# MEL-CEPSTRAL DISTANCE MEASURE FOR OBJECTIVE SPEECH QUALITY ASSESSMENT

Robert F. Kubichek
Electrical Engineering Department
University of Wyoming, Box 3295
Laramie, WY USA
**Phone**: *(307) 766-3182* **E-mail**: *kubichek@corral.uwyo.edu*

## ABSTRACT

Reliable objective measures of speech quality have been sought for many years. Unfortunately, no known method has proven to be sufficiently reliable for use in general speech quality applications. Recently, research has been shifting toward the use of perceptual-based algorithms in an effort to improve the accuracy of objective assessments. This paper proposes a perceptually motivated modification to the well-known Cepstral Distance measure (CD) based on the Mel frequency scale and critical-band filtering. The new objective parameter is referred to as the Mel Cepstral Distance (MCD). This paper measures and compares the performance of the CD and MCD algorithms by applying them to a dataset representing low bit-rate CELP-coded speech with simulated channel conditions.

## 1.0 INTRODUCTION

Many objective speech quality methods have been proposed over the last two decades. Few, however, have demonstrated a capability for accurately predicting subjective listener scores [1]. One exception is the cepstral distance measure (CD), which has yielded promising results for both waveform and non-waveform speech coders [2]-[4]. Advantages of the CD include ease of implementation and limited computational requirements. Unfortunately, as with many other objective techniques, the CD has shown sensitivity to coder type, language, and other impairment conditions that are unrelated to subjective quality. As a result, the algorithm's mean-squared error has been unacceptably high for use in most applications.

To improve the performance of objective speech quality techniques, research in this area is beginning to focus on perceptually-based objective algorithms that incorporate psycho-acoustic models of human hearing. For example, Wang et al [5] have proposed the Bark Spectral Distortion measure (BSD) which computes spectral distortion by estimating a Bark-scaled spectrum, applying critical-band filters, and accounting for loudness and loudness-level effects of the human ear.

In this paper, the CD objective measure is modified to incorporate non-linear frequency dependent effects of the ear by applying Mel-scale critical-band filters in the frequency domain. There are two main goals: the first is to gain a better understanding of the importance of critical-band filters to subjective quality, and the second is to improve the effectiveness of the CD objective measure while retaining its advantages of computational simplicity.

## 2.0 CEPSTRAL DISTANCE

The CD measure uses cepstral coefficients computed from the input (source) and output (processed) speech records on a frame-by-frame basis. For the k-th frame, the root-mean-square difference is computed as

$$CD(k) = \sqrt{\sum_{i=1}^{16} [C_X(i,k) - C_Y(i,k)]^2}\ , \qquad (1)$$

where $C_X(i,k)$ and $C_Y(i,k)$ are the i-th cepstral coefficients of the input and output speech record, respectively. The real cepstral coefficients can be defined as follows:

$$C_x(i,k) = Real[IDFT\{\log|DFT(x_k)|\}], \qquad (2)$$

where "DFT" and "IDFT" are the discrete Fourier transform and its inverse, and $x_k$ is the k-th speech frame (typically 256 samples at 8 kHz sampling rate). Alternatively, cepstral coefficients can be derived from LPC coefficients.

 

The zero-th coefficient is not generally included in (1) since it is primarily affected by system gain rather than system distortion. Further, by excluding high-order coefficients, CD represents a cepstrally smoothed spectral-difference measure which characterizes codec distortion of vocal tract information. An overall cepstral distance measure is formed by averaging CD(k) over all N frames (excluding pauses) in the record:

$$CD = \frac{1}{N} \sum_{k=1}^{N} CD(k) \qquad (3)$$

The average CD value can be mapped into estimated quality (e.g., Mean Opinion Score or Diagnostic Acceptability Measure (DAM)) using regression techniques.

## 3.0 MEL-CEPSTRAL DISTANCE

Cepstrum-based speech parameters are attractive for use in objective speech-quality measures because of the inherent separation of vocal tract and excitation speech components. However, the simple CD measure described above does not account for the well-known variation of the ear's critical bandwidth as a function of frequency. This effect can be included by replacing $C_X(i,k)$ and $C_Y(i,k)$ with coefficients derived from the mel-cepstrum. The mel-cepstrum has been used successfully for speech recognition applications, and has been shown to be a compact representation of perceptually relevant speech characteristics of the short-term speech spectrum [6].

Mel-cepstral coefficients are computed as follows. First the input and output spectra are computed using the DFT. The spectra are then frequency warped and integrated over 20 critical bands using triangular-shaped filters. These filters are approximately 200 Hz wide for frequencies below 1kHz, and range up to 1 kHz wide at 4 kHz. The filters are shown in Figure 1.

A cosine transform of the real logarithm of the output filter powers results in the mel-cepstral coefficients, $MC_x(i,k)$ and $MC_y(i,k)$, which are substituted for $C_x(i,k)$ and $C_y(i,k)$ in equation (1). The formulation for $MC_x(i,k)$ is presented in (4); $MC_y(i,k)$ is determined analogously.

$$MC_x(i,k) = \sum_{n=1}^{20} X_{k,n} \cos\left[i(n-\tfrac{1}{2})\tfrac{\pi}{20}\right], \qquad (4)$$

where i takes values in the range 1,2,...,M, and $X_{k,n}$ is the log-power output of the n-th triangular filter output:

$$X_{k,n} = \log_{10}\left\{\sum_m |X(k,m)|^2 \bullet w_n(m)\right\}. \qquad (5)$$

In this equation, X(k,m) is the Fourier transform of the k-th input speech frame with frequency index m, and $w_n(m)$ is the n-th critical band filter as depicted in Figure 1.

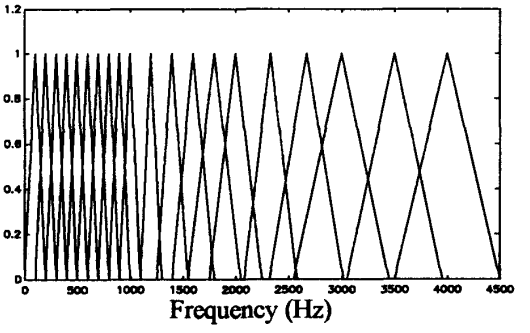


**Figure 1. Mel-scale critical band filter weights.**

## 4.0 ANALYSIS

The CD and MCD objective parameters were applied to a speech dataset consisting of 4800 bps CELP coded speech and subjectively rated using DAM scores. The test data included the following types of simulated channel conditions: "quiet channel", added Gaussian noise at 5 dB signal-to-noise ratio (SNR), 7.5 ms burst errors, Rayleigh fading (60 Hz, 15 dB), and Rician fading (3 path, 12 dB). In addition, both hard and soft Golay error correction are represented, as well as the uncoded speech case. Six sentences from a male talker were processed for the 5 dB SNR and Rayleigh fading cases, while only 2 sentences were available for the 7.5 ms burst error and Rician fading cases.

The objective assessment included the following processing steps: (1) the time delay between the input and output speech records was estimated and removed. Although delay is typically estimated by maximizing the cross-correlation function, this

method does not perform well for non-waveform coded speech. For this paper, delay was estimated by minimizing the input-output cepstral distance.
(2) Pauses in the speech record were removed by eliminating frames with RMS amplitude falling below a manually-selected threshold. This step is required because the effect of pause noise on perceived quality is little understood and quite unpredictable. For example, in some cases, noise added during pauses can actually improve subjective quality. This is why certain network devices (such as echo cancellers or digital circuit multiplication equipment) purposely inject low level "comfort" noise into quiet speech frames.
(3) The CD and MCD values were computed for all speech records and averaged to produce a single objective value for each impairment condition.

Figure 2 shows the results of this procedure. Here, CD and MCD objective values are graphed verses subjective DAM scores. As expected, quality (in terms of DAM score) decreases as distortion (in terms of CD or MCD) increases.

Regression analysis was next used to find the best linear fit between the objective values and subjective DAM scores provided by a human listener panel. Regression results were also used to map the objective measure into equivalent predicted DAM quality scores. The predicted DAM scores are shown in Figure 3 as a function of subjective DAM scores. The squared-correlation coefficient, $\rho^2$, between the objective parameter values (CD and MCD) and the DAM scores was used to measure algorithm performance.

## 5.0 DISCUSSION AND SUMMARY

The results show that the MCD correlation of $\rho^2$=0.667 was significantly better than $\rho^2$=.538 for the CD technique. This suggests that the MCD objective parameter can be used for more accurate and reliable speech quality estimates than the CD. However, as Figure 3 illustrates, there is still significant variability in the predicted quality scores. In fact, consistent squared correlation values in the range of 0.90 to 0.95 would be required to approximate the accuracy of human listener panels [7]. Clearly, the CD and MCD are appropriate only for applications where rough quality estimates are acceptable.

In analyzing these results, it should be kept in mind that the CD and MCD measures may perform much better on other types of speech impairments than were evaluated in this experiment. This test primarily included bit-error type distortions, which are possibly the most difficult condition for objective quality estimation. When applied to other types of impairments such as speech-correlated noise or non-waveform coder distortion, better performance can be expected.

These experimental results also provide insight into the importance of using critical band filters for objective speech quality estimation. In this experiment, two cepstral distance measures were studied. The first was a simple Euclidean distance, while the second combined mel-scale frequency warping and critical band filtering. Although this experiment is not sufficient to determine the relative importance of warping and filtering, their combined effect was clearly to improve the quality prediction. We can speculate that the effect of the triangular filter convolved with the spectrum is to smooth out unperceived detail in the spectrum and making the MCD insensitive to spectral distortion on this level. In contrast, the CD measure includes spectral detail that is not perceived and which therefore contributes to erroneous variability in the objective scores.

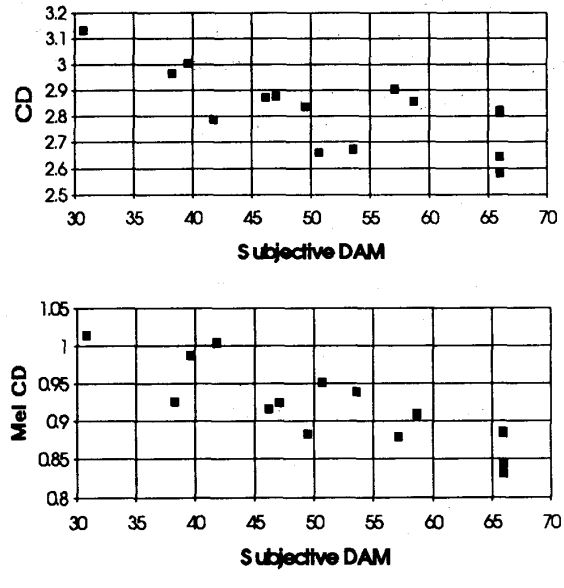


**Figure 2. Cepstral distance and Mel-cepstral distance vs subjective DAM scores.**

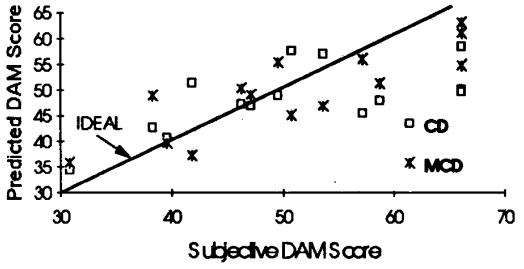


**Figure 3. Predicted vs subjective DAM scores using linear regression of CD and MCD objective measures.**

In summary, this paper describes the benefits of including mel-warped critical band filters in the cepstral distance objective speech quality technique. The improvement in correlation with subjective DAM scores indicates that critical band filtering (and frequency warping) allows better modeling of perceived quality. Introduction of other perceptual-based modifications such as loudness or loudness level (as in [5]) may allow an understanding of the relative importance of these factors to subjective quality. The ultimate goal is to develop a subjective quality model which successfully treats a wide range of impairments and conditions that often foil current objective assessment approaches. These conditions include bit errors (including burst errors), pause effects, language, application, and coder type dependencies.

## REFERENCES

[1] R. Kubichek, "Standards and technology issues in objective voice quality assessment," Digital Signal Processing, Academic Press, vol. 1, no. 2, pp38-54, April 1991.

[2] K. Itoh, N. Kitawaki, and K. Kakehi, "Objective quality measures for speech waveform coding systems," Review Elec. Commun. Lab, vol. 32, no. 2, pp 220-228, Japan NTT, 1983.

[3] N. Kitawaki, K. Itoh, M. Honda, and K. Kakehi, "Comparison of objective speech quality measures for voiceband codecs," Proc. ICASSP, pp 1000-1003, Paris 1982.

[4] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective quality evaluation for low-bit-rate speech coding systems," IEEE Journal on Sel. Areas in Communications, vol. 6, no. 2, pp 242-248, Feb., 1988.

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

[6] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-28, NO. 4, August 1980

[7] R. Kubichek, D. Atkinson, and A. Webster, "Advances in Objective Voice Quality Assessment," Globecom 91, Phoenix, AZ, Dec. 1991.