A COMPUTATIONAL MODEL FOR MOS PREDICTION

Doh-Suk Kim, Oded Ghitza, and Peter Kroon

Acoustics and Speech Research Department
Bell Laboratories, Lucent Technologies
Murray Hill, New Jersey 07974, USA
ds@sait.samsung.co.kr

ABSTRACT

A computational model to predict MOS of processed speech is proposed. The system measures the distortion of processed speech (compared to the source speech) using a peripheral model of the mammalian auditory system and a psychophysically-inspired measure, and maps the distortion value onto the MOS scale. This paper describes our attempt to derive a "universal", database-independent, distortion-to-MOS mapping function. Preliminary experimental evaluation shows that the performance of the proposed system is comparable with ITU-T recommendation P.861 for clean speech sources, and outperforms the P.861 recommendation for speech sources corrupted by either car or babble noise at 30 dB SNR.

1. INTRODUCTION

Up until this day, the most reliable way to evaluate the performance of a speech coding system is to perform subjective speech quality assessment tests such as MOS (Mean Opinion Score) test. Obviously, these tests are expensive both in time and cost, and difficult to reproduce. Thus, it is desirable to replace them with an objective method.

Numerous studies have been conducted with the purpose of finding a distortion measure that will correlate well with subjective MOS measurements, including the PSQM method [1], which was adopted as the ITU-T standard recommendation for telephone band speech (P.861) [2]. To the best of our knowledge, none of these studies yet resolved two major challenges: (1) how to map the distortion value onto the MOS scale, and (2) how to accurately assess the quality of processed speech, where the source had been corrupted by environmental noise.

In this paper, we address these challenges. We consider a system comprising of two stages. The first (termed ASQM, for Auditory Speech Quality Measure) measures the distortion of a processed speech (compared to the source speech) using a peripheral model of the mammalian auditory system and a psychophysically-inspired measure. It will be shown that the robustness of auditory-based representations to environmental noise (as was demonstrated elsewhere, e.g., [3], [4]), results in a distortion measurement that correlates well with subjective quality assessments of speech.

The second stage maps the distortion value onto the MOS scale. Previous studies have been confined to the first stage, and performance was evaluated via correlation analysis of the resulting (objective) distortion measurement with the subjective MOS. One could use the resulting regression line as a distortion-to-MOS mapping function. Such mapping function, however, is database-dependent and one cannot ensure that it will generalize to new speech databases.[1] In this paper we describe our attempt to derive a "universal", database-independent, distortion-to-MOS mapping function. Preliminary experimental evaluation of the proposed framework is reported.

2. OBJECTIVE DISTORTION MEASUREMENT

First, the overall active speech level of the source speech $x(n)$ and the coded $y(n)$ is normalized to -26 dBov using the speech level meter from the ITU software library [5]. Next, the time waveforms of the source and the processed speech are aligned. The level-adjusted and time-aligned signal is then transformed into a sequence of feature vectors using the auditory model. Here, we used the ZCPA (for Zero-Crossings with Peak Amplitudes) auditory model [4]. Finally, the two vector sequences are compared to produce a distortion value which is related to speech quality. Let $X(m,i)$ and $Y(m,i)$ be the auditory representations of source and processed speech at the m-th frame, respectively. Here, $1 \le i \le N_b$ denotes the frequency bin index and N_b is the dimension of frame vector. Then the distortion at the m-th frame is expressed as

$$D(m) = \sum_{i=1}^{N_b} C(m,i)\,|X(m,i) - Y(m,i)|\,. \qquad (1)$$

Here, $C(m,i)$ is an asymmetric weighting factor to account for the psychoacoustic observation, first introduced in the PSQM [1], that additive distortion in the time-frequency domain are subjectively more noticeable than equal amounts of subtractive distortion. $C(m,i)$ is defined as

$$C(m,i) = \left(\frac{Y(m,i)+\epsilon}{X(m,i)+\epsilon}\right)^{\alpha}, \qquad (2)$$

where ϵ is a small number to prevent division by zero and α is a control parameter greater than zero. Although the basic form of the asymmetric measure is adopted from the PSQM, parameters should be optimized for the auditory representations.

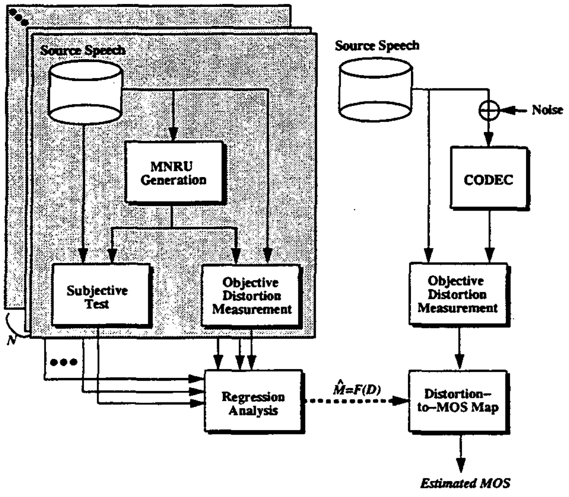
The overall distortion between two sequences X and Y is determined by

$$D \;=\; \gamma D_{sp} + (1-\gamma) D_{nsp} \qquad (3)$$

Doh-Suk Kim is currently with Human and Computer Interaction Laboratory, Samsung Advanced Institute of Technology, P.O. Box 111, Suwon 440-600, Korea.

[1] Here, the term "database" refers to both the speech material and the speech processing systems under evaluation.

0-7803-5651-9/99/$10.00 ©1999 IEEE.



Phase I : Obtaining Distortion–to–MOS Map *Phase II : Evaluation*

Figure 1: Block diagram of the database-independent MOS estimation.

where γ is a weighting factor for active speech frames, and D_{sp} and D_{nsp} are the distortions for the speech portion and the non-speech portions of the signal, respectively. Distortions for the speech and the non-speech are defined as

$$\begin{aligned} D_{sp} &= \frac{1}{\max_m L_Y(m)\cdot T_{sp}} \sum_{\substack{m \\ s.t.\ L_X(m)>K}} D(m) \\ D_{nsp} &= \frac{1}{\max_m L_Y(m)\cdot T_{nsp}} \sum_{\substack{m \\ s.t.\ L_X(m)\leq K}} D(m) \end{aligned} \tag{4}$$

where $L_X(m)$ and $L_Y(m)$ are the the pseudo-loudness of the source speech and the processed speech at the m-th frame, respectively, K is the threshold for speech / non-speech decision, and T_{sp} and T_{nsp} are the number of active speech frames and the number of non-speech frames, respectively. For clean speech, only the active speech frames contribute to the overall distortion metric unless the speech coding algorithm under test generates high-power distortions in the non-speech frames.

3. MOS ESTIMATION

The proposed system for MOS estimation is shown in Fig. 1. It is based on a basic assumption that the subjective MOS scores of MNRU-conditioned [6] speech sentences are consistent across different databases. Consequently, we collected every MNRU condition (speech material and their associated subjective MOS scores) from all databases in our possession. A regression analysis on these MNRU conditions (as a function of SNR) provided the desired distortion-to-MOS mapping function. (Of course, we will update the mapping function once a new MNRU database will be provided). This "Training phase" of the system is shown in the left hand side of Fig. 1. The right hand side describes the proposed usage of the distortion-to-MOS map, which is independent of the speech source material or the nature of the evaluated codec. Note that the objective distortion measurement of the processed speech is with respect to the clean source speech. This is also the case when "processed speech" is the noisy, unprocessed, speech source.

4. EXPERIMENTAL RESULTS

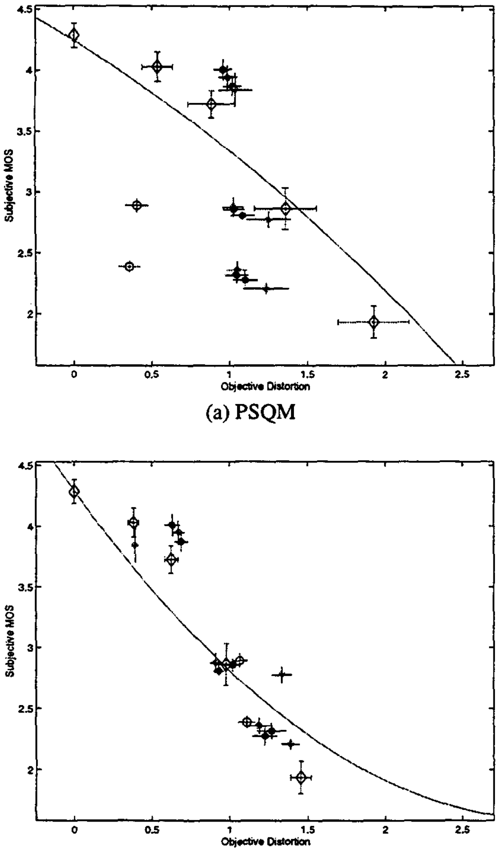
A preliminary experimental evaluation was performed using three databases (termed DB-I, DB-II and DB-III). Databases DB-I and DB-II contain only clean speech material, comprising of 32 speech sentences, spoken by 4 male and 4 female speakers, and 11 different coders, ranging in bit-rate form 32 kb/s to 8 kb/s. Database DB-III contains clean speech as well as noisy speech material (with only four coders). Two kinds of background noise were used, car noise and speech-babble noise (both at 30 dB SNR). The distortion-to-MOS mapping function was obtained from a pool of MNRU conditions collected from all three databases.

The performance of the PSQM and ASQM based systems for all three databases is summarized in Table 1. The first column shows the correlation coefficient, ρ, between the objective distortion measure and its corresponding subjective MOS. The second column shows the root-mean-squared error (RMSE) with respect to the distortion-to-MOS mapping function. The PSQM and the ASQM based systems perform comparably in clean. However, ASQM outperforms PSQM in noisy conditions. In particular, the RMSE of ASQM is significantly smaller than that of PSQM for noisy speech, by 0.195 for car noise and by 0.511 for and babble noise, which demonstrates the robustness of the peripheral auditory model.

This point is also illustrated in Fig. 2, which shows the relationship between the subjective MOS and the objective distortion measurements for the database DB-III, and for the PSQM and ASQM based systems. In each plot, diamonds represent clean source and MNRU conditions, circles represent codecs, and unfilled circles represent the unprocessed noisy source. The value of each point, and the size of the corresponding error-bars, are the mean and the standard deviation computed over all speech sentences. The distortion-to-MOS mapping function, obtained from the MNRU conditions of databases DB-I, DB-II and DB-III, is superimposed on each plot (dashed curve). The data points for all three environmental conditions are plotted. The group of points with MOS values of about 3.8 is for clean speech, those with MOS of 2.8 are for speech with car noise, and those for MOS of 2.3 are for speech with babble noise.

Table 1: Comparison of PSQM and ASQM in terms of correlation coefficient ρ, and RMSE with respect to the distortion-to-MOS regression.

	ρ		RMSE	
	PSQM	ASQM	PSQM	ASQM
DB-I	0.843	0.841	0.344	0.350
DB-II	0.915	0.864	0.263	0.339
DB-III (CLN)	0.969	0.968	0.424	0.469
DB-III (C30)	0.838	0.953	0.445	0.250
DB-III (B30)	0.787	0.986	0.789	0.278



(a) PSQM

(b) ASQM

Figure 2: Relationship between subjective MOS and objective distortion measurement, for (a) PSQM or (b) ASQM based system, for database DB-III. Diamonds represent clean source and the MNRU conditions, circles represent codecs, and unfilled circles represent the unprocessed noisy source. The distortion-to-MOS mapping, obtained from MNRU conditions of databases DB-I, DB-II and DB-III, is superimposed on each plot (dashed line).

5. CONCLUSIONS

This paper described a methodology to predict subjective MOS scores. The method consists of a two stage process. First a distortion measure based on an auditory model, followed by a database independent distortion-to-MOS mapping using MNRU anchors.

Based on evaluation on a small number of databases, the method provides MOS estimates that are highly correlated with MOS scores obtained by real listening tests. The method seems to be robust against environmental noise.

It should be noted that these are preliminary results. Further evaluation is needed, using different databases, to confirm the underlying assumption that distortion-to-MOS mapping based upon MNRU anchor points can be used to map distortion measurements of coded speech. It may very well be that other anchor points may be needed, such as carefully selected, standardized coders. Further evaluation is also needed to confirm the robust performance against noise, using other noise sources for a wide range of SNR values.

6. ACKNOWLEDGMENT

The authors would like to thank M. M. Sondhi and J. H. Hall for stimulating discussions throughout this work, and C. A. Harsin and D. A. Quinlan for providing the noisy speech database.

7. REFERENCES

[1] J. G. Beerends and J. A. Stemerdink, "A perceptual speech-quality measure based on psychoacoustic sound representation," *J. Audio Eng. Soc.*, vol. 42, pp. 115–123, March 1994.

[2] ITU-T Recommendation P.861, *Objective Quality Measurement of Telephone-Band (300-3400 Hz) Speech Codecs.* Geneva, 1996.

[3] O. Ghitza, "Auditory nerve representation as a basis for speech processing," in *Advances in Speech Signal Processing* (S. Furui and M. M. Sondhi, eds.), pp. 453–485, New York: Marcel Dekker, 1992.

[4] D. S. Kim, S. Y. Lee, and R. M. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 1, pp. 55–69, 1999.

[5] ITU-T STL96, *ITU-T Software tool library.* Geneva, May 1996.

[6] ITU-T Recommendation P.810, *Modulated Noise Reference Unit (MNRU)*, Feb. 1996.