Nonintrusive Speech Quality Estimation Using Gaussian Mixture Models

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

***Abstract*—An algorithm for nonintrusive speech quality estimation based on Gaussian mixture models (GMMs) is presented. GMMs are used to form an artificial reference model of the behavior of features of undegraded speech. Consistency measures between the degraded speech signal and the reference model serve as indicators of speech quality. Consistency values are mapped to an objective speech quality score using a multivariate adaptive regression splines function. When tested on unseen data, the proposed algorithm generally outperforms ITU-T standard P.563, which is the current "state-of-the-art" algorithm. The algorithm computes objective quality scores roughly twice as fast as P.563.**

***Index Terms*—Gaussian mixtures, quality assurance, quality measurement, quality of service, speech coding, speech quality, speech transmission, telephony.**

I. Introduction

SPEECH quality is a major contributor to the telecommunication user's perception of quality of service. As communications networks become more heterogeneous, identifying the root cause of voice quality problems can be a challenging task. The evaluation and assurance of speech quality has, consequently, become critically important for telephone service providers. Traditionally, user opinion is measured offline using slow and costly subjective listening tests. In the most common test [1], listeners rate the speech they just heard on a five-point opinion scale, ranging from "bad" to "excellent." The ratings are assigned integer scores ranging from 1 for "bad" to 5 for "excellent." The average of these scores, termed mean opinion score (MOS), is widely used to characterize the quality of telephony equipment and services.

As an alternative to subjective measurement, machine-automated "objective" measurement provides a rapid and economical means to estimate user opinion and makes it possible to perform real-time speech quality measurement on a network-wide scale. Objective measurement can be performed either intrusively or nonintrusively. Intrusive measurement, also called double-ended or input–output-based measurement, is based on measuring the distortion between the received and transmitted speech signals, often with an underlying requirement that the transmitted signal be of high quality ("clean"). Nonintrusive measurement, also called single-ended or output-based measurement, relies only on the received speech signal to estimate its quality.

Manuscript received July 13, 2005; revised August 30, 2005. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Israel Cohen.

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).

Digital Object Identifier 10.1109/LSP.2005.861598

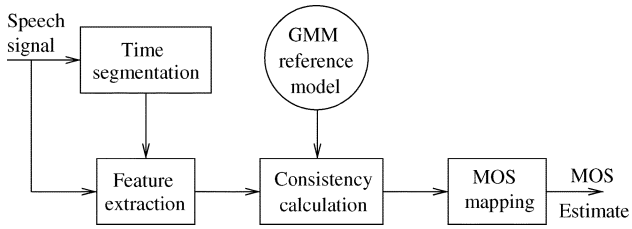

Fig. 1. Architecture of the proposed algorithm.

A handful of nonintrusive measurement schemes have been reported. In [2], comparisons between features of the received speech signal and vector quantizer (VQ) codebook representations of the features of clean speech are used to estimate quality. In [3], the VQ codebook reference is replaced with a hidden Markov model. In [4] and [5], vocal tract modeling and modulation-spectral features derived from the temporal envelope of speech, respectively, provide quality cues for nonintrusive quality measurement. Recently, a nonintrusive method using neurofuzzy inference was proposed [6]. The International Telecommunications Union ITU-T P.563 standard represents the "state-of-the-art" algorithm [7].

This letter presents a novel method [8] of nonintrusive speech quality estimation. Gaussian mixture models (GMMs) trained on features extracted from clean speech signals are used to form a model of normative behavior against which the features of a test speech signal are assessed. A detailed description of the algorithm's functional blocks is presented in Section II. The proposed method is tested on four "unseen" databases and compared to P.563 in Section III.

II. Description of the Proposed Algorithm

The proposed nonintrusive measurement algorithm is designed based on the architecture depicted in Fig. 1. First, perceptual features are extracted from the test speech signal frame by frame. The time segmentation module labels the feature vector of each frame as belonging to one of three possible classes: voiced, unvoiced, or inactive. Offline, high-quality, undistorted speech signals are used to produce a reference model of the behavior of clean speech features. This is accomplished by modeling the probability distribution of the features for each class with a GMM. Features extracted from the test signal are assessed using the reference model, by calculating a "consistency" measure with respect to each GMM. The consistency values serve as indicators of speech quality and are mapped to an estimated MOS value. A detailed description of the algorithm's functional blocks and design considerations are given next.

A. Feature Extraction and Time Segmentation

Perceptual linear prediction (PLP) cepstral coefficients [9] 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: critical band spectral resolution, equal-loudness curve, and intensity loudness power law. The auditory spectrum is approximated by an all-pole autoregressive model, whose coefficients are transformed to PLP cepstral coefficients. 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. As we are interested in measuring quality variation due to the transmission system rather than the speaker, speaker independence is a desirable property. Hermansky [9] suggests that fifth-order PLP coefficients serve well as speaker-independent speech spectral parameters. PLP coefficients were also used in [2].

Time segmentation is employed to separate the speech frames into different classes. It is conjectured that each class exerts different influence on the overall speech quality. 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. The voicing detector further labels active frames as voiced or unvoiced. The VAD from ITU-T G.729B [10], omitting its comfort noise generation functionality, is used here. A more recent improved VAD algorithm may be used to our advantage, though we leave this for future study. Section II-D tests the usefulness of time segmentation.

B. GMM Reference Model

GMMs have been used extensively for speech processing and are introduced here for the sake of notation. Let $\mathbf{u}$ be a K-dimensional vector. A Gaussian mixture density is a weighted sum of M component densities

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

where $\alpha_i \geq 0$, $i = 1, \ldots, M$ are the mixture weights, with $\sum_{i=1}^{M} \alpha_i = 1$, and $b_i(\mathbf{u}), i = 1, \ldots, M$, 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 $\boldsymbol{\lambda}_i = \{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \alpha_i\}$.

GMM parameters are initialized using the k-means algorithm [11] and estimated using the expectation-maximization (EM) algorithm [12]. The EM algorithm iterations produce a sequence of models with monotonically nondecreasing log-likelihood (LL) values. The algorithm is deemed to have converged when the difference of LL values between two consecutive iterations drops below 10^{-3}.

C. Consistency Calculation and MOS Mapping

A GMM is used to model the PLP cepstral coefficients of each class of speech frames. Using clean speech signals, three different Gaussian mixture densities $p_{\text{class}}(\mathbf{u}|\boldsymbol{\lambda})$ are trained. The subscript "class" represents either voiced, unvoiced, or inactive frames. In principle, by evaluating these densities at the degraded PLP cepstral coefficients $\mathbf{x}$ (i.e., $p_{class}(\mathbf{x}|\boldsymbol{\lambda})$), a measure of consistency between the degraded coefficient vector and the clean coefficient model is obtained. Voiced coefficient vectors are applied to $p_{\text{voiced}}(\mathbf{u}|\boldsymbol{\lambda})$, unvoiced vectors to $p_{\text{unvoiced}}(\mathbf{u}|\boldsymbol{\lambda})$, and inactive vectors to $p_{\text{inactive}}(\mathbf{u}|\boldsymbol{\lambda})$.

We make a simplifying assumption that vectors between frames are independent. We expect better performance from more sophisticated approaches that model the statistical dependency between frames, such as Markov modeling. Nevertheless, for the benefit of low computational complexity, we seek to determine how well the simpler approach works. Thus, for a given speech signal, the consistency between the observation and the model is calculated as

$$c_{\text{class}}(\mathbf{x}_1, \ldots, \mathbf{x}_{N_{\text{class}}}) = \frac{1}{N_{\text{class}}} \sum_{j=1}^{N_{\text{class}}} \log\left(p_{\text{class}}(\mathbf{x}_j|\boldsymbol{\lambda})\right) \tag{2}$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_{N_{\text{class}}}$ are the degraded coefficient vectors, and N_{class} is the number of such vectors in the frame class. Larger c_{class} indicates greater consistency. For each class, the product of the consistency measure (2) and the fraction of frames of that class in the speech signal is calculated. The three products for the three classes serve as quality indicators to be mapped to an objective MOS value. In preliminary experiments, we tested two candidate mapping functions: multivariate polynomial regression and multivariate adaptive regression splines (MARS) [13]. With MARS, the mapping is constructed as a weighted sum of basis functions, each taking the form of a truncated spline [14]. Simulation results showed that MARS provides better performance; the results below are all based on using MARS.

D. Algorithm Design Considerations

A preliminary "calibration" experiment was performed to find an effective combination of GMM configuration parameters: M and covariance matrix type. GMMs with 8, 16, or 32 components, and with diagonal or full covariance matrices, are tested. A total of 12 MOS-labeled databases comprised of both clean and degraded speech signals are used. The speech databases include seven ITU-T P-series Supplement 23 (Experiments 1 and 3) multilingual databases [15], two wireless (IS-96A and IS-127 EVRC), and three multilingual databases comprised of speech coded using the ITU-T G.728 speech coder. The databases include speech coded using various standard codecs, speech produced under various channel errors, tandeming, and acoustic noise conditions, and speech degraded by various levels of modulated noise reference unit (MNRU). The combined 12 databases contain 5624 speech file pairs with subjective MOSs ranging from 1 to 4.83. All clean speech files are used for training of the GMMs. We randomly select 90% of the degraded speech files for training of the MARS mapping function, and the remaining 10% are left for testing.

The performance of the algorithm is assessed using the correlation (R) and root-mean-square error (RMSE) between subjective MOS and objective MOS [14]. The "calibration" performance results presented in Table I are for MOSs measured on a "per-file" basis. Columns V, VU, and VUI list test performance figures for reference models designed for voiced frames

TABLE I
PERFORMANCE OF FIFTH-ORDER PLP CEPSTRAL COEFFICIENTS

Matrix type - M	V		VU		VUI	
	R	$RMSE$	R	$RMSE$	R	$RMSE$
Diagonal-8	0.758	0.523	0.816	0.462	0.871	0.392
Diagonal-16	0.763	0.518	0.820	0.458	0.875	0.388
Diagonal-32	0.782	0.499	0.813	0.466	0.880	0.371
Full-8	0.773	0.507	0.825	0.453	0.882	0.377
Full-16	0.774	0.507	0.834	0.441	0.876	0.386

only, voiced and unvoiced frames, and all three frame types, respectively. The largest improvement occurs when all three frame types are taken into consideration. The configuration with 16 Gaussian components and diagonal covariance matrices is deemed to provide a good compromise between accuracy and complexity.

The results presented in Table I are based on an equal number of Gaussian components M across all three frame classes. For clean speech, inactive frames have virtually no signal energy. Such frames ought to be modeled with fewer Gaussian components than voiced or unvoiced frames. A second calibration experiment is performed with 16-component diagonal GMMs for voiced and unvoiced frames and two-, four-, or eight-component diagonal GMMs for inactive frames. The performance results demonstrate that little gain is attained by using higher order GMMs. The simplest configuration with a two-component GMM for inactive frames is preferred.

Note that Gaussian mixture modeling is a data-driven approach that requires a considerable amount of training data. An inherent advantage of the proposed algorithm is that the GMMs are designed using clean speech, which can be obtained in a large quantity at relatively little cost. Subjectively scored (degraded) speech is costly to produce. A smaller amount of such speech can be used for training the MARS mapping function, which has fewer parameters than the GMMs.

III. TEST RESULTS

The proposed algorithm, with a reference model using 16-component diagonal GMMs for voiced and unvoiced frames and a two-component diagonal GMM for inactive frames, is compared to P.563 using four MOS-labeled databases. None of the speech material in these four databases has been applied to the design of the proposed algorithm. The first database [16] contains speech coded with a variety of wireline and wireless codecs. The three other databases comprise speech coded with the 3GPP2 selectable mode vocoder (SMV), standardized as the cdma2000 speech coder.

A. Mixed Database

We use the mixed database described in [16] to assess the performance of the algorithms for different distortion classes. The database contains 240 subjectively scored speech files, covering a total of 60 distortion conditions, grouped into seven major distortion classes. In Table II, performance results for the distortion classes are expressed in terms of RMSE and average absolute error (AAE). The results are obtained after third-order monotonic polynomial regression, applied to eliminate offsets and nonlinearities between the objective and subjective MOS scales, as recommended in [7]. The AAE statistic offers an additional perspective, whereas the correlation coefficient values rank similarly as the MSE values and are omitted for brevity.

The results show that the proposed algorithm achieves lower or comparable RMSE and AAE relative to P.563, with the proposed algorithm looking somewhat more favorable in terms of RMSE than AAE performance. A significant exception occurs with the distortion class containing temporally shifted and front-end clipped speech signals. For such distortions, the poorer performance of the proposed algorithm is not surprising as P.563 is equipped with a functional block that tests for speech interruptions, muting, and time clippings; our scheme currently does not feature this capability. Nonetheless, the results demonstrated are promising, and additional functionalities are left for future study.

B. SMV Databases

Each of the SMV databases comprises 48 different degradation conditions, distributed over 3072 subjectively scored files. Database 1 encompasses tandeming and nominal input level conditions; database 2 covers channel impairments, and database 3 covers noisy environment conditions. SMV is a newer codec than the codecs represented in the training databases. Evaluation using the SMV databases demonstrates the applicability of the proposed algorithm to emerging codec technologies [5].

For this experiment R, RMSE, and algorithm processing times are used to gauge algorithm performance. Table III presents R and RMSE, after third-order monotonic polynomial regression, of "per-condition" MOSs for each database [14]. The column labeled "$\%\uparrow$" lists percentage improvement in R obtained by using the proposed GMM-based method over P.563. The percentage improvement is given by

$$\%\uparrow = \frac{R_{\mathrm{GMM}} - R_{P.563}}{1 - R_{P.563}} \times 100\% \tag{3}$$

and indicates percentage reduction of P.563's performance gap to perfect correlation. The column labeled "$\%\downarrow$" lists

TABLE II
PERFORMANCE ON MIXED DATABASE

Distortion Classes	*RMSE*		*AAE*	
	Proposed	P.563	Proposed	P.563
MNRU	0.571	0.678	0.451	0.586
Clear Channel	0.572	0.676	0.420	0.491
Tandeming	0.537	0.577	0.453	0.424
Bit Errors	0.444	0.633	0.283	0.472
Frame Errors	0.414	0.422	0.378	0.349
Amplitude Variations	0.087	0.080	0.136	0.083
Temporal Clippings	0.550	0.428	0.849	0.421

TABLE III
PERFORMANCE ON SMV DATABASES

Database	P.563		Proposed			
	R	RMSE	R	%↑	RMSE	%↓
1	0.863	0.253	0.928	47.1	0.187	26.1
2	0.835	0.274	0.814	-12.7	0.290	-5.9
3	0.748	0.273	0.910	64.3	0.171	37.4

percentage reduction in P.563's RMSE by using the proposed scheme.

Note that the proposed algorithm outperforms P.563 on databases 1 and 3 by as much as 64% in R and 37% in RMSE. The results for database 3 suggest that the proposed method may be more effective than P.563 for speech in noisy environment conditions. The algorithm achieves somewhat poorer results than P.563 on database 2. Degradation conditions in database 2 encompass frame errors (0, 1, 3, and 5% frame error rate). The results in Table II also suggest that the proposed algorithm may underperform P.563 for such degradation conditions.

Finally, processing time is used to gauge computational complexity. Here, the ANSI-C reference implementation of P.563 is used. Computation time for the proposed algorithm comprises the time for feature extraction and time segmentation and calculation of the consistency values and the MARS mapping. The ANSI-C implementation of the G.729B VAD algorithm is used. The remainder of the algorithm is implemented using Matlab version 6.5 Release 13. P.563 is clearly advantaged in this comparison. Simulations are run on a PC with a 2.8-GHz Pentium 4 processor and 2 GB of RAM. Processing times for the two algorithms are shown in Table IV. For this comparison, three files are randomly selected, one from each of the three SMV databases. The processing times for the proposed algorithm are expressed as percentage reduction in processing time relative to P.563. Note that the proposed algorithm is capable of reducing the processing time of P.563 by roughly 35%–45%. A complete C implementation of the proposed algorithm would surely reduce its computation time; a 50% reduction would be quite achievable. Clearly, the proposed method offers low complexity and accurate measurement of speech quality.

TABLE IV
ALGORITHM PROCESSING TIMES—SMV DATABASES

Database	File length (seconds)	P.563 (seconds)	Proposed (% reduction)
1	7.20	4.37	36.8
2	5.89	4.10	44.7
3	7.91	5.50	44.6

IV. CONCLUSION

A novel nonintrusive speech quality estimation algorithm is proposed based on GMMs. The algorithm provides competitive quality estimates relative to the current "state-of-the-art" algorithm while requiring considerably lower computational complexity. Further testing and refinement will likely lead to a more comprehensive and robust algorithm. Moreover, the simplicity and modular architecture of the proposed algorithm makes it readily adaptable to wideband speech quality estimation.

REFERENCES

[1] ITU-T Rec. P.800, Methods for Subjective Determination of Transmission Quality, Int. Telecommun. Union, Geneva, Switzerland, Aug. 1996.

[2] C. Jin and R. Kubichek, "Vector quantization techniques for output-based objective speech quality," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, May 1996, pp. 491–494.

[3] W. Li and R. Kubichek, "Output-based objective speech quality measurement using continuous hidden Markov models," in *Proc. 7th Int. Symp. Signal Processing Applications*, vol. 1, Jul. 2003, pp. 389–392.

[4] 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.

[5] 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.

[6] G. Chen and V. Parsa, "Nonintrusive speech quality evaluation using an adaptive neurofuzzy inference system," *IEEE Signal Process. Lett.*, vol. 12, no. 5, pp. 403–406, May 2005.

[7] ITU-T P.563, Single Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony Applications, Int. Telecommun. Union, Geneva, Switzerland, May 2004.

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

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

[10] ITU-T Rec. G.729-Annex B, A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to Recommendation V.70, Int. Telecommun. Union, Geneva, Switzerland, Nov. 1996.

[11] A. Gersho and R. Gray, *Vector Quantization and Signal Compression*. Norwell, MA: Kluwer, 1992.

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

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

[14] 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.

[15] ITU-T Rec. P. Supplement 23, ITU-T Coded-Speech Database, Int. Telecommun. Union, Geneva, Switzerland, Feb. 1998.

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