PERFORMANCE OF THE MODIFIED BARK SPECTRAL DISTORTION AS AN OBJECTIVE SPEECH QUALITY MEASURE

Wonho Yang, Majid Benbouchta and Robert Yantorno

Speech Processing Lab
Electrical & Computer Engineering, Temple University, Philadelphia, PA 19122-6077
wonho@astro.temple.edu, mbenbouc@nimbus.temple.edu, ryantorn@nimbus.temple.edu

ABSTRACT

The Modified Bark Spectral Distortion (MBSD), used for an objective speech quality measure, was presented previously [1]. The MBSD measure takes into account the noise masking threshold in order to use 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, performance of the MBSD is reported in terms of frame sizes, speech classes, and spectral regions. The performance of the MBSD is not very sensitive to the frame size. The performance of the MBSD for voiced speech is almost the same as for non-silent speech. The high frequency region appears to play an important role in human perception of speech quality.

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 [2].

Among the various different objective speech quality measures, we have been interested in the perceptual distortion measures such as Bark Spectral Distortion (BSD) [3] and Perceptual Speech Quality Measure (PSQM) [4]. Since the development of the BSD, it has become a good candidate for a highly correlated objective quality measure, according to several researchers [5][6][7]. 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 [3].

Even though the conventional BSD measure showed a relatively high correlation with MOS scores, there are areas for possible improvement. Motivated by the transform coding of audio signals, which uses the noise masking threshold [8], 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 threshold replaces the empirically derived distortion threshold value used in the conventional BSD [3]. The concept of a noise masking threshold was also used to improve speech quality in coder development [9]. 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 investigate the performance of the MBSD in terms of different frame sizes, speech classes, and spectral regions.

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. 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. This procedure is same as that of the BSD. 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 in 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 [3]. The importance of defining an appropriate perceptual distortion metric was discussed in [10]. An initial attempt to search for a proper metric is addressed in this paper. It has been determined that the most appropriate metric is the average difference of two loudnesses; details of this will be discussed later.

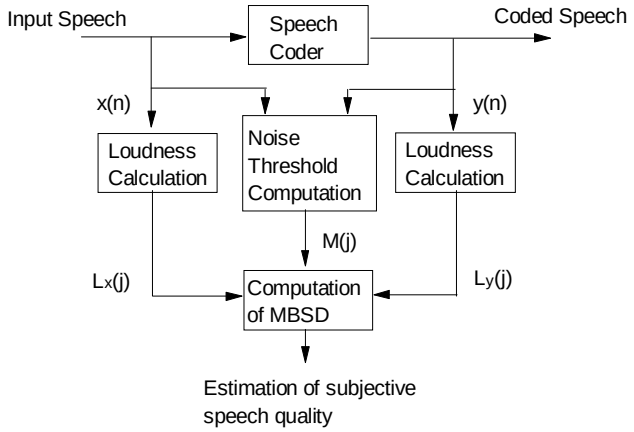

Figure 1. Block diagram of MBSD method

The noise masking threshold is estimated by critical band analysis, spreading function application and absolute threshold consideration [8]. 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 (2). 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, while the BSD measure is the average squared Euclidean distance of estimated loudness; see equation (1).

$$BSD = \frac{\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{K}\left[L_x^{(j)}(i) - L_y^{(j)}(i)\right]^2}{\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{K}\left[L_x^{(j)}(i)\right]^2} \quad (1)$$

$$MBSD = \frac{1}{N}\sum_{j=1}^{N}\left[\sum_{i=1}^{K} M(i)\left|L_x^{(j)}(i) - L_y^{(j)}(i)\right|^n\right] \quad (2)$$

where,

N : number of frames processed
K : number of critical bands
$M(i)$: Indicator of 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. First, we used various different distortion metrics to search for a proper metric in the MBSD. Second, we compared the performance of the BSD with that of the MBSD. Third, we investigated the performance of the MBSD with various frame sizes, speech classes and spectral regions.

For the first and the second experiments, we computed the BSD and the MBSD measures frame by frame, with the frame length of 160 samples. Each frame was weighted by a Hanning window. We processed only voiced frames. This was based on two reasons. One is that the conventional BSD showed a better performance with voiced portions of speech only [3]. The other reason is that research [11] has shown that degradation in quality of LPC speech is not due to coding the unvoiced portion of speech. This suggests that measuring the speech quality for unvoiced speech is not necessary. We used a speech data set which included MNRU distortions and various different types of speech coders. Since the BSD measure is a comparison measure of two speech utterances, we estimated the MOS difference rather than MOS of the original speech and the coded speech with a second-order regression analysis. The reason why we used MOS difference rather than MOS will be explained later. In our experiment, 64Kbps PCM was regarded as original speech.

3.1. Search for a proper metric

In the BSD, the squared Euclidean distance was used for the distortion metric, but it was never determined if this was the most appropriate metric. In order to determine a proper metric which will match the human perception of distortion in the MBSD,

various metrics have been examined. These metrics are limited by the variation of the first and the second norms. The results of the experiments are shown in Table 1. Results of this experiment indicate the importance of a proper metric. Depending upon the metric, the correlation coefficient could vary by 0.01 to 0.05. The average difference of estimated loudness of the 5th metric in the Table 1 (Metric I) showed the highest correlation coefficient. For this reason we have chosen to use this metric for the rest of the experiments. Currently, the validation of this metric is being examined with different speech databases.

3.2. Correlation coefficients with MOS difference

Correlation coefficients with MOS scores have been the traditional evaluation tool for the performance of objective speech quality measures. In this section, we suggest 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. One reason for this claim is based on the observation of the 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. In other words, the MOS test is an absolute test while objective speech quality measures are comparison measures. Therefore, it would be more appropriate to evaluate an objective speech quality measure with MOS difference. The second reason is that the correlation coefficients obtained using the MOS difference (Metric II, Table 1) are higher than those obtained using the MOS (Metric I, Table 1). Therefore, we have chosen to determine the correlation coefficients using the MOS difference for evaluating the performance of objective speech quality measures for the remainder of the experiments.

Table 1. Correlation Coefficients with MOS (I) versus Correlation Coefficients with MOS difference (II) for various metrics

Metric	1	2	3	4	5
I	0.940	0.931	0.917	0.911	0.898
II	0.956	0.946	0.938	0.931	0.906

Metric 1 is eqn. 2 (MBSD) with n = 1.
Metric 2 is eqn. 2 (MBSD) where n = 2.
Metric 3 is eqn. 2 (MBSD) divided by average loudness of original speech and n = 2.
Metric 4 is eqn.2 (MBSD) divided by total loudness squared of original speech and n = 2.
Metric 5 is eqn. 1 (BSD).

3.3. Comparison of MBSD with BSD

Table 2. shows the correlation coefficients of the BSD and the MBSD. Note, the BSD used here had no empirically determined power threshold. The MBSD showed higher correlation coefficients than the BSD. In addition, the performance of the MBSD measure is more consistent for both male and female speech than that of BSD. From these preliminary results, we consider that the MBSD is an improvement over the conventional BSD.

Table 2. Correlation coefficients of BSD and MBSD

	MIXED	FEMALE	MALE
BSD	0.898	0.938	0.854
MBSD	0.956	0.969	0.945

3.4. Performance of MBSD with frame sizes, speech classes, and spectral regions

In order to examine the performance of the MBSD, we performed several experiments. We were interested in the performance of the MBSD with different frame sizes, speech classes and spectral regions.

Table 3. summarizes the performance of the MBSD with different frame sizes and speech classes. The frame size was varied from 40 samples to 400 samples. Speech was classified by hand-labeling silence, voiced, unvoiced and transition regions of speech. According to this table, the MBSD showed the best performance with the frame size of 160 samples and processing all of non-silent regions. It should be noted that the performance of the MBSD is not very sensitive to the frame size variation in the range between 40 samples and 400 samples.

Table 3. Correlation coefficients of the MBSD with different frame sizes and speech classes

	FRAME SIZE (samples)		
SPEECH CLASS	40	80	160
VOICED	0.956	0.956	0.955
UNVOICED	0.604	0.657	0.691
TRANSITIONAL	0.627	0.731	0.794
NON-SILENT	0.943	0.955	0.957
	FRAME SIZE (samples)		
SPEECH CLASS	240	320	400
VOICED	0.954	0.954	0.953
UNVOICED	0.718	0.736	0.745
TRANSITIONAL	0.816	0.709	0.674
NON-SILENT	0.956	0.955	0.954

The last experiment with spectral regions showed very interesting results. We investigated the sensitivity of spectral regions to human perception of speech quality. The spectral regions are divided into three regions: low frequency, mid frequency and high frequency. Table 4. shows the corresponding critical bandwidths, frequency bandwidths and correlation coefficients of each spectral region. According to these results, the high frequency region

appears to play an important role in the human perception of speech quality. We know that the low frequency region plays an important role in speech intelligibility. However, our results demonstrate that the high frequency region is important for speech quality, in contrast with the importance of low frequency region for intelligibility.

Table 4. Critical Bandwidth and Correlation Coefficients for Spectral Regions

	Critical bands	Frequency bandwidth	Correlation coefficients
low frequency region	1 - 8	100-1080	0.554
mid frequency region	9 - 13	1081-2320	0.926
high frequency region	14 - 18	2321-4400	0.953
all frequency regions	1 - 18	100-4400	0.956

4. CONCLUSION

The MBSD is a modified conventional BSD, which incorporates the noise masking threshold. This modification replaces an empirically determined power threshold in the BSD. The MBSD uses a new distortion metric, which has been determined by comparing the performances of various different metrics. The distortion in the MBSD is defined as the average difference of loudnesses. In order to optimize the performance of the MBSD, the frame size of the MBSD was varied for different speech classes. It was found that the performance of the MBSD is not very sensitive to frame size and the performance of the MBSD with non-silent regions is slightly better than that with voiced regions. According to our experiments, we suggest that the MBSD shows an appropriate performance when it processes non-silent regions with the frame size of 160 samples.

From the performance of the MBSD with spectral regions, human perception of speech quality is highly sensitive to high frequency regions.

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

[3] 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

[4] 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

[5] 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

[6] 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

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

[8] 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

[9] 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

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

[11] G. Kubin, B. S. Atal and W. B. Kleijn, "Performance of noise excitation for unvoiced speech," IEEE Speech Coding Workshop, pp. 35-36, 1993