Speech Quality Measure for VoIP using Wavelet based Bark Coherence Function

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

ASSP Lab. Dept. of Electrical and Electronic Eng. Yonsei University, Seoul, Korea 120-749
*Center for Signal Processing Research Yonsei University, Seoul, Korea 120-749 latest@assp.yonsei.ac.kr

Abstract

The Bark Coherence Function (BCF) [1] defines a coherence function with loudness speech as a new cognition module, robust to linear distortions due to the analog interface of digital mobile system. Preliminary experiments have shown the superiority of BCF over current measures. In this paper, a new BCF suitable for VoIP is developed. The new BCF is based on the wavelet series expansion that provides good frequency resolution while keeping good time locality. The proposed Wavelet based Bark Coherence Function (WBCF) is robust to variable delay often observed in internet telephony such as VoIP. We also show that the refinement of time synchronization after signal decomposition can improve the performance of the WBCF. The regression analysis was performed with VoIP speech data. The correlation coefficients and the standard error of estimates computed using the WBCF showed noticeable improvement over the PSQM that is recommended by ITU-T.

1. Introduction

Monitoring the quality of speech communication systems is essential for maintaining required quality of service. Information of the speech quality has been traditionally provided by human listeners. But evaluation by repeated listening tests at various sites may be associated with many problems due to its expensive and time-consuming nature. As a result, it has been an issue of importance to develop an objective measure that can be used to predict the subjective assessment of speech quality [2].

A variety of methods based on human perception have been suggested to accomplish this objective, that include Bark Spectral Distortion (BSD) [3], Perceptual Speech Quality Measure (PSQM) [4] and Measuring Normalizing Block (MNB) [5]. 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 packet-based speech transmission systems such as VoIP (Voice over Internet Protocol) [6][7], because they were not developed mainly to evaluate packet-based speech transmission system. Irrelevancy is due to the property of packet-based speech system that may introduce variable delay and jitter. Variable delay often results in inaccurate time alignment between original and distorted speech, so, the current methods measure under estimated speech quality because long sections of speech end up being misaligned [8]. The analog interface of digital speech communications system introduces significant spectral distortions [1][8]. But, 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 robust to variable delay and the linear filtering by the analog interfaces placed between measurement points.

In this paper, a new methodology for speech quality measure is presented. As a new cognition module of the perceptual quality measurement system, the Wavelet based Bark Coherence Function (WBCF) is defined. The Bark Coherence Function (BCF) for the objective quality measure has been well reported in [1]. The WBCF computes coherence function with perceptually weighted speech after wavelet series expansion [10]. Discrete wavelet series expansion satisfies two conflict requirements of signal processing; archive good time resolution while keeping good time locality as well. By using the WBCF, it is possible to alleviate the effects of variable delay of packet-based end-to-end system and linear distortion caused by the analogue interface of the communications systems. It is noted that the accuracy of the new quality measurement system is always comparable to the PSQM that is recommended by ITU-T [11].

The outline of the paper is as follows. We start with a section about the effects of variable delay and linear distortion on objective measurement in section 2 and section 3 respectively. In section 4, we describe the proposed method, the WBCF. Section 5 then describes the performance of the WBCF compared with current objective measures. Section 6 provides our conclusions

2. Variable delay with VoIP

The original and distorted speech signals are to be time aligned before the quality is evaluated by the measurement method. Cross correlation function between the temporal envelopes of the two signals is frequently used. The delay of one signal relative to the other is given by the position of the maximum of the correlation function. This approach is effective to time-invariant bulk delay system such as
waveform coders that well preserve waveforms of speech. However, new packet based speech communication system may occur variable bulk delay. This is a common phenomenon in VoIP that aims at providing a comparable voice quality at a lower cost. Internet was designed to deliver messages reliably regardless of the delay. VoIP generally use User Datagram Protocol (UDP) that does not guarantee delivery, preservation of sequence for real time audio stream instead of Transmission Control Protocol (TCP) that provides for the re-transmission of lost data. So, the incoming packets are buffered considering a different amount of delay and late packet dropping for a continuous audio stream. There are trades off between the buffer size and the end-to-end delay. The large buffer size is good for severe jitter but increases the end-to-end delay. Dynamic buffer resizing that minimize bulky delay and produce good speech quality is frequently occurred during silence interval and imperceptible. The delay change during speech period can be observed with packet loss or late arrival.

The cross correlation function already mentioned is not appropriate for VoIP system. If this method is applied to the signals that suffer from variable delay, the uncertain delay is given by the position of the maximum of the correlation function that is affected by the concentration of energy in speech. The sub-optimal approach is to shift the delay by the position of minimum objective distortion. But, it needs much computing power and cannot estimate subjective quality precisely. So, it is desirable to develop new objective measure to robust variable delay.

3. Linear distortion

In the digital speech communication, speech is encoded by the speech coder at the sender site and sent to the receiver via a transmission channel. The decoded speech is heard to a subscriber of the public switched telephone network (PSTN). Most standard speech coders for the commercial communication have excellent quality. Normally, speech is distorted when the received speech frame cannot be reconstructed due to channel impairments. On the other hand, linear filtering by the analog interface in the communication system is not likely to degrade the subjective quality. It may even improve aspects of speech quality. However, as noted in [1][8], the presence of significant linear filtering brings negative effects to the speech quality measurement system.

4. Wavelet based bark coherence function

The block diagram of the WBCF is shown Figure 1. There are three major processing steps: wavelet series expansion, time alignment, and BCF. The wavelet series expansion is similar to an octave-band filter bank since each successive highpass output contains an octave of the input bandwidth. Original and distorted signals are split first via a two-channel filter bank, the the highpass and lowpass versions are downsampled and lowpass versions split again using the same filter bank, and so on. This leads to a hierarchy of resolution, also called a multiresolution decomposition. The decomposed signals achieve good frequency resolution while keeping good time locality as well.

The time alignment that exploits cross correlation function between the decomposed original and distorted signals follows. The uncertain delay is given by the position of the maximum of the correlation function that is affected by the concentration of energy in the whole bandwidth signals. The alignments with multiresolution decomposed signals give more accurate delays because the energy of signals disperses in each octave bands.

After, time alignment, the Wavelet based BCF is computed. Although octave bands from wavelet series expansions is similar to bark scale that have good frequency resolution at low frequency and poor at high frequency, bark scale transform is exploited for high correlated objective quality. This procedure is almost same with our previous method [1].

The BCF was motivated by the Coherence Function (CF) method [9] developed by Bell Northern Research that is a measure of signal-to-distortion ratio (SDR) weighted to account for hearing sensitivity, noise threshold effects and the handset receiver sensitivity. We improved the CF using psychoacoustics and defined the BCF. The ordinary magnitude-squared coherence function (MSC) is defined as

\[
\gamma^2_{xy}(f) = \frac{|S_{xy}(f)|^2}{S_{xx}(f)S_{yy}(f)}
\]  

(1)

where \(S_{xy}(f)\) is cross power spectrum between input, \(x(t)\) and output, \(y(t)\), \(S_{xx}(f)\), \(S_{yy}(f)\) are auto power spectra of \(x(t)\), \(y(t)\) respectively. It can be interpreted as the correlation between the input and output signal at frequency \(f\). Next, coherent signal power, \(CP(f)\) and non-coherent distortion power, \(NCP(f)\) are estimated from \(\gamma^2_{xy}(f)\) via (2) and used to develop the Distortion-to-Signal Ratio (DSR), shown in (3) [14]

\[
CP(f) = \gamma^2_{xy}(f)|S_{xy}(f)|^2
\]

\[
NCP(f) = |1-\gamma^2_{xy}(f)||S_{xy}(f)|^2
\]  

(2)

\[
DSR(f) = \frac{NCP(f)}{CP(f)} = \frac{1-\gamma^2_{xy}(f)}{\gamma^2_{xy}(f)}
\]  

(3)

In ideal noise-free, distortion-free environments, one can obtain \(\gamma^2_{xy}(f)=1\) for all \(f\), so the DSR is zero. On the other hand, the DSR will become infinite, if only the noise signals are observed at the output. Since most systems in real environments are not linear, the DSR value stays in between zero and infinite.
In order to compute the DSR in loudness domain, the WBCF is newly defined. In the WBCF, the perceptual model of the conventional BSD [3] is used: critical band analysis, equal loudness preemphasis and intensity loudness law. Critical band analysis is that the human auditory system is known to have a poorer frequency resolution at high frequencies than at low frequencies. Equal loudness preemphasis incorporates the fact that the ear is not equally sensitive to energy at different frequencies. Intensity loudness law simulates the nonlinear relation between the intensity of sound and its perceived loudness. From the Eq. (1) to (3) are redefined with decomposed perceptually weighted speech. The WBCF at stage $j$ is defined as

$$WBCF(b)^j = \frac{|L_{sx}(b)^j|^2}{L_{sy}(b)^j L_{sy}(b)^j}$$  \hspace{1cm} (4)$$

where $b$ is the bark frequency, $L_{sx}(b)^j$ and $L_{sy}(b)^j$ are auto bark spectra of decomposed signal $x(t)$ and $y(t)$ at stage $j$, respectively, and $L_{sy}(b)^j$ is a cross bark spectrum between decomposed $x(t)$ and $y(t)$ at stage $j$. The auto bark spectrum follows the definition in Ref. [3]. 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 [3][12], followed by the loudness compression. Figure 2 shows a block diagram of cross bark spectrum between two signals.

![Figure 2. Block diagram of cross bark spectrum.](image)

Finally, the Wavelet based Bark Distortion-to Signal Ratio (WBDSR) is defined as the amount of non-linear distortion normalized to signal power on loudness domain:

$$WBDSR = \sum_{j=b}^{J} \sum_{b(j)=0}^{b(j)} \frac{1-WBCF(b)^j}{WBCF(b)^j}$$  \hspace{1cm} (5)$$

where $sb(j)$ is the start of bark frequency at stage $j$ and $eb(j)$ is the end of bark frequency at stage $j$ respectively, $w_b$ is weighting factor for high correlated objective quality and $J$ is level of wavelet expansion.

5. Experiment and results

In order to evaluate the performance of the WBCF, the regression analysis was performed with VoIP and CDMA PCS. Previous studies often used a second or third order polynomial for the regression analysis, which does not increase monotonically. But in most subjective tests, MOS lies in the range of 1-5, while the perceptual measure such as the PSQM usually covers from zero to infinite. Thus, common polynomial cannot explain the value of objective measure outside the vertex of regression function. We used a new regression function with asymptotes at 1 and 5:

$$\hat{y} = 1 + \frac{4}{1 + \exp(ax + b)}$$  \hspace{1cm} (6)$$

where $a$ and $b$ are coefficients of the regression analysis, $x$ is objective speech quality and $\hat{y}$ is estimated subjective speech quality. Pearson correlations between estimated subjective quality and real MOS was used throughout this paper. The correlation coefficient will describe the linear relationship between the objective measure and the MOS. The closer to 1 the correlation coefficient is, the better the objective speech quality measure is at estimating the subjective quality. The standard error of estimates (SEE) [13] was also represented. The SEE is an unbiased statistic for the estimate of the deviation from the regression function between estimated MOS and the actual MOS.

Table 1 shows correlation coefficient and the SEE per speech after regression analysis using the Eq. (6) with VoIP system that is commercially used and based on H.323 [6]. The speech coders was 6.3 kbps G.723.1 that is frequently used in many VoIP system. Two female and two male sentences (Korean) were selected in NTT CD-ROM. Because the status of real internet was not easily controlled for experimental purpose, we used the NIST Net (National Institute Standards and Technology Network Emulation Tool) [15] that is a general-purpose tool for emulating performance dynamics in IP networks. This tool is designed to allow controlled, reproducible experiments with network performance sensitive/adaptive applications and control protocols in a simple laboratory setting. The recording procedure is shown in Figure 3. Speech was encoded by coders at H.323 system A and speech packets were transmitted to NIST Net. Then, the packet was delayed and/or dropped at specified delay time and packet loss rate. The transmitted packets were reconstructed at H.323 system B. The specified network parameters were that the averaged delay time was 150ms, the standard deviation of delay time (jitter) was from 0 to 60ms and the packet loss rate was from 0 to 25%. Informal listening test with hifi-head phone was performed by 35 normal persons who were volunteers.

![Figure 3. Block diagram VoIP speech data.](image)

The PSQM I and the BCF I exploited cross correlation function for time alignment. The delay for the PSQM II and BCF II was given by the position of minimum objective distortion respectively. The WBCF used only cross correlation function. As shown in Table 1, the performance of PSQM and BCF are fairy good. The sub-optimal time alignment method that was used for the PSQM II and BCF II improved the performance of current measurement, but not enough for VoIP application. The WBCF has noticeable improvement over the PSQM and the BCF, because it is robust to variable delay due to wavelet expansion and time alignment between decomposed signals. The haar filters were used for wavelet expansion. These filters are very simple, but
exploited for high correlated objective quality compared with other filters such as Daubechies and Symlets. The analysis frame size was 480 samples for 8KHz speech and silence period was excluded for analysis. The wavelet expansion level was 5.

Table 1. Correlation coefficient and SEE with VoIP

<table>
<thead>
<tr>
<th></th>
<th>Correlation Coefficient</th>
<th>SEE</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSQM I</td>
<td>0.832</td>
<td>0.600</td>
</tr>
<tr>
<td>PSQM II</td>
<td>0.849</td>
<td>0.692</td>
</tr>
<tr>
<td>BCF I</td>
<td>0.812</td>
<td>0.631</td>
</tr>
<tr>
<td>BCF II</td>
<td>0.822</td>
<td>0.615</td>
</tr>
<tr>
<td>WBCF</td>
<td>0.891</td>
<td>0.492</td>
</tr>
</tbody>
</table>

Table 2 shows the result of the second experiment. The correlation coefficient and the SEE were computed for verification of robustness to linear distortion of the WBCF. Speech data sets were recorded with CDMA PCS 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. 35 volunteers attended informal subjective test where test materials of two groups were mixed and played randomly. 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. 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 WBCF demonstrated its excellent performance throughout the cases considered. Even with each group of data, the WBCF showed the results that are better or comparable to the PSQM.

Table 2. Correlation coefficient and SEE with PCS

<table>
<thead>
<tr>
<th></th>
<th>Correlation Coefficient</th>
<th>SEE</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSQM</td>
<td>A 0.718 B 0.908</td>
<td>0.504 0.442</td>
</tr>
<tr>
<td></td>
<td>A+B 0.568</td>
<td>0.766</td>
</tr>
<tr>
<td>WBCF</td>
<td>A 0.894 B 0.875</td>
<td>0.324 0.511</td>
</tr>
<tr>
<td></td>
<td>A+B 0.881</td>
<td>0.441</td>
</tr>
</tbody>
</table>

6. Conclusion

A new BCF incorporated with the wavelet series expansion, referred to as WBCF, has been presented. Wavelet series expansion provides good frequency resolution and good time locality at the same time, so that time alignments with decomposed signals are done in a convenient way. The time alignment with multi-resolution decomposed signals gives more accurate prediction of perceptual quality of the synthesized speech. Experimental results conducted with VoIP speech data showed that the WBCF outperformed the PSQM and the conventional BCF. The WBCF is also robust to linear distortion due to analog interface of communication system because it is based on the BCF.

7. References