

B. Feature Extraction

Perceptual linear prediction (PLP) cepstral coefficients [17] serve as primary features and are extracted from the speech signal every 10 milliseconds. We have experimented with different perceptual features (e.g., mel frequency cepstral coefficients, RASTA-PLP [18], PLP), feature representations (e.g., direct-form coefficients, cepstral coefficients, line spectral frequencies), and model orders. Fifth order PLP cepstra is chosen as it strikes a balance between performance and complexity.

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 p^{th} order PLP cepstral coefficients $\mathbf{x} = \{x_i\}_{i=0}^{p}$. The zeroth coefficient serves as a measure of the signal (log-)energy [19]. 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 notation $\bar{\mathbf{x}} = \{\bar{x}_i\}_{i=0}^{p}$ represents the PLP vector averaged over a given set of frames.

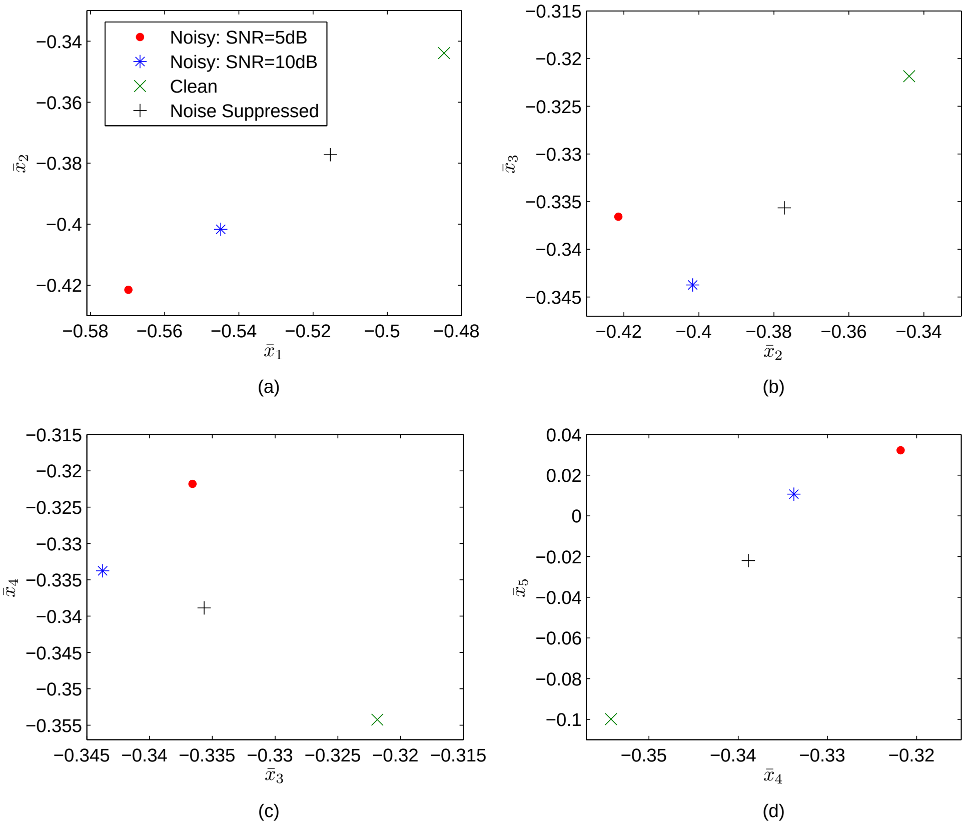

Figure 3. PLP cepstral behavior for clean speech (×), speech corrupted by background noise with an SNR of 5 dB (•) and 10 dB (⋆), and noisy speech processed by a noise reduction algorithm (+). Cepstral coefficients are averaged over 1000 active speech frames. The plots depict (a) $\bar{x}_2$ versus $\bar{x}_1$, (b) $\bar{x}_3$ versus $\bar{x}_2$, (c) $\bar{x}_4$ versus $\bar{x}_3$, (d) $\bar{x}_5$ versus $\bar{x}_4$.

In preliminary experiments, we found that the behavior of PLP cepstra is affected not only by additive noise (also shown in [20] for linear prediction coefficients) but also by different noise suppression algorithms. Figure 3 (a)-(d) illustrates this behavior for clean, noisy, and noise suppressed speech. The coefficients depicted in Fig. 3 are averaged over one thousand active speech frames. Note that PLP cepstral coefficients lie in distinct areas of the cepstral vector space with lower quality speech (e.g., SNR=5 dB case in Fig. 3) lying further away from the clean speech cluster. As can be seen, similar trends are found for all PLP cepstral coefficients. Different “distances” are obtained for different noise reduction algorithms and different noise levels; this serves as motivation to use the Kullback-Leibler distance as an indicator of speech quality.

Lastly, the mean cepstral deviation ($\bar{\sigma}$) of the test signal is computed. The mean cepstral deviation is the average of all “per-frame” deviations (σ_m) of the PLP coefficients (excluding the zeroth coefficient). The deviation for the m^{th} frame is defined as

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

and $p = 5$. Previously, $\bar{\sigma}$ has shown to be related to the spectral flatness of the signal [7]. Here, $\bar{\sigma}$ is calculated for active and inactive frames separately ($\bar{\sigma}_{active}$ and $\bar{\sigma}_{inactive}$, respectively). Our experiments have shown that the spectral flatness measure assists in discriminating between clean, noisy, and enhanced speech; similar findings are reported in [21].

C. GM Reference Models and Parameter Estimation

Gaussian mixture models are used to model the PLP cepstral coefficients of active and of inactive speech frames. A Gaussian mixture density is a weighted sum of M component densities

$$p(\mathbf{x}|\boldsymbol{\lambda}) = \sum_{i=1}^{M} \alpha_i b_i(\mathbf{x}), \tag{2}$$

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

Offline, six different Gaussian mixture densities, $p_{model,class}(\mathbf{x}|\boldsymbol{\lambda})$ are trained. The subscript “model” represents either clean, noisy, or noise suppressed; the subscript “class” represents either active or inactive frames.

Online, the expectation-maximization (EM) algorithm [22] is used to train a GM density on features extracted from the test signal; a separate model is found for active and for inactive frames ($\tilde{p}_{class}(\mathbf{x}|\tilde{\boldsymbol{\lambda}})$). Experiments on our databases show that if the EM algorithm is initialized using the *k-means* algorithm it converges in approximately 17 iterations for the active models and in 7 iterations for the inactive models. Faster implementations or alternate initialization schemes may also be tested; this investigation, however, is left for future study.

D. KLD Calculation and MOS Mapping

The Kullback-Leibler distance measures the "distance" between two probability density functions $p_1(x)$ and $p_2(x)$ by

$$D(p_1, p_2) = \int p_1(x) \ln \frac{p_1(x)}{p_2(x)} dx. \tag{3}$$

$D(p_1, p_2)$ describes how well $p_2(x)$ approximates $p_1(x)$. Here, the KLD is calculated between the online-estimated model ($\tilde{p}$ for short) and the three reference models (p for short), for active and inactive frames. Commonly, the Monte Carlo method is used to compute the integral in (3); this, however, is prohibitively expensive for online quality measurement. We experiment with two fast approximations of the KLD; one (termed $D1$) assumes equal number of Gaussian components between reference and test models $M = \tilde{M}$ [23], while the other ($D2$) allows for $M \neq \tilde{M}$ [24]. $D1$ is given by

$$D1(p, \tilde{p}) = \sum_{i=1}^{M} \alpha_i \log \frac{\alpha_i}{\tilde{\alpha}_i} + \sum_{i=1}^{M} \alpha_i \; D(b_i(\mathbf{x}), \tilde{b}_i(\mathbf{x})) \tag{4}$$

where

$$D(b_i(\mathbf{x}), \tilde{b}_i(\mathbf{x})) = \frac{1}{2}\Big(\log \Big(\frac{\det \tilde{\boldsymbol{\Sigma}}_i}{\det \boldsymbol{\Sigma}_i} \Big) + \text{trace}(\tilde{\boldsymbol{\Sigma}}_i^{-1} \boldsymbol{\Sigma}_i) + (\tilde{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)^T \tilde{\boldsymbol{\Sigma}}_i^{-1} (\tilde{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i) - K \Big) \tag{5}$$

is the KLD between two K-variate Gaussian densities. The approximation described in [24] is posed as a linear programming problem. Many algorithms are available to solve the problem efficiently, however, they are often complex and time consuming. A simplification is performed and $D2$ is defined here as

$$D2(p, \tilde{p}) = \sum_{i=1}^{M} \sum_{j=1}^{\tilde{M}} \alpha_i \tilde{\alpha}_j D(b_i(\mathbf{x}), \tilde{b}_j(\mathbf{x})). \tag{6}$$

Note from (3)-(6) that $D1$ and $D2$ are asymmetric measures, i.e., $D(p, \tilde{p}) \neq D(\tilde{p}, p)$. We symmetrize the measures according to [25], i.e.,

$$D_{sym}(p, \tilde{p}) = \frac{1}{\dfrac{1}{D(p, \tilde{p})} + \dfrac{1}{D(\tilde{p}, p)}}. \tag{7}$$

Symmetric measures are termed $D1_{sym}$ and $D2_{sym}$. Performance of the four measures is described in Section V.

As a final step, the six computed KLDs, together with $\bar{\sigma}_{active}$ and $\bar{\sigma}_{inactive}$, are mapped to $\widehat{MOS}$. We experiment with several different candidate mapping functions: linear, multivariate polynomial and support vector regression (SVR). Simulation results showed that a radial basis SVR, with parameters optimized via linear search, provides least estimation error. The results to follow are all based on using SVR. The reader is referred to [26] for a more comprehensive SVR review.

IV. Algorithm Design Considerations

The KLD measure $D1$ described in (4) requires that both the GM reference model and the online-estimated model have the same number of Gaussian components. A larger number of components may hamper online parameter estimation. It is observed that speech databases used for subjective listening-quality assessment contain files that are on average 7 seconds long with an activity ratio of 60-85%. GM models with 6 and 2 components are thus chosen for active and inactive models, respectively. This choice results in a training ratio (ratio between number of frames in the test signal and number of parameters estimated during training) of approximately 10. For the KLD measure $D2$ described in (6), we experiment with reference models with $6 \leq M \leq 16$ for active frames and $2 \leq M \leq 6$ for inactive frames. Superior performance is attained with 10 and 4 components, respectively. Moreover, we allow the number of GMM components for the test signal to vary such that the training ratio is kept above 10. It is observed that for most signals on our test databases (described in Section V) the chosen number of components is 6 and 2, for active and inactive frames, respectively.

V. Experimental Results

In this section, experimental results are presented. Section V-A describes the databases used for training and testing of the proposed algorithm. Section V-B presents the experimental results and Section V-C discusses algorithm computational complexity.

A. Database Description

The NOIZEUS database [5] is used to design the GM reference models. The database is comprised of clean speech, speech corrupted by four types of noise (babble, car, street, and train) at two low SNR levels (5 and 10 dB) and noisy speech processed by 13 different noise suppression algorithms. The noise suppression algorithms fall under four classes: spectral subtractive, subspace, statistical-model based, and Wiener algorithms. A description of the algorithms can be found in [5]. To train the MOS mapping function, a proprietary subjectively scored database is used. The database is comprised of speech corrupted by car and street noise at SNR=15 dB and office noise at SNR=20 dB and processed by the SMV speech codec; a total of 960 speech files are available.

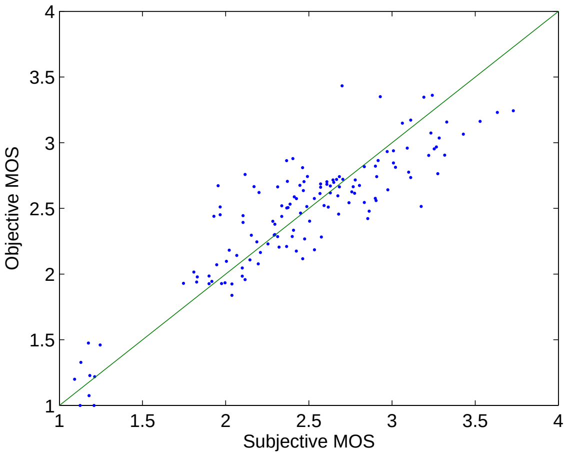

Figure 4. Per-condition objective MOS versus subjective MOS for the combined test datasets using the proposed algorithm.

Three datasets not used in training (i.e., unseen) are used for testing. The first dataset (DS1) is comprised of speech corrupted by four noise sources (babble, car, street and hoth) at three SNR levels (0 dB, 10 dB, and 20 dB). The second (DS2) has noisy speech files (babble, street, car) at three SNR levels (0 dB, 10 dB, and 20 dB) processed by two noise suppression algorithms (SMV and Adobe Audition® with its "reduction level" parameter set to 75%). The third (DS3) is comprised of noisy speech signals (car, hoth, babble at 10 dB and 20 dB) processed by three speech codecs (G.711, G.729, and AMR) with packet loss concealment (PLC) capabilities. Random and bursty losses are simulated at 2% and 4%. A silence insertion concealment scheme is also present. Dataset DS3 is used to test the robustness of the proposed algorithm to alternate (unseen) methods of speech enhancement, in particular, packet loss concealment. The combined three test datasets consist of 1080 speech files covering 135 degradation conditions.

B. Test Results

Table II presents "per-condition" correlation (R) and root-mean-square error (ϵ) between subjective MOS and P.563 objective MOS, for the three unseen datasets. Results are obtained after 3^{rd} order monotonic polynomial regression, as recommended in [10]. The table also reports the percentage improvement, relative to P.563, attained by the proposed algorithm for the four KLD measures described in Section III-D. The columns labeled "$\%R$" and "$\%\epsilon$" list the percentage increase in R and percentage reduction in ϵ, respectively. For measure $D2_{sym}$, R and ϵ are also shown to ease comparison. As can be seen, the proposed algorithm outperforms P.563 on all three datasets; as much as 60% increase in R and 37% decrease in ϵ can be attained. The plot in Fig. 4 depicts objective MOS ($D2_{sym}$) versus subjective MOS where each data point represents one of the 135 degradation conditions available in the combined test dataset.

Furthermore, it is observed that for datasets DS1 and DS2, similar performance is attained for asymmetric and symmetric measures. This is due to the fact that when p and $\tilde{p}$ are similar (i.e., test signal is "consistent" with one of the reference models, as expected for DS1 and DS2) the KLD takes on small values and $D(p,\tilde{p}) \approx D(\tilde{p},p) \approx D_{sym}(\tilde{p},p)$. On the other hand, when a test signal is not as consistent with the reference model (e.g., noisy speech processed by a PLC algorithm, as in DS3) the KLD takes on larger values and $D(p,\tilde{p}) \neq D(\tilde{p},p)$. In this case, the symmetric measure performs better. Another example of this behavior can be observed with unseen test signals corrupted by speech-correlated noise (MNRU); measure $D2$ results in $R = 0.782$ and $\epsilon = 0.685$ while $D2_{sym}$ in $R = 0.955$ and $\epsilon = 0.326$. For comparison purposes, P.563 achieves $R = 0.9142$ and $\epsilon = 0.443$.

C. Algorithm Processing Time

Processing time is also an important figure of merit for gauging algorithm performance. We use the ANSI-C reference implementation of P.563. With the exception of the VAD algorithm (taken from the ANSI-C reference implementation of the AMR codec), the remainder of 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. Here, processing time is defined as the time it takes to process ten speech files randomly selected from the three unseen test sets. The ten files combined have a total length of 57.77 seconds. For P.563, a processing time of 13.75 seconds is attained. The proposed algorithm (using $D2_{sym}$) has a processing time of 9.04 seconds, an approximate 35% reduction. A slight decrease in processing time of 0.15 seconds can be attained by using $D1_{sym}$. Note that a complete C implementation of the proposed algorithm would further increase the speedup.

Table III describes the percentage of the total processing time used by each module in the proposed algorithm. As can be seen, the computational complexity of the proposed algorithm is mainly attributable to voice activity detection and level normalization and IRS filtering. A more efficient VAD algorithm and implementation would further decrease algorithm processing time. Experiments also show that only a slight decrease in performance is attained if level normalization and IRS filtering is not performed; this can result in a 44% reduction in processing time relative to P.563.

VI. Beyond Subjective Quality Estimation

Due to the properties of the PLP cepstra described in Section III-B, the KLD can serve purposes other than MOS estimation. Sections VI-A and VI-B describe two alternate applications; namely, degradation classification and component quality estimation, respectively.

A. Experimenting with Degradation Classification

In some instances (e.g., testing the applicability of a double-ended algorithm) it is desirable to detect if noise

TABLE II.
PERFORMANCE OF P.563 AND THE PROPOSED ALGORITHM ON THREE UNSEEN DATASETS

Unseen Dataset	P.563		$D1$		$D1_{sym}$		$D2$		$D2_{sym}$			
	R	ϵ	$\%R$	$\%\epsilon$	$\%R$	$\%\epsilon$	$\%R$	$\%\epsilon$	R	$\%R$	ϵ	$\%\epsilon$
DS1	0.838	0.355	8.1	22.6	6.6	17.8	8.7	24.3	0.926	10.4	0.246	30.6
DS2	0.631	0.492	31.1	27.6	32.6	29.5	35.1	32.6	0.865	37.0	0.318	35.2
DS3	0.527	0.293	38.2	19.4	52.1	29.6	49.5	27.6	0.846	60.5	0.183	37.2
Average	–	–	**34.6**	**23.5**	**42.4**	**29.5**	**42.3**	**30.1**	–	**48.8**	–	**36.2**

TABLE III.
ALGORITHM PROCESSING TIMES

Processing Module	Time (s)	%
Level normalization & IRS	1.30	14.4
PLP calculation	0.91	10.1
Cepstral deviation calculation	0.01	0.1
Voice activity detection	5.90	65.3
GMM parameter estimation (EM)	0.68	7.5
KLD calculation & MOS mapping	0.24	2.6
Total	**9.04**	**100**

suppression has occurred or if the input signal is noisy. As a simple proof-of-concept experiment, a three-node classification tree [27] is designed to detect whether a signal is noisy or if it has been processed by a noise suppression algorithm. In this experiment, the KLD is computed only between the online derived model and the clean reference model, for both active and inactive frames. We test the designed tree on 96 *unseen* speech signals: 48 noise-suppressed signals and 48 signals corrupted by babble and car noise at 0 dB and 5 dB; all 96 signals were correctly detected.

B. Component Quality Estimation

It is known that certain noise suppression algorithms can introduce unwanted artifacts such as "musical noise." As mentioned in Section II, with Recommendation P.835, noise suppressed signals are rated based on speech content alone (SIG), on background noise alone (BCK), and on speech plus noise content (OVRL). Currently, objective measurement algorithms (both single- and double-ended) can only attempt to estimate overall quality. Devising an algorithm capable of also estimating signal distortion and background intrusiveness 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 whilst reducing background noise to nonintrusive levels. In [5], the NOIZEUS database is used to evaluate six double-ended objective estimates of SIG and BCK ($\widehat{SIG}$ and $\widehat{BCK}$, respectively). The study makes use of the original clean signal as a reference and low correlations with subjective quality are reported ($R < 0.65$).

Due to the modular architecture of the proposed algorithm, a simple extension can be implemented to allow for single-ended measurement of BCK and SIG. In particular, two new SVR mapping functions are obtained. To estimate signal distortion, a 4-dimensional SVR is devised to map the KLDs computed from active frames (relative to the three reference models) and $\bar{\sigma}_{active}$ into $\widehat{SIG}$. To estimate background intrusiveness, a 5-dimensional SVR is designed to map the KLDs computed from inactive frames, $\bar{\sigma}_{inactive}$, and an estimated SNR to $\widehat{BCK}$. Here, we use the SNR estimated by the AMR VAD algorithm.

Since only the NOIZEUS database contains subjective SIG and BCK scores, we use 10-fold cross validation to measure the performance of the proposed scheme. The NOIZEUS database is randomly divided into 10 data sets of almost equal size. Training and testing is performed in 10 trials, where, in each trial, one of the data sets serves as a test set and the remaining 9 are combined to serve as a training set. Each data set serves as a test set only once. The ten resulting R's and ϵ's are averaged to obtain the cross-validation performance figures. The proposed method attains $R = 0.80$, $\epsilon = 0.33$, and $R = 0.74$, $\epsilon = 0.39$, for SIG and BCK, respectively. The results are encouraging given that the original clean signal is *not* available as a reference.

VII. CONCLUSION

A low-complexity single-ended speech quality estimation algorithm is proposed. The algorithm provides superior quality estimates relative to P.563 for several commonly encountered distortions, such as those caused by noise suppression or packet loss concealment algorithms. It is also demonstrated that, besides offering the conventional function of measuring the overall quality of a noise suppressed speech signal, the algorithm is also capable of measuring signal distortion and background intrusiveness. This functionality is not available with current state-of-art ITU-T standard algorithms.

REFERENCES

[1] 3GPP2 C.S0030-0, "Selectable mode vocoder (SMV) service option for wideband spread spectrum communication systems," Jan. 2004.

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

[3] ITU-T P.835, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithms," Intl. Telecom. Union, 2003.

[4] V. Mattila, "Objective measures for the characterization of the basic functioning of noise suppression algorithms," in *Proc. Int. Conf. on Measurement of Speech and Audio Quality in Networks*, May 2003.

[5] Y. Hu and P. Loizou, "Evaluation of objective measures for speech enhancement," in *Proc. of the Int. Conf. on Spoken Language Processing*, 2006, pp. 1447–1450.

[6] T. Rohdenburg, V. Hohmann, and B. Kollmeir, "Objective perceptual quality measures for the evaluation of noise reduction schemes," in *Proc. 9th Intl. Workshop on Acoustic Echo and Noise Control*, 2005, pp. 169–172.

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

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

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

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

[11] T. H. Falk, Q. Xu, and W.-Y. Chan, "Non-intrusive GMM-based speech quality measurement," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, vol. 1, March 2005, pp. 125–128.

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

[13] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.

[14] ITU-T P.56, "Objective measurement of active speech level," Intl. Telecom. Union, Switzerland, 1993.

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

[16] 3GPP2 TS 26.094, "Adaptive multi-rate (AMR) speech codec: voice ativity detector (VAD), release 6," Dec. 2004.

[17] H. Hermansky, "Perceptual linear prediction (PLP) analysis of speech," *J. Acoust. Society of America*, vol. 87, pp. 1738–1752, April 1990.

[18] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. on Speech and Audio Processing.*, vol. 2, no. 4, Oct. 1994.

[19] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing*. Prentice-Hall, 2001.

[20] T.-H. Hwang, L.-M. Lee, and H.-C. Wang, "Cepstral behaviour due to additive noise and a compensation scheme for noisy speech recognition," *IEE Proc. - Vision, Image, and Signal Proc.*, vol. 145, no. 5, pp. 316–321, Oct. 1998.

[21] V. Grancharov, J. Plasberg, J. Samuelsson, and B. Kleijn, "Generalized postfilter for speech quality enhancement," *IEEE Trans. on Audio, Speech and Language Proc.*, to appear.

[22] A. Dempster, N. Lair, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statistical Society*, vol. 39, pp. 1–38, 1977.

[23] M. Do, "Fast approximation of Kullback-Leibler distance for dependence trees and hidden markov models," *IEEE Signal Proc. Letters*, vol. 10, no. 4, pp. 115–118, April 2003.

[24] Z. Liu and Q. Huang, "A new distance measure for probability distribution function of mixture type," in *Proc. Intl. Conf. on Acoustics, Speech, and Signal Proc.*, vol. 1, June 2000, pp. 616–619.

[25] D. Johnson and S. Sinanovic, "Symmetrizing the Kullback-Leibler distance," Rice University, Tech. Rep., 2001.

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

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

Tiago H. Falk was born in Recife, Brazil, in September 1979. He received the B.Sc. degree from the Federal University of Pernambuco, Brazil, in 2002, and the M.Sc. (Eng.) degree from Queens University, Canada, in 2005, all in electrical engineering. He is currently pursuing the Ph.D. degree at Queen's University, Canada. His research interests include multimedia quality measurement, speech enhancement, and speaker recognition. Mr. Falk is a board member of the International Speech Communication Association Student Advisory Committee, a Student member of the Institute of Electrical and Electronics Engineers, and a Student Member of the Brazilian Telecommunication Society. He 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.

Hua Yuan was born in Shanghai, China, in 1974. He received the B.S. degree from Fudan University, China, and the M.A.Sc. degree from Ryerson University, Toronto, Canada, in 1997 and 2004, respectively, both in electrical engineering. He is currently pursuing the Ph.D. degree at Queen's University, Kingston, Canada. From 1997 to 2001, he was with Alcatel Shanghai Bell Corporation, as a Software Developer in the Mobile Communication Department. He had been involved in mobile core network systems research and development. He was a Research Assistant with the Communications and Signal Processing Applications Laboratory (CASPAL) at Ryerson University from 2002 to 2004. Since January 2005, he has been with the Multimedia Coding and Communications Laboratory (Mc^{2}L) at Queens University. His research interests include speech processing, perceptual audio coding, image processing and retrieval.

Wai-Yip Chan, also known as Geoffrey Chan, received his B.Eng. and M.Eng. degrees from Carleton University, Ottawa, and his 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. He has held positions in academia and industry, namely: McGill University, Illinois Institute of Technology, Bell Northern Research, and Communications Research Centre. His research interests are in multimedia signal coding and communications. He is an associate editor of EURASIP Journal on Audio, Speech, and Music Processing. He has helped organize IEEE sponsored conferences in speech coding, image processing, and communications. He held a CAREER Award from the National Science Foundation.