

COMPARISON OF TWO OBJECTIVE SPEECH QUALITY MEASURES: MBSD AND ITU-T RECOMMENDATION P.861

Wonho Yang and Robert Yantorno

Speech Processing Lab

Electrical & Computer Engineering, Temple University, Philadelphia, PA 19122-6077

wonho@astro.temple.edu, ryantorn@nimbus.temple.edu

<http://nimbus.temple.edu/~ryantorn/speech/>

ABSTRACT

The Modified Bark Spectral Distortion (MBSD), used for an objective speech quality measure, was presented previously [1, 2]. The MBSD measure estimates speech distortion in loudness domain taking into account the noise masking threshold in order to include only audible distortions in the calculation of the distortion measure. Preliminary simulation results have shown improvement of the MBSD over the conventional BSD. In this paper, the performance of the MBSD is improved by scaling the noise masking threshold. Finally, this new MBSD (MBSD II) measure is compared with the ITU-T Recommendation P.861 objective speech quality measure for telephone-band speech [3].

1. INTRODUCTION

Development of an objective speech quality measure that correlates well with subjective speech quality measures has been considered important because subjective tests are expensive and time-consuming. Since objective measures are easy to implement and less time-consuming, a good objective speech quality measure would be a valuable assessment tool for speech coder development, speech codec deployment on communication systems, and even for speech codec selection. In reality, various types of objective speech quality measures have been used to improve speech quality in Analysis-By-Synthesis (ABS) speech coders [4].

Among the various different objective speech quality measures, we have been interested in perceptual distortion measures such as the Bark Spectral Distortion (BSD) [5] and the Perceptual Speech Quality Measure (PSQM) [6]. PSQM has been recommended as an objective quality measurement of telephone-band speech codecs by the ITU [3]. Since the development of the BSD, it has become a good candidate for a highly correlated objective quality measure, according to several researchers [7, 8, and 9]. The BSD measure is based on the assumption that speech quality is directly related to speech loudness, which is a psychoacoustical term, defined as the magnitude of auditory sensation. The BSD measure is the average squared Euclidean distance of estimated loudness of the original and the coded utterances. In order to calculate loudness, the speech signal is processed using results of

psychoacoustic measurements, which include critical band analysis, equal-loudness preemphasis and intensity-loudness power law [5].

Even though the conventional BSD measure showed a relatively high correlation with Mean Opinion Score (MOS) – the most popular subjective speech quality measure, there are areas for possible improvement. Motivated by the transform coding of audio signals, which uses the noise masking threshold [10], the MBSD measure has incorporated this concept of a noise masking threshold into the conventional BSD measure, where any distortion below the noise masking threshold is not included in the BSD measure. This new addition of the noise masking threshold replaces the empirically derived distortion threshold value used in the conventional BSD [5]. The concept of a noise masking threshold was also used to improve speech quality in coder development [11]. It was shown that coding gain could be obtained with no loss of speech quality, by transmitting only spectral samples above the noise masking threshold. This implies that the noise below the noise masking threshold is not perceptible. Therefore, the noise spectral components below the noise masking threshold are excluded in the calculation of the MBSD measure because these components are considered inaudible.

In this paper, we describe the MBSD measure and show the effect of the noise masking threshold. The performance of the MBSD is improved by scaling the noise masking threshold and comparing this new MBSD measure with that of the ITU-T Recommendation P.861.

2. MBSD MEASURE

The block diagram of the MBSD measure is shown in Fig. 1. There are three major processing steps: loudness calculation, noise masking threshold computation, and computation of MBSD. The loudness calculation transforms speech signal into loudness domain. In order to transform speech into the loudness domain, the speech signal is processed in several steps: critical band analysis, equal-loudness preemphasis and intensity-loudness power law. The procedure of loudness transformation is the same as that of the BSD [5]. However, there are two differences

between the conventional BSD and the MBSD. First, the MBSD uses the noise masking threshold for the determination of audible distortion, while the BSD uses an empirically determined power threshold. Second, the computation of distortion of the BSD is different from that of the MBSD. The BSD defines the distortion as the average squared Euclidean distance of estimated loudness, while the MBSD defines the distortion as the average difference of estimated loudnesses. The determination of a perceptual distortion metric in the loudness domain was not investigated for the BSD. The importance of defining an appropriate perceptual distortion metric was discussed in [12]. An initial attempt to search for a proper metric in the MBSD is addressed in [2]. The most appropriate metric was determined to be the average difference of two loudnesses.

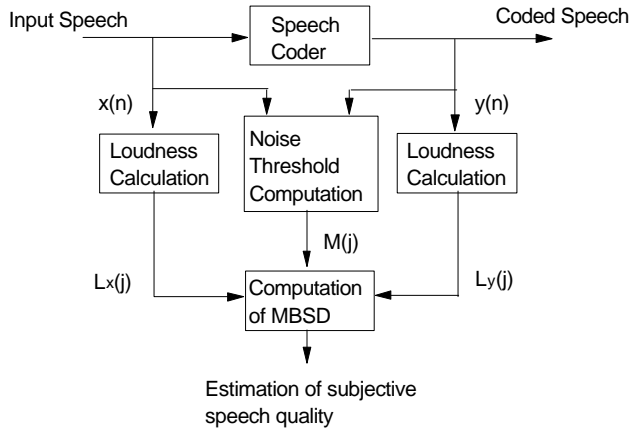


Figure 1. Block diagram of MBSD method.

The noise masking threshold is estimated by critical band analysis, spreading function application and absolute threshold consideration [10]. This noise masking threshold estimation considers tone-masking noise and noise-masking tone. The loudness of the noise masking threshold is compared to the loudness difference of the original and the coded speech to determine if the distortion is perceptible. When the loudness difference is below the loudness of the noise masking threshold, this loudness difference is imperceptible. Therefore, it is not included in the calculation of the MBSD.

In order to formally define the distortion for the MBSD, an indicator of perceptible distortion $M(i)$ is introduced, where i is the i -th critical band. When the distortion is perceptible, $M(i)$ is 1, otherwise $M(i)$ is 0. The indicator of perceptible distortion is obtained by comparing the loudness to the noise masking threshold. The calculation of the MBSD is given by equation (1). Imperceptible distortion is excluded in the MBSD calculation when $M(i)$ is zero. The MBSD is then defined as the average difference of estimated loudness which is perceptible.

$$MBSD = \frac{1}{N} \sum_{j=1}^N \left[\sum_{i=1}^K M(i) \left| L_x^{(j)}(i) - L_y^{(j)}(i) \right| \right] \quad (1)$$

where,

N : number of frames processed

K : number of critical bands

$M(i)$: Indicator of perceptible distortion at i -th critical band

$L_x^{(j)}(i)$: Bark spectrum of j -th frame of original speech

$L_y^{(j)}(i)$: Bark spectrum of j -th frame of coded speech

3. RESULTS AND DISCUSSION

In order to examine the performance of the MBSD, we performed several different types of experiments. Some initial results regarding the performance of the MBSD have been previously reported [2]. In this paper, we show that the noise masking threshold plays an important role in estimating perceptual speech quality. The performance of this new MBSD (MBSD II) is compared with that of the ITU-T Recommendation P.861.

For the experiments, we computed the MBSD measures frame by frame, with the frame length of 320 samples overlapping by a half frame. Each frame was weighted by a Hanning window. We processed only non-silence frames. We used a speech data set which included 5 MNRU conditions and various different types of speech coders such as ADPCM, GSM, IS54, FS1016, LD-CELP and CELP. An objective quality measure is a comparison measure of two speech utterances whereas the MOS is an absolute measure. For the experiments outlined here, the MOS difference between the original speech and the coded speech is used for the evaluation of objective speech quality measures with a second-order regression analysis. In our experiment, 64Kbps PCM was regarded as original speech.

3.1. Effect of Noise Masking Threshold

Since the MBSD uses the noise masking threshold, which determines if the distortion is perceptible, it is worthwhile to study the effect of the noise masking threshold on the performance of the MBSD. In order to examine the effect of the noise masking threshold, we compared the performance of the MBSD without the noise masking threshold and with the noise masking threshold. The estimated distortion for the MBSD without noise masking threshold has been computed by setting $M(i)$, the indicator of perceptible distortion, to 1. Figure 2 shows the performance of the MBSD without the noise masking threshold. According to Figure 2, the MBSD without the noise masking threshold overestimates some distortions because it simply calculates the loudness difference without considering perceptual distortion. Figure 3 shows the performance of the MBSD with the noise masking threshold using the same speech data set. It clearly shows that the overestimated distortion has been decreased, and the MBSD with the noise masking threshold gives a higher correlation (0.9418 versus 0.8630 for MBSD without the noise masking threshold) with subjective quality measure. Therefore, the noise masking threshold plays an important role in estimating perceptually relevant distortion of objective speech quality measure.

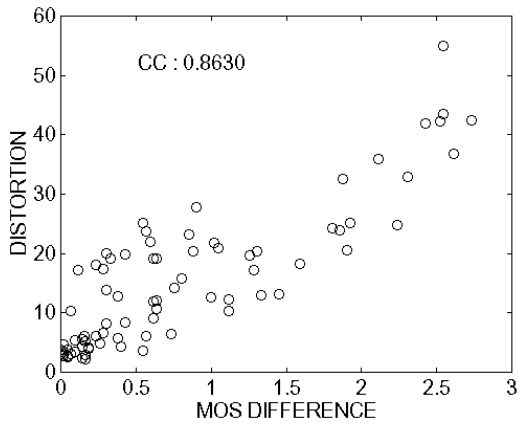


Figure 2. Plot of MBSD without noise masking threshold versus MOS difference for various coders.

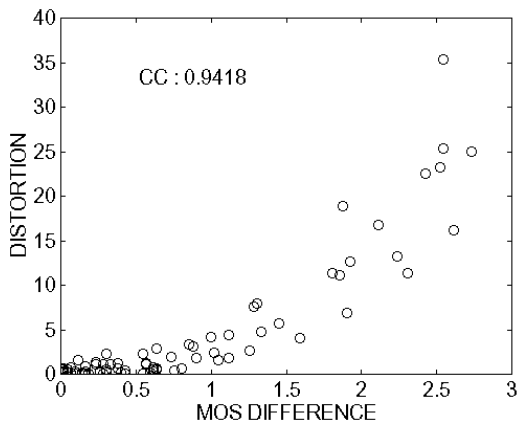


Figure 3. Plot of MBSD with noise masking threshold versus MOS difference for various coders.

3.2. Comparison of MBSD with P.861

Correlation coefficients with MOS scores have been the traditional evaluation tool for the performance of objective speech quality measures. It has been suggested that it is more appropriate to use correlation coefficients with MOS difference rather than the MOS for evaluation of the performance of objective speech quality measures [2]. One reason for this claim is based on the observation that there is a difference between the MOS test and objective speech quality measures. While the subjects in a MOS test determine the speech quality without the reference speech, objective speech quality measures are based on the distortion using a reference. Table 1 shows the correlation coefficients of the MBSD and P.861. Two different correlation coefficients can be defined. One is the correlation coefficient associated with each utterance, defined as Per Speech. The other is the correlation

coefficient associated with each condition or each coder and defined as Per Coder. For the situation where objective measures are used for the coder evaluation, the correlation coefficient associated with Per Coder can be used. However, the correlation coefficients for each coder may increase the value of the correlation coefficient by compensating for two oppositely correlated components. So, we report both correlation coefficients for the evaluation of objective quality measures. Even though the correlation coefficient of the MBSD per speech is slightly better than P.861, the performance of the MBSD per coder is not as good as P.861.

Table 1. Correlation coefficients of MBSD and P.861.

	Per Speech	Per Coder
MBSD	0.9001	0.9582
P.861	0.8933	0.9801

3.3. Improvement of MBSD by Scaling Noise Masking Threshold

It has been shown that there is an improvement of the performance of the MBSD by using noise masking threshold. However, since the noise masking threshold calculation is based on the psychoacoustics in which single tones and narrow band noises are usually used, the noise masking threshold might not be very accurate if it is directly applied to nonstationary signals such as speech. So, we examined the performance of the MBSD by scaling the noise masking threshold. In other words, $M(i)$, the indicator of perceptible distortion, is determined by comparing the loudness difference to the scaled noise masking threshold. Figure 4 shows the relationship between the performance of the MBSD and scaling factor. Note that the scaling factor of 0.7 gives the highest correlation coefficient per coder.

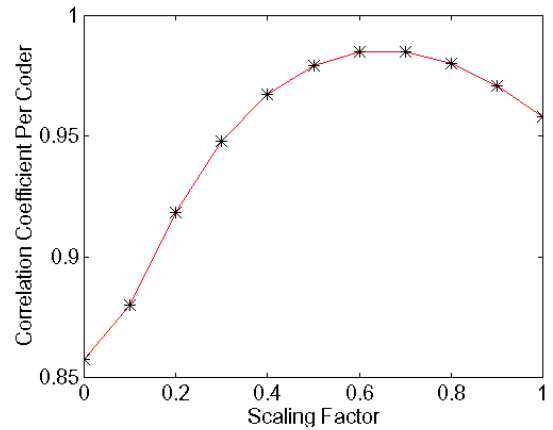


Figure 4. Relationship between the performance of the MBSD and the scaling factor.

Note that the MBSD with scaling factor of 0.7 is identified as the MBSD II. Table 2 shows the correlation coefficients of MBSD II and P.861. The performance of the MBSD II per coder is as good as P.861 and the performance of the MBSD II per speech is clearly better than P.861.

Table 2. Correlation coefficients of MBSD II and P.861.

	Per Speech	Per Coder
MBSD II	0.9252	0.9851
P.861	0.8933	0.9801

4. CONCLUSION

The MBSD is a modified version of the conventional BSD, which incorporates the noise masking threshold. The noise masking threshold plays an important role in estimating perceptual distortion in the MBSD. The MBSD II performance is improved over the MBSD by adopting a scaling factor of 0.7. Its performance per coder is as good as ITU-T Recommendation P.861 and its performance per speech is better than P.861. Currently, the performance of the MBSD measures is being examined with other speech databases.

Acknowledgment:

We wish to thank Peter Kroon of Lucent Technologies for supplying original and coded speech and associated MOS scores.

5. REFERENCES

[1] W. Yang, M. Dixon and R. Yantorno, "A modified bark spectral distortion measure which uses noise masking threshold," IEEE Speech Coding Workshop, pp. 55-56, Pocono Manor, 1997

[2] W. Yang, M. Benbouchta and R. Yantorno, "Performance of the modified bark spectral distortion as an objective speech quality measure," ICASSP, vol. 1, pp. 541-544, Seattle, 1998

[3] ITU-T Rec. P.861, "Objective quality measurement of telephone-band speech codecs," Geneva, 1996

[4] D. Sen and W. H. Holmes, "Perceptual enhancement of CELP speech coders," ICASSP, vol. 2, pp. 105-108, 1994

[5] S. Wang, A. Sekey and A. Gersho, "An objective measure for predicting subjective quality of speech coders," IEEE J. on Select. Areas in Comm., vol. SAC-10, pp. 819-829, 1992

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

[7] K. Lam, O. Au, C. Chan, K. Hui, and S. Lau, "Objective speech quality measure for cellular phone," ICASSP, vol. 1, pp. 487-490, 1996

[8] M. M. Meky and T. N. Saadawi, "A perceptually-based objective measure for speech coders using abductive network," ICASSP, vol. 1, pp. 479-482, 1996

[9] S. Voran and C. Sholl, "Perception-based objective estimators of speech quality," IEEE Speech Coding Workshop, pp. 13-14, Annapolis 1995

[10] J. Johnston, "Transform coding of audio signals using perceptual noise criteria," IEEE J. on Select. Areas in Comm., vol. SAC-6, pp. 314-323, 1988

[11] D. Sen, D. H. Irving and W. H. Holmes, "Use of an auditory model to improve speech coders," ICASSP, vol. 2, pp. 411-414, 1993

[12] S. Voran, "Estimation of perceived speech quality using measuring normalizing blocks," IEEE Speech Coding Workshop, pp. 83-84, Pocono Manor 1997