ENHANCED NON-INTRUSIVE SPEECH QUALITY MEASUREMENT USING DEGRADATION MODELS

Tiago H. Falk and Wai-Yip Chan

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

ABSTRACT

The speech quality estimation scheme in [1] is improved with the addition of a reference model of the behavior of speech degraded by different transmission and/or coding schemes. Moreover, via maximization of a mutual information measure, we validate the use of segmental SNR as a measure of the amount of multiplicative noise present in the test signal. These two additions result in an algorithm that is more accurate and more robust to certain distortion conditions. When tested on unseen data, the proposed algorithm outperforms the current "state-of-art" P.563 algorithm while requiring considerably lower computational complexity.

1. INTRODUCTION

Speech quality has long been recognized as a pivotal factor in voice telecommunication but speech quality measurement has remained labor intensive. In the most common test [2], 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. Non-intrusive objective algorithms require only the degraded (processed) signal as input, while intrusive algorithms input both the clean (unprocessed) and degraded signals. Research into non-intrusive quality measurement usually entails comparisons between the test signal and normative behavior of clean speech, amongst other distortion-sensitive features. In [1], features of the received speech signal are compared to Gaussian-mixture probability models (GMMs) that serve as artificial reference models of clean speech behavior. In [3], degradations are detected when calculations of vocal-tract parameters yield implausible results relative to normal speech. Modulation-spectral features derived from the temporal envelope of speech are used in [4] to detect improbable movement speeds of the human articulatory system. The International Telecommunications Union ITU-T P.563 standard represents the "state-of-art" algorithm [5]. P.563 tests for inconsistencies in vocal tract analysis, for high levels of noise, and for speech interruptions, mutes and time clippings.

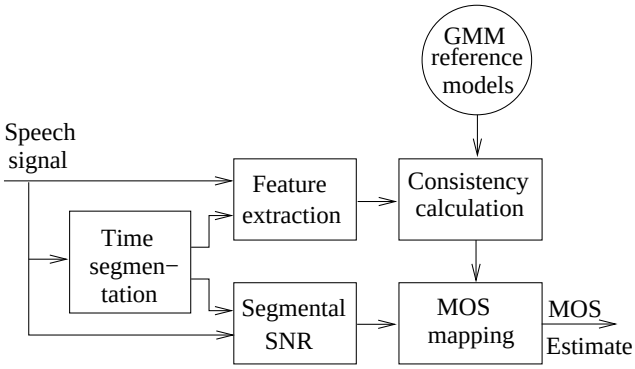

Fig. 1. Architecture of the proposed algorithm.

This paper proposes a more accurate and robust GMM-based speech quality measurement algorithm. Robustness and increase in accuracy are 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. Simulation shows that our approach offers accurate and yet low-complexity measurement of speech quality.

2. ALGORITHM DESCRIPTION

The overall architecture of the proposed algorithm is depicted in Fig. 1. Perceptual features are first extracted from the test speech signal every 10 milliseconds. The time segmentation module labels the feature vector of each frame as belonging to one of three possible classes: voiced, unvoiced, or inactive. Offline, two reference models are created. One uses high-quality, undistorted speech signals to produce a reference model of the behavior of clean speech

features. A second uses speech signals corrupted by different coding and/or transmission distortions to produce a reference model of degraded speech behavior. In both instances, this is accomplished by modeling the probability distribution of the features for each of the three classes with a GMM. Features extracted from the test signal are assessed using the reference models by calculating "consistency" measures with respect to each of the six GMMs. The segmental SNR block is added to compensate for files with high multiplicative noise. The calculated values serve as speech quality indicators and are mapped to an estimated MOS value.

2.1. Feature Extraction and Time Segmentation

As in [1], 5^{th} order perceptual linear prediction (PLP) cepstral coefficients [6] are used as speaker independent perceptual features. Here, as coding and/or transmission degradations may affect the energy of the signal, the zeroth cepstral coefficient is kept as an energy measure. Time segmentation is employed to separate the speech frames into different classes and is performed using a voice activity detector (VAD) and a voicing detector. The VAD from ITU-T G.729B [7] is used here.

2.2. GM Reference Models and Consistency Calculation

Gaussian mixture models are used to model the PLP cepstral coefficients of each class of speech frames. A Gaussian mixture density is a weighted sum of M component densities $p(\mathbf{x}|\boldsymbol{\lambda}) = \sum_{i=1}^{M} \alpha_i b_i(\mathbf{x})$, where $\alpha_i \geq 0, i = 1, ..., M$ are the mixture weights, with $\sum_{i=1}^{M} \alpha_i = 1$, and $b_i(\mathbf{x})$, $i = 1, ..., M$, 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 $\boldsymbol{\lambda}_i = \{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \alpha_i\}$. Using clean speech signals, three different Gaussian mixture densities, $p_{clean,class}(\mathbf{x}|\boldsymbol{\lambda})$ are trained. The subscript "class" represents either voiced, unvoiced, or inactive frames. For the degradation model, $p_{degraded,class}(\mathbf{x}|\boldsymbol{\lambda})$ are found.

For the benefit of low computational complexity, we make a simplifying assumption that vectors between frames are independent. Improved performance may be achieved with more sophisticated models, such as Markov models, where statistical dependencies between frames can be considered. Nonetheless, the simplifying assumption has been shown in [8] to provide accurate speech quality estimates. Thus, for a given speech signal, the consistency between the observation and the models can be defined as

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

where $\mathbf{x} = \mathbf{x}_1, \ldots, \mathbf{x}_{N_{class}}$ are the PLP coefficient vectors, N_{class} is the number of such vectors in the frame $class$, and the subscript "model" represents either the clean or the degradation reference model. In total, six consistency measures are calculated.

2.3. Segmental SNR and MOS Mapping

In preliminary experiments without the "segmental SNR" block (as depicted in Fig. 1), large estimation errors were found on files degraded by modulated noise reference unit (MNRU) [9]. Fig. 2 (a) depicts absolute estimation errors for a test database. Note that the largest errors occur at low-SNR MNRU degradations (conditions 46-48). It is known that MNRU introduces multiplicative noise in a manner similar to logarithmically companded PCM and ADPCM systems [9]. In P.563, several features are used to describe multiplicative noise.

Our ultimate goal is to achieve accurate low complexity speech quality measurement. To use a plethora of features to estimate multiplicative noise is unfeasible. To this end, normalized mutual information (NMI) is used to detect the single best feature extracted by P.563. Here, NMI is the mutual information between a feature (X) and subjective MOS (Y), normalized by the entropy of X or Y, whichever is smaller. The feature with the highest NMI is ranked the most important. For MNRU-degraded files, estimated segmental SNR (SegSNR) is detected as being the most important feature. The calculations performed in the "segmental SNR" block are similar to those described in [5]. Fig. 2 (b) depicts estimation errors after SegSNR is used. Note a decrease in absolute error for MNRU conditions 46-48. It is important to emphasize that other conditions are not affected (negatively) with the addition of the "segmental SNR" block.

Lastly, the "MOS mapping" block shown in Fig. 1 is responsible for mapping the six consistency measures and SegSNR to an objective MOS. We experiment with multivariate polynomial regression and multivariate adaptive regression splines (MARS) [10]. Simulation results showed that MARS provides better performance; the results below are all based on using MARS.

3. ALGORITHM DESIGN CONSIDERATIONS

Preliminary "calibration" experiments were carried out in order to find an effective combination of GMM configuration parameters: M and covariance matrix type. For voiced and unvoiced PLP frames we experiment with diagonal matrices and M=8, 16, or 32, and M=2, 3, 5 for full covariance matrices. For inactive PLP frames, we only experiment with diagonal matrices and M=2, 3, or 6. In this calibration experiment, a total of ten 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 [11], two wireless (IS-96A and IS-127 EVRC), and a mixed

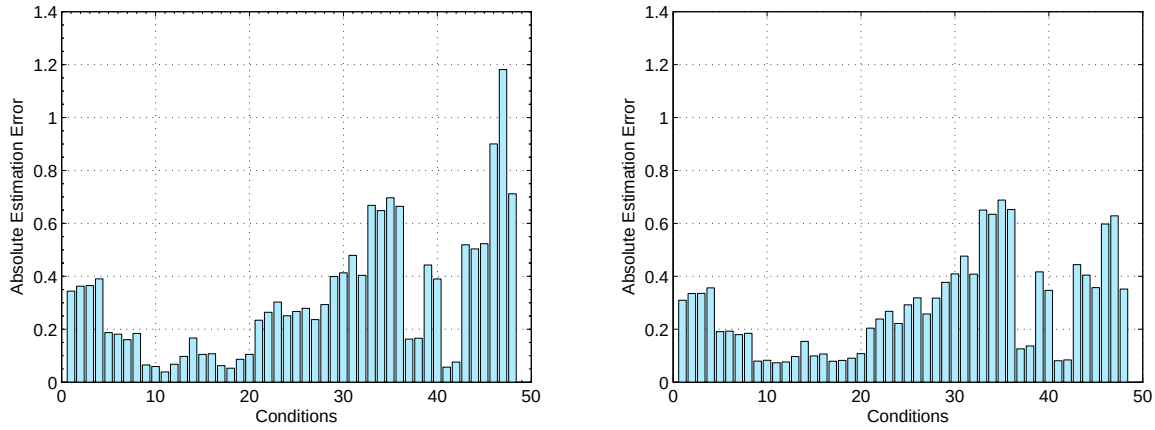

Fig. 2. "Per-condition" absolute MOS estimation error (a) without SegSNR and (b) with SegSNR.

database [12]. The combined ten databases contain 1760 degraded files and 1136 clean files. The degraded files include speech coded using various reference codecs (e.g., G.726, G.728, G.729, IS-54, and GSM-FR), single or in tandem, corrupted by various channel errors and acoustic noise conditions, and speech degraded by various levels of MNRU. 90% of the files are used to train the GM models and 10% is left for testing. The clean and degradation GM models are evaluated at the feature vectors of the test signal ($\mathbf{x}$) for each speech class, i.e., $c_{clean,class}(\mathbf{x})$ and $c_{degraded,class}(\mathbf{x})$ are calculated, respectively. Experiments suggest the use of 3 full GMM components for voiced frames, and 32 and 6 diagonal components for unvoiced and inactive frames, respectively. These parameters strike a balance between algorithm complexity and performance.

4. EXPERIMENT RESULTS

The aforementioned databases and combinations of M and matrix types are used to design the clean and degradation reference models. A separate set of seven MOS-labeled databases is used to train the MARS mapping. Three databases comprise speech coded with the 3GPP2 Selectable Mode Vocoder (SMV). Experiment 1 encompasses tandeming and nominal input level conditions, experiment 2 covers channel impairments, and experiment 3 noisy environment conditions. Three multilingual databases are comprised of speech coded using the G.711, G.726 and the G.728 speech coders, alone and in various different tandeming schemes. The seventh database includes speech coded with various different speech coders (e.g., G.723.1, G.729, G.729E, and GSM-EFR), under various channel degradation conditions. In all seven databases, speech degraded by different levels of MNRU are also included.

To test the algorithm on "unseen" data, each database is separated into two disjoint sets; one is kept for training, the other for testing. Each SMV and the mixed database are comprised of speech files from four female and four male speakers. For each database, files from one female and one male speaker are kept for training and the remaining are used for testing. The multilingual databases are each comprised of files from two male speakers and two female speakers and are used entirely for training.

Table 1 presents "per-condition" correlation (R) and root-mean-square error (ϵ) between subjective MOS and objective MOS, for each of the datasets. The results are obtained after 3^{rd} order monotonic polynomial regression, as recommended in [5]. 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

$$\% \uparrow = \frac{R_{GMM} - R_{P.563}}{1 - R_{P.563}} \times 100\% \qquad (2)$$

and indicates percentage reduction of P.563's performance gap to perfect correlation. The column labelled "$\%\downarrow$" lists percentage reduction in ϵ, relative to P.563, by using the proposed scheme.

The proposed algorithm outperforms P.563 on all four test sets. An average improvement in R of 45% and an average decrease in ϵ of 28% is achieved. We emphasize the algorithm's enhanced performance by comparing results for the SMV-2 database. In [1], P.563 outperformed the GMM-based algorithm by 13% in R and 6% in ϵ. Degradation conditions in SMV-2 encompass frame errors with 1, 3, or 5% frame error rates. By introducing the degradation reference model, the algorithm becomes more robust to these distortion conditions. The result is an improvement of ap-

Table 1. Performance of P.563 and the proposed algorithm on "unseen" datasets

Unseen Dataset	P.563		Proposed			
	R	ϵ	R	$\%\uparrow$	ϵ	$\%\downarrow$
SMV-1	0.846	0.267	0.937	59.1	0.175	34.5
SMV-2	0.848	0.265	0.861	8.6	0.254	4.2
SMV-3	0.795	0.251	0.959	80.0	0.116	53.8
Mixed	0.912	0.242	0.942	34.1	0.199	17.8
Average	–	–	–	**45.4**	–	**27.5**

proximately 22% in R and 10% in ϵ relative to [1]. It is also worth mentioning that the performance gains attained on SMV-3 suggest that the proposed method may be more effective than P.563 for speech under noisy environment conditions.

A final test is performed on a *de facto* unseen database with speech files coded using newer codecs (e.g., a cable VoIP speech coder) than the codecs represented in the GMM and MARS training datasets. Evaluation using this database demonstrates the applicability of the proposed algorithm to emerging codec technologies. On this database, P.563 achieves $R = 0.916$ and $\epsilon = 0.218$. The proposed algorithm achieves $R = 0.924$ and $\epsilon = 0.207$, a 10% improvement in correlation and a 5% decrease in root-mean-square error.

4.1. A Note on Complexity

The computational complexity of the proposed algorithm is mainly attributable to the time segmentation module. While training the GMMs and the MARS mapping is somewhat involved, this is performed offline and does not pose a serious burden. The proposed algorithm depends only on seven features while P.563 depends on over 50 parameters. Since all of the processing modules in the proposed algorithm have similarly complex blocks present in P.563, it is estimated that the proposed algorithm uses roughly 15% of the load required by P.563. More precise comparison figures are currently being investigated.

5. CONCLUSION

An enhanced non-intrusive speech quality estimation algorithm is proposed. It has been shown that, by adding a degradation reference model, measurement accuracy is enhanced and the GMM-based algorithm becomes more robust to certain distortion conditions. Further improvement is attained when segmental SNR, used as a measure of the amount of multiplicative noise present in the signal, is considered. The algorithm provides competitive quality estimates relative to the current "state-of-art" algorithm, whilst requiring considerably lower computational complexity.

6. REFERENCES

[1] 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.

[2] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality," International Telecommunication Union, Geneva, Switzerland, Aug. 1996.

[3] 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.

[4] D.-S. Kim, "ANIQUE: An auditory model for single-ended speech quality estimation," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, pp. 821–831, Sept. 2005.

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

[7] 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.

[8] T. H. Falk and W.-Y. Chan, "Non-intrusive speech quality estimation using Gaussian mixture models," to appear in *IEEE Signal Processing Letters*, Feb. 2006.

[9] ITU-T Rec. P.810, "Modulated noise reference unit - MNRU," International Telecommunication Union, Geneva, Switzerland, Feb. 1996.

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

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

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