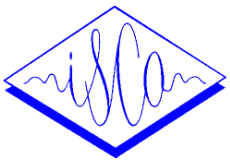

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

A Bark Coherence Function For Perceived Speech Quality Estimation

Sang-Wook Park, Seung-Kyun Ryu, Young-Cheol Park and Dae-Hee Youn*

ASSP Lab., Dept. of Electrical and Computer Eng., Yonsei University
134 Shinchon-dong Sudaemon-ku, Seoul 120-749, Korea
phone: +82-2-361-2863, Fax: +82-2-312-4584, E-MAIL: latest@lethe.yonsei.ac.kr
*Center for Signal Processing Research, Yonsei University

ABSTRACT

A new methodology for perceptual quality measure is presented. The new method defines the bark coherence function (BCF) as a new cognition module. False prediction errors are occasionally observed in previously developed perceptual measures when they are applied to the end-to-end speech quality measurement of communication systems. Those errors are mainly caused by the linear distortion of the analogue interface of the system being evaluated. The BCF itself normalizes those effects of linear filtering, so that it is ideal for the speech quality assessment of mobile communication systems. In addition, the proposed scheme does not require the local as well as global scaling, so that it is robust to the difference between the original and received speech levels. To evaluate the performance of the new perceptual model, the regression analysis was performed with CDMA digital cellular, CDMA personal communication service (PCS) and speech codec's. The correlation coefficients computed using the BCF showed noticeable improvements over the PSQM that is recommended by ITU-T. Robustness of the BCF to various conditions was also tested.

1. INTRODUCTION

Monitoring the speech quality of mobile communication systems is essential for maintaining required quality of service. Information of the speech quality has traditionally been provided by human listeners. But evaluation by repeated listening tests at various sites within the metropolitan areas is almost impossible due to its expensive and time-consuming nature. As a result, it has been an issue of importance to develop a objective speech quality measure that can be used to predict the subjective assessment of speech quality.

A variety of perceptual methods have been suggested to accomplish this objective, that include bark spectral distortion (BSD) [1], perceptual speech quality measure (PSQM) [2] and measuring normalizing block (MNB) [3]. These methods employ the perceptual transformation and the cognition module as two vehicles for the perceptual measure. The perceptual transformation is the representation of an audio signal in a way that is approximately equivalent to the human hearing process. In the cognition module, the distance is measured in order to seek the difference between two perceptually transformed signals. Regardless of the fact that these methods have been successful in some applications, they are often irrelevant to the end-to-end quality measurement of mobile communication systems because they were developed mainly to evaluate telephone-band speech codes [4]. Irrelevancy is due to the linear filtering by the analog interface of the communication systems. The analog interface introduces significant spectral distortions [4]. However, these distortions cannot be considered as negative aspects because, sometimes, they can even improve the perceived speech quality. Thus, it is necessary to develop a perceptual quality measure that is independent of the linear filtering by the analog interfaces placed between measurement points. Furthermore, it is desirable that the measurements can be applicable to as many types of systems as possible.

In this paper, a new methodology for perceptual quality measure is presented. As a new cognition module of the perceptual quality measurement system, the bark coherence function (BCF) is defined. The BCF is based on the well-known magnitude square coherence (MSC) function. However, in prior to the MSC computations, the signals are perceptually weighted using a psychoacoustic model. Later, the auto and cross bark spectra are computed from the weighted signals in order to obtain the BCF. By using the BCF, it is possible to alleviate the effects of the linear distortion caused by the analogue interface of the communication systems being evaluated. As a result, large false errors can be prevented. Also, it is noted that the accuracy of the new quality measurement system is always comparable to the PSQM that is recommended by ITU-T, even in such cases that linear distortions are not existing.

The outline of the paper is as follows. We start with a section about the problem of objective measure on end-to-end mobile system. In section 3, we describe the proposed method, the BCF. Section 4 then describes the performance of the BCF compared with current objective measures. Section 5 provides our conclusions.

2. LIMITATION OF CURRENT PERCEPTUAL QUALITY MEASURE

2.1 Linear Filtering

In the digital cellular system, speech is encoded by speech codec's at mobile stations, then sent to a based station via wireless channels. The decoded speech is heard to a subscriber of the public switched telephone network (PSTN). Figure 1 shows a block diagram for the transmission of speech in the digital mobile system.

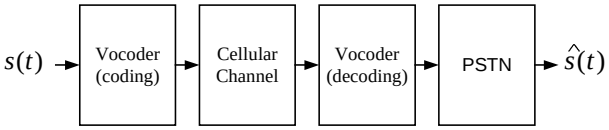

Figure 1: Transmission of speech on a digital mobile system.

Most standard speech coders for commercial mobile communication have almost transparent quality. Normally, speech is distorted when the received speech frame cannot be reconstructed due to channel errors. Linear filtering by the analog interface in the communication system is not likely to degrade the subjective quality [4]. However, as noted in [4], the presence of significant linear filtering brings negative effects to the speech quality measurement system. In this section, examples are given to show that conventional perceptual quality measurement methods may produce large estimation errors due to the presence of analog filtering circuit.

We recorded speech data from Korean CDMA personal communication service (PCS). Two groups of data were recorded via two different types of signal paths: from mobile to private branch exchange (PBX) through the PSTN (group A), and from mobile to the PSTN (group B). In particular, the path from PSTN to PBX exhibited a band-limited frequency response [4]. Subjective evaluation using the recorded data was performed with 30 normal, naïve, paid listeners. Figure 2 shows the histogram of mean opinion score (MOS) and perceptual speech quality measure (PSQM) over the two groups. The PSQM has been recommended as an objective quality measurement of telephone-band speech codec's by ITU [5].

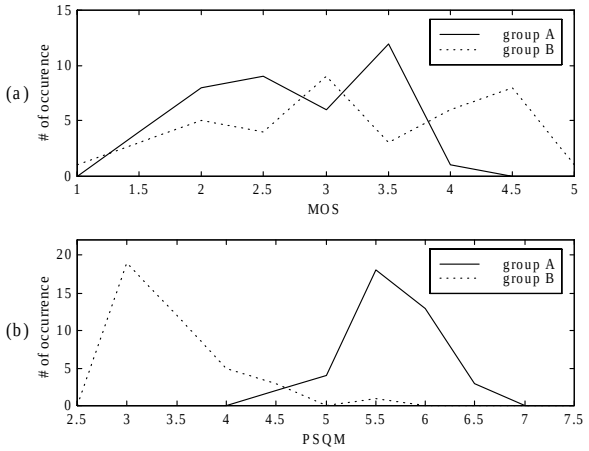

Figure 2: Distribution of subjective quality and objective quality over CDMA speech data. (a) MOS distribution (b) PSQM distribution

MOS distribution does not reveal any significant difference of speech quality between two groups of data. But the PSQM shows a big difference. Given the PSQM estimates, we may be mislead to judge that the signal path of group B would guarantee better speech quality than that of group A. The reason for this false judgement is that the PSQM provides biased estimates due to the linear filtering of the PBX attached to PSTN. According to our experiment, most of psychoacoustic-motivated perceptual quality measurement methods showed the same results that may mislead us to false judgements. Thus, it can be said that these measurement results can not efficiently predict the subjective quality under different end-to-end measurement conditions.

2.2 Scaling

Global scaling is made in compensation for the overall gain of the end-to-end system. Also, signals are locally scaled within each frame as compensation for slow variation of gain. Usually, distorted speech is corrupted by an additive noise occupying certain frequency band. But since the total energy increases as the noise is added, the scaling will reduce the energy level of the speech outside the frequency band where the noise is present. Thus, the scaling process can cause a new spectral distortion when the speech is corrupted by the noise. Figure 3 shows an example of scaling. In Figure 3 (a), the solid line indicates the original bark spectrum, the dotted line is unscaled and distorted, and the dashed line is scaled and distorted spectra. The speech was corrupted by a noise in high frequency band. The listener may complain about this noise. After the scaling was done, as shown in Figure 3 (b), new spectral distortion occurred in other frequency region while the distortion is still present in the noise region. Considering that the conventional methods measure mainly the spectral difference, it can be said that the scaling process can increase the possibility of poor estimation of speech quality. To overcome this difficulty, the cognition module should be independent of the signal level.

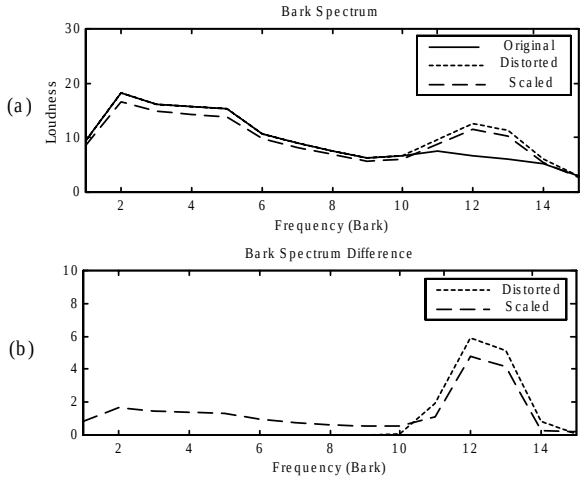

Figure 3: Scaling problem by positive noise. (a) bark spectra of original (solid), distorted (dotted) and scaled (dashed) speech. (b) bark spectrum difference between distorted and original speech and scaled and original speech.

3. BARK COHERENCE FUNCTION

The current issue of psychacoustically motivated quality measure is the cognition module, rather than the perceptual transformation. The perceptual transformation for the modeling of the human hearing has been well established [6][7]. It is clear that the frequency resolution of human hearing is not uniform and the sensitivity of ear is a function of frequency. Also, the perceived loudness is related to the signal intensity in a non-linear way. These features of the human hearing are modeled and implemented via the perceptual transformation.

On the other hand, the cognition module models the information processing of human brain. But since the

complicated auditory processing stages in the human brain are far from being finally explored, a satisfying model will not be available in the short or even long term. Thus, the cognition module is necessarily implemented as a rough simplification of the information processing of human brain.

In this paper, we propose a new cognition module for the perceptual quality measurement system, referred to as the bark coherence function (BCF). The BCF is based on the magnitude-squared coherence (MSC) function. The BCF is based on the magnitude-squared coherence (MSC) function computed in bark domain. The MSC is defined as [8]

$$\gamma_{xy}^2(f) = \frac{|S_{xy}(f)|^2}{S_{xx}(f)S_{yy}(f)} \qquad (1)$$

where $S_{xy}(f)$ is cross-spectrum between input $x(t)$ and output $y(t)$, $S_{xx}(f)$ and $S_{yy}(f)$ are auto-spectra of $x(t)$ and $y(t)$, respectively. In ideal noise-free, distortion-free environments, one can obtain $\gamma_{xy}^2(f)$ =1 for all *f*. On the other hand, the MSC will become zero, if only the noise signals are observed at the output. Since most systems in real environments are not linear, the MSC value stays in between zero and unity.

The real end-to-end mobile communication systems may be simplified as the figure shown below.

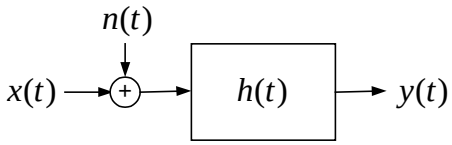

Figure 4: Simplified blockdiagram of mobile system.

In Figure 4, $x(t)$ is the input speech, $n(t)$ is a non-linear additive noise due to channel errors and/or speech coder, $h(t)$ is the transfer function of the analog part of communication system which can be modeled as a linear system, and $y(t)$ is a distorted speech heard to listener. The MSC itself is independent of the linear filtering due to analog system.

In order to transform the MSC to a perceptual quality measure, the BCF is defined. The same perceptual model that is used in the conventional BSD [1] is employed in this method.

The BCF is defined as

$$BCF(b) = \frac{|L_{xy}(b)|^2}{L_{xx}(b)L_{yy}(b)} \qquad (2)$$

where b is the bark frequency, $L_{xx}(b)$ and $L_{yy}(b)$ are auto bark spectra of $x(t)$ and $y(t)$, respectively, and $L_{xy}(b)$ is a cross bark spectrum between $x(t)$ and $y(t)$. The auto bark spectrum follows the definition in Ref. [1]. In addition, the cross bark spectrum is defined as cross spectrum in the loudness domain. Given the cross spectrum between $x(t)$ and $y(t)$, the cross bark spectrum is computed using perceptual weighing filters [1], followed by the loudness compression. Figure 5 shows a block diagram of bark cross spectrum between two signals. To transform speech into the loudness domain, several steps are involved: critical band analysis, equal-loudness pre-emphasis and intensity-loudness power law.

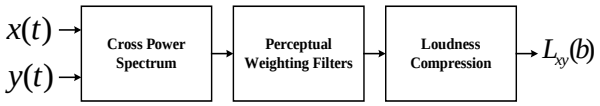

Figure 5: Block diagram of Cross bark Spectrum.

Finally, the bark distortion-to signal ratio (BDSR) is defined as the amount of non-linear distortion normalized to signal power on loudness domain:

$$BDSR = \sum_{b=0}^{N} \frac{1 - BCF(b)}{BCF(b)} \qquad (3)$$

The BDSR represents a measure of perceptual speech quality which can be applicable to any mobile communication system.

4. EXPERIMENT AND RESULTS

In order to evaluate the performance of the BCF, the regression analysis is performed with CDMA digital cellular, CDMA PCS and telephone-band speech coders. Pearson correlations are used throughout this paper. The correlation coefficient will describe the linear relationship between the objective measure and the MOS.

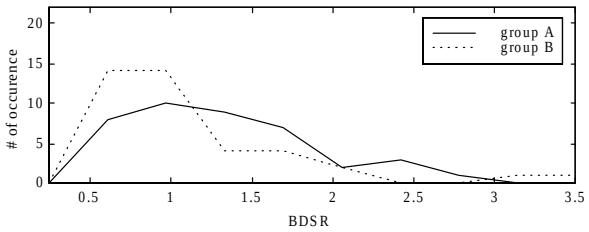

Figure 6: Distribution of BDSR.

Figure 6 shows distribution of BDSR measured with the same data used for Figure 2. Now, it can be clearly seen that the BCF of the two data exhibit great deal of similarity to the MOS. Which means that the effects of the analog system are greatly removed by the BCF measure. Now, it is important to note that the BCF can be considered as a perceptual measure that can account for the effect of linear filtering by the analogue interface in the end-to-end measurement system.

Table 1 shows the correlation coefficient. Test I was conducted with a speech data obtained in the Korean CDMA PCS, same as Figure 2, and Test II was conducted with Korean Advanced Mobile Phone Service (AMPS). Data sets were recorded in Seoul, Korea, where millions of subscribers use mobile phones. In order to record speech data under various situations that include multi-path fading, slow/fast fading, low power and high power, the whole setup was installed in a van moving around metropolitan areas. The original clean speech was played using a portable DAT to a mobile phone via an interface kit. Simultaneously, the distorted speech was recorded from a telephone hybrid unit to the DAT. Similar experiments were performed under down-link conditions. The performance of the PSQM was excellent with each group of data, but when the two groups were mixed, the correlation coefficient was significantly

low. However, the BCF demonstrated its excellent performance throughout the cases considered. Even with each group of data, the BCF showed the results that are better or comparable to the PSQM.

Test	Correlation Coefficient					
	PSQM			BCF		
	A	B	A+B	A	B	A+B
I	0.798	0.911	0.558	0.881	0.926	0.903
II	0.890	0.904	0.852	0.923	0.880	0.891

Table 1: Correlation coefficient with Korean CDMA PCS and Korean AMPS.

Table 2 shows the results of regression analysis performed with another Korean CDMA digital cellular (Test III) and Korean CDMA PCS (Test IV). The recording conditions were almost the same as Test I and II. Superiority of the BCF was again demonstrated in these experiments. In these data sets, variable delay was observed. Variable delay is a characteristic of many packet-based transmission systems and often degrades the performance of perceptual quality measurement systems. Variable delay was more frequently observed in Test III because the number of subscribers of this particular system is almost 8.5 millions. So, the performance of two methods with Test III is lower than that of Test IV. In order to cope with the variable delay, we chose the minimum value of the PSQM and the BCF. The correlation shows the BCF is robust to variable delay

In Test IV, we found strong background noise. It has been know that the background noise in silence periods disturbs perceptual quality measurement in mobile communication systems. That's why the PSQM showed low correlation coefficient compared with that of the BCF. In fact, the local scaling was performed when we computed ND of the PSQM. Disturbing background noise was scaled down, so that its contribution was minimized in prior to the computation of ND. However, those efforts were not so successful in the PSQM. On the other hand, the BCF has no need to global and local scaling. But it showed excellent performance even when the background noise occurred.

Test	Correlation Coefficient	
	PSQM	BCF
III	0.692	0.840
IV	0.929	0.958

Table 2: Correlation coefficient with Korean CDMA Digital cellular and Korean CDMA PCS.

The last experiment was with telephone-band speech codec's, and the results are shown in Table 3. Test V was performed with 2.4kbps ~ 16kbps coders, DAT loop, modulated noise reference unit (MNRU) (5-35dB) and Test VI was with 2.4kbps ~ 32kbps coders, MNRU (5-25dB). North American English was used and subjective tests were performed at different US labs. In Test VI, 64Kbps PCM was regarded as original speech. As can be seen from the results, although the BCF is not primarily aiming at the assessing the speech quality of telephone-band codec's, its performance is still better than or comparable to the PSQM.

Test	Correlation Coefficient	
	PSQM	BCF
V	0.916	0.913
VI	0.878	0.910

Table 3: Correlation coefficient with Speech Coders.

5. CONCLUSION

In this paper, we presented a new methodology of perceptual speech quality measurement. The presented method employs the bark coherence function (BCF) defined in parallel to the MSC in loudness domain. The most important advantage of the BCF is that it is not affected by the analog interface in communication systems, so that its performance is independent of the linear distortions that have little effects on the perceived speech quality. Tests were conducted with CDMA system and telephone-band speech coders. Test results verified robustness of the BCF to the linear distortion due to the presence of analogue interface. In addition, tests with telephone-band codec showed that the BCF was also an excellent test metric for accessing the speech quality of codec's.

6. Acknowledgement

We wish to thank Peter Kroon of Lucent Technologies and Joshua Rosenbluth of AT&T for supplying English original and coded speech and associated MOS scores.

7. REFERENCES

[1] Shihua Wang, et al, "An Objective Measures for Predicting Subjective Quality of Speech", *IEEE J. Select. Areas Commun*., vol 10, No5, pp819-829, June 1992.

[2] J. G. Beerends and J. A. Stemrdink, "A perceptual speech-quality measured based on a psychoacoustic sound representation," *J. Audio Eng. Soc*. vol 42, No 3., pp115-123, March, 1994.

[3] S. Voran, "Objective estimation of perceived speech quality, Part I : Development of the measuring normalizing block technique," *IEEE Trans. on Speech and Audio Processing*, vol. 7, No. 4, pp371-382, July 1999.

[4] A. Rix, R. Reynolds, and M. Hollier, "Robust perceptual assessment of end-to-end audio quality", *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp39-42, October 1999..

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

[6] E. Zwicker and H. Fastl, *Psychoacoustics Facts and Models*, Springer-Verlag, 1990.

[7] Markus Hauenstein, "Application of meddis' inner hair-cell model to the prediction of subjective speech-quality", *in Proc. IEEE Int. Conf. Acoust., Speech Signal Process*, pp 545-548, 1998.

[8] Julius S. Bendat and Allan G. Piersol, *Engineering Applications of Correlations and Spectral Analys*is, John Wiley & Sons, 1980.