NON-INTRUSIVE GMM-BASED SPEECH QUALITY MEASUREMENT

Tiago H. Falk, Qingfeng Xu, and Wai-Yip Chan

Department of Electrical and Computer Engineering
Queen's University, Kingston, Ontario, Canada
E-mail: {falkt, xuq, chan}@ee.queensu.ca

ABSTRACT

We propose a non-intrusive speech quality measurement algorithm based on using Gaussian-mixture probability models of features of undegraded speech signals as an artificial reference model of "clean" speech behaviour. The consistency between the features of the test speech signal and the reference model serves as an indicator of speech quality. Consistency measures are calculated and mapped to an objective speech quality score using a multivariate adaptive regression splines function. Simulation results show that the proposed method offers accurate and yet low-complexity measurement of speech quality.

1. INTRODUCTION

Speech quality is a major contributor to the end user's perception of quality of service. As networks become more heterogeneous and complex, and new technologies interoperate with legacy equipment, 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.

Voice quality is a subjective opinion, based on the user's reaction to the speech signal they actually heard. Subjective methods make use of a listener panel to measure speech quality on an integer scale from 1 to 5, with 1 corresponding to unsatisfactory speech quality and 5 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 frequent or rapid application. These shortcomings can be overcome by using objective measurement methods, which replace the listener panel with a computational algorithm. Objective methods aim to deliver MOSs that are highly correlated with the MOSs obtained from subjective listening experiments.

Objective quality assessment tests can be classified as *intrusive* or *non-intrusive*. Intrusive measurement depends on some form of distance metric between the input (clean) and output (degraded) speech signals to predict the subjective MOS. In some situations an intrusive approach may not be applicable because the input speech signal may be unavailable. Non-intrusive measurement depends only on the degraded speech signal and is a more challenging approach to objective speech quality estimation. Non-intrusive models have been proposed in [2, 3, 4], but only recently has ITU-T released P.563 as its non-intrusive objective quality measurement standard algorithm [5]. P.563 resulted from a collaboration of Psytechnics' NiQA algorithm [6], SwissQual's NiNA [7], and Opticom's P3SQM.

The signal parameterization in P.563 is divided in three independent functional blocks corresponding to the main classes of distortion; they are: vocal tract analysis, high additional noise, and speech interruptions, muting and time clippings. A total of 51 characteristic signal parameters are calculated. Based on a restricted set of 8 key parameters, a dominant distortion class is selected. The key parameters and the selected distortion class are used for adjusting the speech quality model. Furthermore, for each distortion class, a linear combination of parameters is used to generate an intermediate quality rating that, together with other additional signal features are combined to calculate the (raw) objective quality score.

We propose a novel method for non-intrusive objective speech quality assessment. The method is based on comparing features extracted from degraded speech to an artificial reference model of clean speech signals. The degree of mismatch serves as an indicator of speech quality. The reference model employs Gaussian mixture (GM) models trained on features extracted from a dataset of clean speech signals. Simulation shows that our approach offers accurate and yet low-complexity measurement of speech quality.

2. GMM-BASED NON-INTRUSIVE SPEECH QUALITY ESTIMATION

The proposed non-intrusive measurement algorithm is designed based on the architecture depicted in Fig. 1. Non-intrusive methods do not have access to the clean speech signal. Our approach uses high-quality, undistorted speech

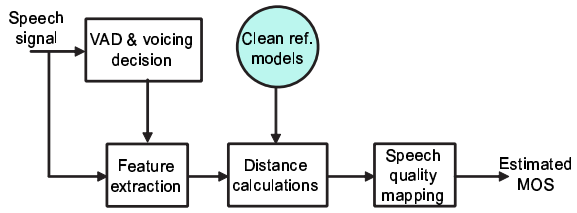

Fig. 1. Algorithm architecture

signals to produce an artificial model of the behavior of clean speech. Features extracted from a test speech signal are compared to the artificial reference model and the degree of mismatch serves as an indicator of speech quality. GMs are used to model the probability distribution of clean speech features.

Hermansky [8] suggests that 5^{th} order Perceptual Linear Prediction (PLP) coefficients can serve as speaker independent speech spectral parameters. PLP analysis uses three psychoacoustic concepts to derive a representation of the auditory spectrum. These concepts are 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 cepstral coefficients. PLP analysis is more consistent with the behavior of the human ear than the traditional linear predictive analysis [8]. PLP analysis is computationally efficient and permits a compact representation of speech.

Furthermore, it is shown in [9] that time segmentation improves objective quality assessment performance. The speech signal is processed through a voice activity detector (VAD) and then a voicing detector. The VAD identifies each speech frame as *active* or *inactive*. The voicing detector further labels active frames as *voiced* or *unvoiced*. Time segmentation separates the different classes of speech frames as they exert different influence on the overall speech quality estimate.

GM models have been used extensively for speech processing and have also shown to be useful in intrusive measurement algorithms [10]. GM models are introduced here only 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

$$p(\mathbf{u}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}) = \sum_{i=1}^{M} \alpha_i b_i(\mathbf{u}) \qquad (1)$$

where $\alpha_i \geq 0, i = 1, ..., M$ are the mixture weights, with $\sum_{i=1}^{M} \alpha_i = 1$, and $b_i(\mathbf{u})$, $i = 1, ..., M$, are the K-variate Gaussian densities with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. We experiment with 8, 16 and 32 Gaussian components, and with diagonal and full covariance matrices. The EM (expectation-maximization) algorithm [11] is used to train the GM densities, i.e., to estimate the weights, means and covariances of the Gaussian components.

GM densities are used to model the PLP coefficients of the different classes of speech frames. Using clean speech signals, we train three different Gaussian mixture densities, $p_{class}(\mathbf{u}|\boldsymbol{\lambda})$. The subscript *class* represents either "voiced", "unvoiced" or "inactive" and we use $\boldsymbol{\lambda}$ to denote $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}\}$ throughout the remainder of this paper. In principle, by evaluating these densities at the degraded PLP coefficients $\mathbf{x}$ (i.e., $p_{class}(\mathbf{x}|\boldsymbol{\lambda})$) we measure how consistent is the degraded prediction coefficient vector with the clean prediction coefficient model. Degraded voiced vectors are applied to $p_{voiced}(\mathbf{u}|\boldsymbol{\lambda})$, unvoiced vectors to $p_{unvoiced}(\mathbf{u}|\boldsymbol{\lambda})$ and inactive vectors to $p_{inactive}(\mathbf{u}|\boldsymbol{\lambda})$.

In practice, for a given speech signal, there is not one degraded PLP coefficient vector, but instead a sequence of degraded vectors. We need to evaluate

$$p_{class}(\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{N_{class}}\}|\boldsymbol{\lambda}), \qquad (2)$$

where N_{class} is the number of degraded PLP coefficient vectors of a specific speech class. Assuming independence of the vectors between frames, the likelihood probability can be expressed as

$$p_{class}(\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{N_{class}}\}|\boldsymbol{\lambda}) = \prod_{j=1}^{N_{class}} p_{class}(\mathbf{x}_j|\boldsymbol{\lambda}). \qquad (3)$$

We measure consistency between the observation and the model using a normalized log-likelihood of the observed data

$$c_{class}(\mathbf{x}) = \frac{1}{N_{class}} \sum_{j=1}^{N_{class}} \log(p_{class}(\mathbf{x}_j|\boldsymbol{\lambda})). \qquad (4)$$

Larger c_{class} indicates greater consistency. For each class, the product of the consistency measure (4) and the fraction of frames of that class in the speech signal is calculated. The three products for the three classes are mapped to objective MOS using multivariate adaptive regression splines (MARS) [12]. MARS models are designed based on the subjective MOSs of degraded speech, as we illustrate in the experimental results presented below.

3. EXPERIMENTAL RESULTS

We compare the proposed GMM-based algorithm to P.563 using speech databases that have been evaluated in MOS tests. The performance of the algorithms is assessed by the correlation between subjective MOS w_i and objective MOS y_i, using Pearson's formula

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

where $\bar{w}$ is the average of w_i, and $\bar{y}$ is the average of y_i. We also use computer processing time as a measure of algorithm complexity.

A total of thirteen databases comprised of both clean and degraded speech signals are used in the training of our algorithm. Clean speech is used for training the GMMs, and degraded speech for training the MARS model. The speech databases include seven ITU-T P-series Supplement 23 multilingual databases [13], two wireless (IS-96A and IS-127 EVRC), a mixed wireline-wireless, and three multilingual databases comprised of speech coded using the ITU-T G.728 speech coder. The databases include speech subjected to various channel errors, tandeming, acoustic noise, and reference conditions. The combined thirteen databases contain 5864 speech file pairs.

For testing, we use three databases comprised of speech coded with the 3GPP2 Selectable Mode Vocoder (SMV). Each of these databases has 3072 subjectively scored degraded speech files. Experiment 1 tests tandeming and nominal input level conditions; experiment 2 tests channel impairments, and experiment 3 noisy environment conditions. None of the speech material in the SMV databases was applied to the design of the GMM-based algorithm.

In [5], it is suggested that offsets and non-linearities between the scales of objective and subjective MOSs be eliminated by applying a 3^{rd} order monotonic function to map the (raw) objective scores onto the subjective scale. Following the suggestion, we design 3^{rd} order polynomial mappings, by performing regression with a monotonicity constraint, between (raw) objective scores and subjective MOSs.

Fig. 2 exhibits correlation performance results for P.563 and the proposed GMM-based algorithm, for with and without regression mapping. Without regression mapping, the GMM method achieves correlation R comparable to P.563 on SMV experiments 1 and 2. An increase in R of approximately 45% can be seen on experiment 3. With regression mapping, the GMM method outperforms P.563 by approximately 5% and 22% on experiments 1 and 3, respectively. The results suggest that the proposed method may be capable of predicting MOSs for speech under noisy conditions more effectively than P.563.

Processing time is also an important figure of merit for gauging algorithm performance. We use the ANSI-C reference implementation of P.563. The computation time for the GMM-based method encompasses the time for PLP calculation, VAD and voicing decision, separation of voiced, unvoiced and inactive frames, calculation of likelihoods and MARS mapping. With the exception of the VAD algorithm (taken from the ANSI-C reference implementation of ITU-T G.729B [14]), the rest of the GMM-based algorithm is implemented using Matlab version 6.5 Release 13. 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 1. Results are given for a randomly selected speech file 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. The proposed method is capable of reducing the processing time to less than half of that of P.563. The reduction numbers are very conservative. A complete C implementation of the proposed algorithm would surely increase the percentage reduction.

4. CONCLUSION AND FURTHER INVESTIGATION

A novel non-intrusive speech quality estimation algorithm is proposed based on Gaussian mixture models. We have modeled the behavior of clean speech using GMMs and compared features extracted from degraded speech signals to the artificial reference model. The degree of consistency with the model served as an indicator of speech quality. Simulations with PLP coefficients have shown that our approach outperforms P.563 by up to 44.74% increase in R, for SMV coded speech under noisy conditions; the proposed algorithm is comparable to P.563 under various other conditions. Furthermore, an average 40% reduction in processing time was obtained compared to P.563.

Currently, our approach uses only GMMs to model the behavior of clean speech. It can be inferred that if the behavior of degraded speech is also modeled and consistency measures with respect to both clean and degraded models are computed, a more powerful and robust algorithm can be constructed. Other avenues such as incorporating different features are also being pursued. The accuracy and computational simplicity of the proposed method will be further improved.

5. REFERENCES

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

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

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

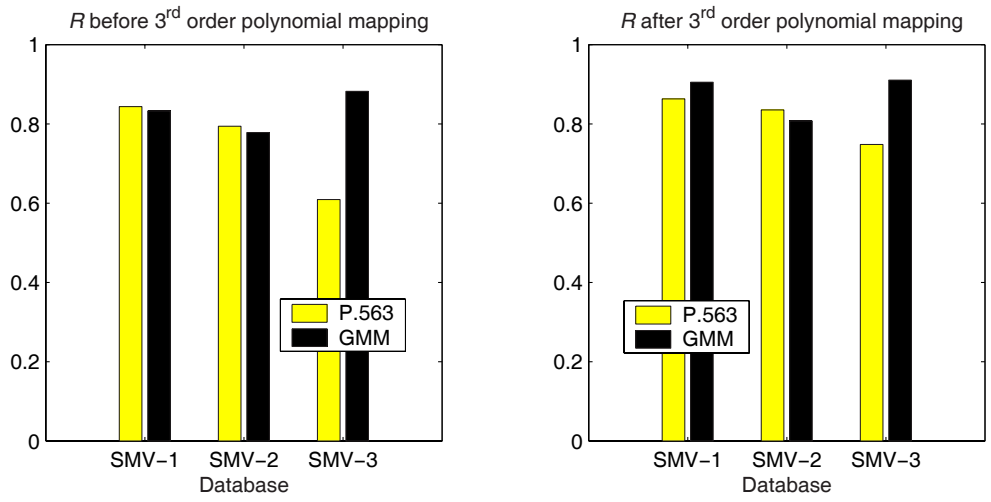

Fig. 2. Comparison of the proposed GMM-based algorithm with P.563, for with (right) and without (left) 3^{rd} order polynomial regression mapping.

Table 1. Algorithm processing time

			GMM - Diagonal (% reduction)			GMM - Full (% reduction)	
Database	File length (sec.)	P.563 (sec.)	$M = 8$	$M = 16$	$M = 32$	$M = 8$	$M = 16$
SMV - Expt. 1	7.20	4.37	52.5	36.8	5.6	52.5	36.8
SMV - Expt. 2	5.89	4.10	58.1	44.7	17.2	58.0	44.6
SMV - Expt. 3	7.91	5.50	58.5	44.6	17.4	58.3	44.5

[4] D.-S. Kim and A. Tarraf, "Perceptual model for non-intrusive speech quality assessment," in *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 3, May 2004, pp. 1060–1063.

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

[6] Psytechnics Limited, "NiQA - Product Description," Tech. Rep., January 2003. [Online]. Available: http://www.psytechnics.com/pages/products/niqa.php

[7] SwissQual Inc., "NiNA - SwissQual's non-intrusive algorithm for estimating the subjective quality of live speech," Tech. Rep., June 2001. [Online]. Available: http://www.swissqual.com/HTML/ninapage.htm

[8] H. Hermansky, "Perceptual linear prediction (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[9] W. Zha and W.-Y. Chan, "A data mining approach to objective speech quality measurement," in *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, May 2004, pp. 461–464.

[10] T. H. Falk, W.-Y. Chan, and P. Kabal, "Speech quality estimation using Gaussian mixture models," in *Proc. of the Int. Conf. on Spoken Language Processing*, Oct. 2004, pp. 2013–2016.

[11] 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.

[12] J. H. Friedman, "Multivariate adaptive regression splines," *The Annals of Statistics*, vol. 19, no. 1, pp. 1–141, March 1991.

[13] ITU-T Rec. P. Supplement 23, "ITU-T coded-speech database," International Telecommunication Union, Geneva, Switzerland, Feb. 1998.

[14] ITU-T Rec. G.729 - Annex B, "A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70," International Telecommunication Union, Geneva, Switzerland, Nov. 1996.