ICSV13 - Vienna
The Thirteenth International Congress on Sound and Vibration
Vienna, Austria, July 2-6, 2006

OBJECTIVE SPEECH QUALITY MEASURE USING THE EMPIRICAL MODE DECOMPOSITION METHOD

Kamel Yahiaoui and Abdelaziz Ouamri

Signals and Images Laboratory, Department of Electronic, Sciences And Technology
University of Oran, Algeria
kamel_yh@yahoo.fr, Ouamri@ustomail.univ-usto.dz

Abstract

The evaluation of speech codecs can only be performed subjectively by listening tests. For practical reason, these listening tests cannot be performed in real time or in repetitive way.

In the present paper, we describe a new objective measure to estimate the subjective quality obtained in the listening tests of speech codecs used in telephonic band (300-3400 Hz). This objective measure uses the recent time-frequency analysis called the Empirical Mode Decomposition (EMD), and estimates the amount of perceptible distortion by means of perceptuel models that we have adjust to the refined nature of Hilbert spectrum obtained by the EMD method.

The obtained results indicate a satisfactory performance of our measure compared to the recent objective measures, with possibilities of improvement by integrating other processes made possible by the use of the EMD analysis method.

1. INTRODUCTION

Several methods to evaluate the objective speech quality are available in the literature. The main purpose of these objective measures is to produce a more convenient alternative to the subjective tests that are expensive and time-consuming.

Among these objective measures, one finds the traditional distortions as the SNR [1], the segmental SNR [1], the Itakura distance [2], and the cepstral distance [3]. These measures of quality applied to low bit rate speech codecs produce an estimation that is not well correlated with the human perception.

That s why, a more elaborated measures based on the mathematical models of the human auditory system have been achieved. These measures try to simulate the

human ear in order to extract only the audible distortions. Among these measures one mention: the MBSD (Modified Bark Spectral Distance) [4], the EMBSD (Enhanced Modified Bark Spectral Distortion) [5], the PSQM (Perceptual Speech Quality Measure) [6], and the PESQ (Perceptual Evaluation of Speech Quality) [7] that is the last standard of the objective speech quality measure of the ITU. These objective measures called perceptual measures of the quality present a better performance than the traditional measures.

In this paper, we describe a new perceptual objective speech quality measure, based on the algorithm of EMD, using the Johnston s perceptual model [8] and the PEAQ s perceptual model [9] with some modifications. The performance of our measure has been evaluated with the mean opinion score (MOS) estimated by the PESQ measure, and compared to the performance of the MBSD and the EMBSD.

2. DESCRIPTION OF THE PROPOSED PERCEPTUAL OBJECTIVE MEASURE

The proposed measure compares the time-frequency representation of the distortion due to the low bit rate codec, to that of the original speech signal in a perceptual domain, and determines the perceived distortion. The obtained distortion is then mapped to the MOS scale using a regression curve.

The structure of the proposed Perceptual measure is detailed below.

2.1. IRS filtering

The degraded and the original speech signals should first be filtered using receiving characteristics appropriate for a telephone handset. The ITU-T recommends the use of the modified IRS receiving characteristics defined in Annex D/P.830 [10] as receiving frequency characteristics of a telephone handset.

2.2. Speech activity detect (VAD)

Because periods of silences (either background noise in the real conditions) carry less perceptual information than periods of speech, the estimation of the distortion is only done in segments where the speech is active. The effect of the noise in segments of silence is assumed to be negligible.

The discrimination between segments of speech and segments of silence is achieved by means of a Vocal Activity Detection (VAD) process [11]. The VAD process can be seen as a decision problem, in which detector decides between silence or active speech. Only pauses larger than 200 ms are considered as being periods of silence

2.3. Time alignment

The segments of speech active of the degraded signal and those of the original signal must be aligned before evaluating the distortion, This alignment is determined

by using the delays that give the maximal correlation between every segment of speech active of the original signal and the one correspondent of the degraded signal.

2.4. Calculation of the distortion

Because of the non-linearity of the time-frequency analysis method EMD [12] used in the proposed objective measure, the direct comparison by subtraction between the two time-frequency representations of the two signals (original and degraded) deals a false estimation of the true distortion. To circumvent this problem, the distortion is first calculated in the temporal domain then translated in the time-frequency representation.

2.5. Time-frequency mapping

The passage from the temporal domain to the time-frequency domain is effectuated by the EMD analyses. The EMD decomposes the speech signal $x(n)$, considering its local oscillations and its temporal envelope, in a sum of Frequency-Amplitude Modulated components.

$$x(n)=\sum_{k} \gamma_k(n) \,.\, \cos(\varphi_k(n)) \tag{1}$$

It works like an adaptatif filter bank that determines intrinsically the different instantaneous cuts-offs frequencies. This decomposition simulates remarkably the filter bank behaviour of the auditory system, and it replaces efficaciously the gammatone filter bank used in some perceptual models.

Theses AM-FM components are then analyzed using the Hilbert transform to determine the temporal envelopes γ_k and the instantaneous phases φ_k. The result is a time-frequency representation of the instantaneous energy density called Hilbert spectrum, having a good resolution both in time and in frequency.

This analysis method EMD based on the temporal envelope and on the locals oscillations of the speech signal is more adequate to simulate the spectral analysis done in the auditory system, than the transformations based on an mathematical expansion on a orthogonal basis of a fixed amplitudes and a fixed frequencies.

2.6. Frequency warping

A frequency warping allowing passing from the hertz scale to the critical band scale (bark) is achieved, in order to take in consideration frequency sensitivity of the ear.

This is effectuated by grouping the instantaneous energy densities of the Hilbert spectrum in critical bands.

2.7. Premasking

The Premasking is considered as the inertia of the ear to perceive a sound, or as

the weak temporal resolution of the ear to localize temporally a sound. That is why, the weak sounds emitted rightly before more powerful sounds are masked. The Premasking effect lasts about 20 ms, but its most meaningful phase is only 4 ms. The effect of sounds in this interval is determined by the mean of their sounds level rather than by their instantaneous energy densities.

To simulate this effect of premasking, the Hilbert spectrum is quantified every 4 ms, therefore, one will have a representation in cells of 4 ms width and 1 bark of height. The instantaneous energy densities present in every cell are averaged, the result is then an energy density by cell.

2.8. Frequency spreading

In order to reproduce the effect of frequency spreading that occurs in the ear, the spreading function is used.

The used spreading function is the one of the Johnston model [8]:

$$C(i,n) = \sum_{j=1}^{18} S(i,j)B(j,n) \qquad j = 1:18 \tag{2}$$

where $C(i, n)$ represent the spread energy density in the critical band i and in the index of cells n, and $S(i,j)$ is a spreading matrix defined by :

$$\begin{aligned} S(i,j) &= 15.81 + 7.5(i-j+0.474) \\ &\quad - 17.5\sqrt{1+(i-j+0.474)^2} \\ \text{for } &|j-i| \le 25 \end{aligned} \tag{3}$$

where i is the frequency in bark of the masked signal, and j is the frequency in bark of the masker signal.

2.9. Postmasking

The postmasking effect corresponds to the diminution of the masker effect in the ear. The temporal spreading function used is a modified version of the one of the PEAQ postmasking model. This modification of the PEAQ postmasking model is due to the fact that it presents a too much pronounced effect of postmasking. The reduction of the postmasking effect of the PEAQ postmasking model is achieved by making the sound level decreasing varying with time.

The postmasking is achieved via the following relation:

$$C_T(i,n) = a(i).C(i,n_0(i)) + (1-a(i)).C(i,n) \tag{4}$$

where $C_T(i,n)$ is the temporal spread of the energy density with $C_T(i,0) = 0$ as initial condition, $n_0(i)$ is the index of the last cell before the sound level decreasing, and $a(i)$ is a parameter that depends in the time constant of the filter :

$$a(i) = k \,.\, \exp\left(\frac{-t}{\tau(i)}\right) \tag{5}$$

$$\text{with} \quad t = \frac{1}{F'_{cel}} = \frac{(n - n_0(i)) \,.\, L_{cel}}{F_e} \tag{6}$$

where F_{cel} is the cells rate by seconds, F_{cel} is the length in samples of a cell, k is a parameter obtained experimentally (k = 1.2), and F_e is the sampling rate.

The final spread energy density in the cell defined by the pair (i,n) is:

$$E(i, n) = \max(\, C(i, n) \,, C_T(i, n) \,) \tag{7}$$

2.10. Intensity warping

In order to simulate the sensitivity of the ear to the loudness level of sounds, the finale spread energy densities are transformed in the loudness domain, by means of Equal-Loudness Contours [13]. The result is a time-frequency representation that reflects the auditory image of a sound signal in the ear.

2.11. Calculation of the noise masking threshold

The noise masking threshold is determined from the auditory image of the original signal in the perceptual domain. This threshold differs depending on whether one has a noise masking a tone or a tone masking a noise. According to the Johnston model, the masking threshold is below the loudness level of a masker by an offset $O(i)$ given by :

$$O(i) = \alpha*(14.5 + i) + (1 - \alpha)*5.5 \tag{8}$$

where α is the tonality factor that depends on the masker structure to be tonal or noise-like. We have modified this factor by a parameter k = 0.4, to take in count the refined structure of the Hilbert spectrum:

$$\alpha = \min\left(\frac{k \,.\, SFM_{dB}}{SFM_{dB\max}}, 1\right) \tag{9}$$

where SFM_{dB} is the spectral flatness measure of the final spread energy density, and SFM_{dBmax} is set to 60 dB for the entirely tone-like signal.

The total noise masking threshold is determined by:

$$NMT(i,n) = \max\,(E_{sone}(i,n) - O(i),\, AT(i)) \tag{10}$$

where $AT(i)$ is the absolute threshold. Then, $NMT(i,n)$ is a time-frequency representation of the noise masking threshold of the masker. NMT represents the sensation surface of the original signal. All noise being under the level of the sensation surface of the original signal is considered as inaudible.

2.12. Disturbance processing

Finally, the auditory image of the distortion is compared with the sensation surface of the original signal. All zones where the auditory image of the distortion is lower than the sensation surface of the original signal are considered as a masked distortions and will be therefore imperceptible to the ear.

The total audible distortion is then calculated as the mean of the audible distortions of all speech active segments.

3. RESULTS

The performance of the proposed objective speech quality measure has been evaluated with a database witch consists of 40 sentences coded by eight codecs (G721, G726, GSM 6.10, CELP, Speex, MPEG1-Layer3, SBC, DSP Group TrueSpeech 8Kbits). with different bit rates from 2 to 32 kpbs.

The Lack of real subjective scores forced us to evaluate our objective measure using the PESQ (ITU standard) as MOS scores. We consider this approach as an alternative way to examine the validation of the obtained results.

The obtained performance is illustrated in the figure 1. The correlation coefficient obtained is 91.13%.

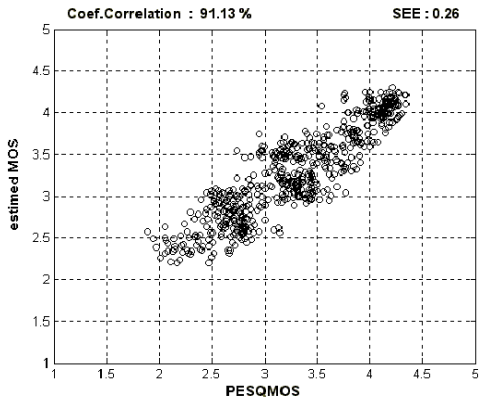

Fig. 1. *Obtained performance with the proposed method judged with the PESQMOS scores.*

The performance of the proposed objective speech quality measure was evaluated in terms of correlation coefficient between the obtained results and the PESQMOS scores calculated by the PESQ algorithm. For comparison, the figure 2 and the figure 3 represent the performance obtained with the two measures MBSD and EMBSD respectively.

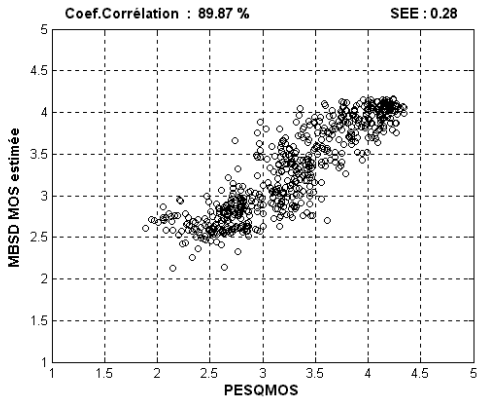

Fig. 2. *Obtained performance with the MBSD measure judged with the PESQMOS*

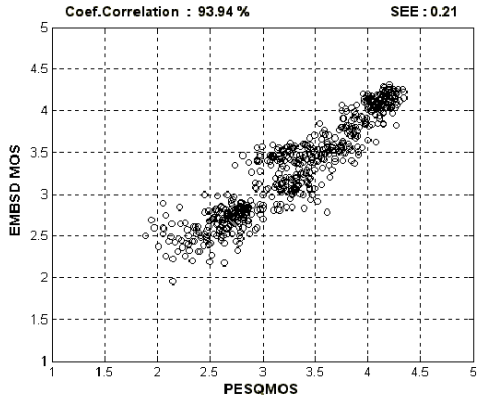

Fig. 3. *Obtained performance with the EMBSD measure judged with the PESQMOS*

Table.1. *Performance of the proposed perceptual objective measure in terms of correlation coefficient with PESQMOS scores. The performances of EMBSD and MBSD measures are shown for comparison*

	Proposed	MBSD	EMBSD
Correlation / PESQMOS	0.9113	0.8987	0.93.94

Table 1 shows the performance of the proposed perceptual objective measure compared to that of the EMBSD and the MBSD measures. It indicates a good performance of the proposed measure compared to that obtained by the MBSD measure, and a quite comparable performance to that obtained by the EMBSD measure. The obtained performance is a quite promising result, considering that other improvements can be carried on the proposed objective measure due to the remarkable temporal precision offered by the Hilbert spectrum, things that will improve furthermore its performance.

4. CONCLUSION

In this paper, we have proposed a new objective speech quality measure. The proposed method uses the recent time-frequency analysis EMD method based on the temporal envelope and the local oscillations of speech signal in contrast to the short-time Fourier representations typically used in conventional objective models. We found that the EMD method reflects more the characteristics of the human auditory system and provides a very refined time-frequency representation of the speech signal with few harmonics and a remarkable precision. The integration of auditory models has therefore been very easy with a possibility of integration of new process that would permit to benefit on the best of the time-frequency precision offered by the analysis method EMD, thing that was not possible with Fourier transform.

5. REFERENCES

[1] S. R. Quackenbush, T. P. Barnwell III, and M. A. Clements, Objective Measures of Speech Quality, Prentice Hall, Englewood Cliffs, 1988.

[2] F. Itakura, Line spectrum representation of linear predictive coefficients of speech signals, Journal Acoustical Society America, vol. 57, p. S35, Apr. 1975. abstract.

[3] A H. Gray and J. D. Markel, Distance measures for speech processing, IEEE Trans. Acoust., Speech and Signal Processing, vol. ASSP-24, pp. 380-391, Oct. 1976.

[4] W. Yang, M. Dixon, et R. Yantorno, A Modified Bark Spectral Distortion Measure Which Uses Noise Masking Threshold, IEEE Speech Coding Workshop, pp. 55-56, Pocono Manor, 1997.

[5] Wonho Yang, Enhanced Modified Bark Spectral Distortion (EMBSD) : an objective speech quality measure based on audible distortion and cognition model . PhD thesis, Temple University, May 1999.

[6] Recommandation UIT-T P.861, Mesure Objective de la Qualité des Codecs Vocaux Fonctionnant en Bande Téléphonique (300-3400 Hz) (PSQM) . Février 1998.

[7] Recommandation UIT-T P.862, Evaluation de la Qualité Vocale Perçue (PESQ) . Février 2001.

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

[9] Recommandation UIT-R BS.1387, 'Méthode de Mesure Objective de la Qualité du Son Perçu (PEAQ) . 2001.

[10] Recommandation UIT-T P.830, Evaluation Subjective de la Qualité des Codecs Numériques à Bande Téléphonique et à Large Bande . Février 1996.

[11] Recommandation UIT-T P56, Mesure Objective du Niveau Vocal Actif . Mars 1993.

[12] N.E. Huang, Z. Shen, S.R. Long, M.L. Wu, H.H. Shih, Q. Zheng, N.C. Yen, C.C. Tung et H.H. Liu, The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis . Proc. Roy. Soc. London A, Vol. 454, pages 903 995, 1998.

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