Quality enhancement of CELP coded speech by using an MFCC based Gaussian Mixture Model

D.G.Raza and C.F.Chan

Department of Computer Engineering & IT
City University of Hong Kong, Hong Kong
raza.dar@student.cityu.edu.hk
itcfchan@cityu.edu.hk

Abstract

At low bit rates CELP coders present certain artifacts generally known as hoarse and muffing characteristics. An enhancement system is developed to lessen the effects of these artifacts in CELP coded speech. In enhancement system, the high frequency components (4kHz-8kHz) are reinserted to reduce the muffing characteristics. This is achieved by using an MFCC based Gaussian Mixture Model. The hoarse characteristics are reduced by re-synthesizing the CELP reproduced speech with harmonic plus noise model. The pair-wise listening experiment results show that the re-synthesized wideband speech is preferred over the CELP coded speech. The enhanced speech is affirmed to be pleasant to listen and exhibits the naturalness of the original wideband speech.

1. Introduction

The two main artifacts that degrade the quality of CELP [1] (Code Excited Linear Prediction) coded speech at low bit rates are hoarse and muffing characteristics. A careful study shows that the cause of the hoarse (also known as coding noise) artifacts is primarily due to the parameter quantization and codebook excitation signal that is used to excite the synthesis filter in CELP coding of speech. Muffing characteristics are due to lack of high frequency components in the narrowband (0-4kHz) speech signal and this becomes more prominent after the coding/decoding process of speech signal. The removal of the higher frequency consonants such as /sh/, /s/ and /f/ severely affects the naturalness of speech and reduces the fricative differentiation in speech signal. In this paper we present the enhancement system, which is designed to reduce the effects of these artifacts and thus improve the quality of CELP coded speech. In enhancement system, the CELP coded narrowband speech is re-synthesized with harmonic plus noise synthesizer [2][3] to improve the pitch periodicity of voiced speech. The high frequency components are reinserted by using an MFCC based Gaussian mixture model to make it a wideband speech (0-8kHz). This improves the intelligibility as well as the naturalness of speech. It also introduces the crispy characteristic of wideband speech and thus lessens the listening fatigue. As the telecommunication industry is using huge installation based on CELP coders (FS1016 CELP coder, QCELP, LD-CELP, VSELP, CS-ACELP) and the demand for higher quality of speech is increased with the emergence of application such as streaming speech over internet, internet telephony therefore it is required to further improve the quality of CELP coders.

2. Enhancement System

The proposed enhancement system is depicted in figure 1. The narrowband CELP reproduced speech is first analyzed by a harmonic plus noise analyzer and lowband information are extracted which includes fundamental frequency or pitch (ω_{o}), 10LSP (Line Spectrum Pairs), harmonic magnitudes, phases, voiced/unvoiced information for each harmonic and Mel-Frequency Cepstrum Coefficients (MFCC) [5][6]. The lowband feature vector (21 MFCC coefficients) is then fed to a Gaussian mixture model to estimate statistically a wideband spectrum envelope. The predicted wideband spectrum envelope is then sampled at pitch harmonics to obtain the highband information that includes highband harmonic magnitudes and V/UV information of each harmonic. Finally the lowband and estimated highband information are fed to harmonic plus noise synthesizer to synthesize a wideband speech (0-8kHz).

3. GMM Model

The highband information is estimated from the wideband spectrum envelope, which is obtained from an MFCC based Gaussian mixture model [4]. The wideband spectrum envelope in the GMM is preserved in the form of 18-dimenstinal LSP vectors obtained from wideband speech database. A finite mixture of Gaussian densities can be expressed mathematically as given in (3.1) and (3.2)

$$p(x) = \sum_{j=1}^{Q} \pi_j p(x, \theta_j) \qquad 3.1$$

$$\sum_{j=1}^{Q} \pi_j = 1 \qquad 3.2$$

Where Q is the number of normal component densities in a Gaussian mixture model and π_j are the mixing proportions of each components density in a mixture model. The $p(x,\theta_j)$ is an individual component density in a mixture model and is completely defined by the parameter vector θ_j and is given in (3.3)

$$\theta_j = (\mu_j, \Sigma_j) \qquad j = 1, 2 \ldots\ldots Q \qquad 3.3$$

Where μ_j, and Σ_j are the mean vector and covariance matrix of the j^{th} component density of a Gaussian mixture model. The probability of an input vector $\boldsymbol{x}$ for the j^{th} component density can be determined by the equation (3.4).

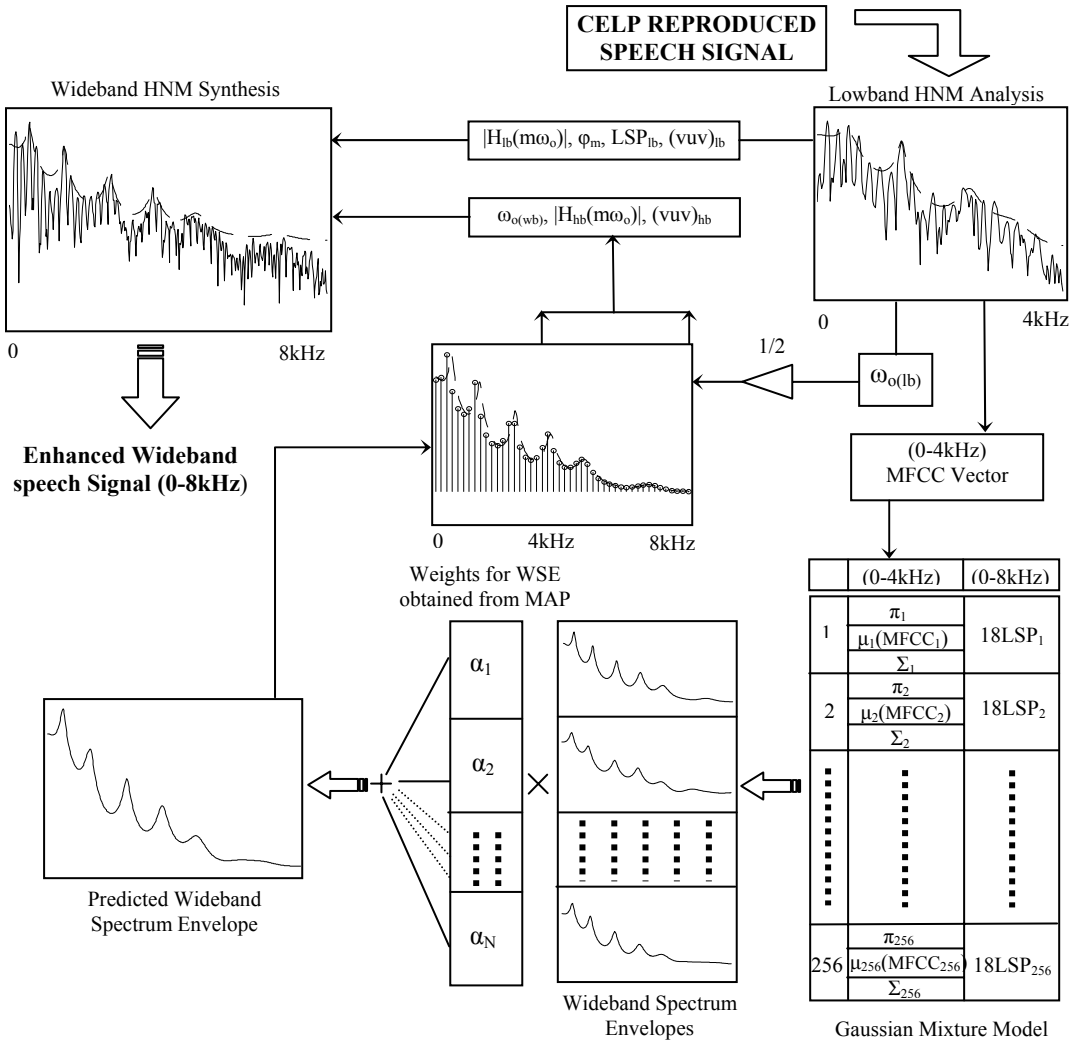

Figure1. Enhancement System

$$p(x,\theta_j)=\frac{\pi_j}{\left(\sqrt{2\pi}\right)^n\left|\Sigma_j\right|^{1/2}}\exp[-\frac{1}{2}(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j)] \qquad 3.4$$

4. Training of GMM

The lowband (0-4kHz) MFCC coefficients are used as a training data for GMM model. There are two main advantages of using MFCC coefficients as training vectors for GMM model. The first, they are completely de-correlated coefficients hence permits the use of a diagonal covariance matrix in a GMM model. Secondly, few number of MFCC coefficients can be used to obtain satisfactory results. This reduces the considerable computation and memory requirements. A phonetically well-balanced wideband (0-8kHz) speech corpus, contributed by many male and female speakers was collected. To include microphone characteristic, the speech database is recorded by using different microphones. A lowband speech corpus is obtained from the wideband speech corpus. This is obtained by lowpass filtering with linear phase FIR filter and decimation by a factor of 2. A bank of 22 triangular filters spaced on Mel-Scale is used to compute the lowband 22-MFCC coefficients. The first MFCC coefficient is dropped being of highly dynamic. For each lowband MFCC coefficients its corresponding wideband information in the form of 18LSP parameters are also obtained from the wideband speech corpus. These parameters (lowband 21-dimensional MFCC vectors and wideband 18-dimensional LSP vectors) are then used to train a Gaussian Mixture Model. The mel-cepstrum distortion was used as classification criteria [6] for lowband MFCC coefficients as given in equation (4.1)

$$M_{cd}=\sqrt{\sum_{i=1}^{C}\left[MFCC_x(i)-MFCC_y(i)\right]^2} \qquad 4.1$$

The LBG algorithm with split initialisation was used to classify training data into required number of clusters. Then sample mean vectors, diagonal covariance matrix and component weights are obtained from each cluster. These parameter values are then used as initial values for EM algorithm [7]. The EM algorithm has two steps; the E-step or

expectation step uses equation (4.2) to classify the input vectors.

$$y_i \in C_k \Leftrightarrow k = \arg\max_j \; [\log \pi_j - \log|\Sigma_j| - (y_i - \mu_j)^T \Sigma_j^{-1} (y_i - \mu_j)] \quad 4.2$$

Where $\boldsymbol{y_i}$ is the input vector and $\boldsymbol{C_k}$ is the k^{th} component density. The M-step updates the parameters according to equations (4.3) and (4.4). Where N_k is the number of training vectors belong to cluster k and N is the total number of spectral vectors in training data.

$$\pi_k = \frac{N_k}{N}, \quad \mu_k = \frac{1}{N_k} \sum_{y_i \in c_k} y_i \quad 4.3$$

$$\Sigma_k = \frac{1}{N_k} \sum_{y_i \in c_k} (y_i - \mu_k)(y_i - \mu_k)^T \quad 4.4$$

5. Wideband Spectrum Envelope Prediction

The GMM model is used as a tool for recovering statistically the Wideband Spectrum Envelop (WSE). Let x be the lowband MFCC coefficient vector obtained from lowband analysis of speech then the maximum a posterior probability (MAP) for this vector can be calculated by the equation (5.1)

$$p_j(x) = p(\theta_j \mid x) = \frac{\frac{\pi_j}{(2\pi)^{p/2}|\Sigma_j|^{1/2}} \exp[-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)]}{\sum_{i=1}^{Q} \frac{\pi_i}{(2\pi)^{p/2}|\Sigma_i|^{1/2}} \exp[-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)]} \quad (5.1)$$

Where $p_j(x) = p(\theta_j \mid x)$ is the posterior probability of the lowband input MFCC coefficient vector x for the j[th] class. Based on the posterior probabilities, the first N classes having the highest MAP are then selected for interpolation. The weighing factors α_l are obtained from the MAP of N classes by using equation (5.2). The α_l are used to weight the mean vectors of the component densities. In equation (5.3), N is the number of the component densities used for interpolation and in this case it was set equal to 2. The value of M can be used with in the range of $1 \le N \le 4$. Increasing the number of the component densities for interpolation beyond this range reduces the dynamic variation of predicted wideband spectrum envelope.

$$\alpha_l = \frac{[1 - p_l(x)]}{\sum_{i=1}^{N} [1 - p_i(x)]}, 1 \le l \le N \quad 5.2$$

$$E(y) = \sum_{l=1}^{N} \alpha_l \times \mu_l \quad , \sum_{l=1}^{N} \alpha_l = 1 \quad 5.3$$

Here μ_l are the mean vectors (wideband spectrum envelopes) obtained from plotting the18LSP parameters to 128-points. The predicted WSE $E(y)$ can be seen as composed of the weighted mean vectors of the component densities of GMM model that were used for interpolation. Fig.2 shows the original and predicted WSE. The performances of trained GMM model is then evaluated in predicting the wideband spectrum envelopes. The average spectral distortion is calculated between the original wideband spectrum envelopes and the predicted wideband spectrum envelopes by using the equation (5.4). Table 1

GMM Model	**Average Highband (4kHz-8kHz) Spectral Distortion SD_{AVE} dB**	
Classes	Inside the Training Data	Outside the Training Data
256	3.46	3.57

Table 1. Performance of GMM model

shows the performance evaluation of an MFCC based GMM model.

$$SD_{ave} = \left[\frac{1}{N} \sum_{n=1}^{N} \frac{1}{\pi} \int_{\frac{1}{\pi}}^{\frac{1}{\pi}} \left(10\log|H_n(\omega)| - 10\log|\widehat{H}_n(\omega)| \right)^2 d\omega \right]^{1/2} \quad 5.4$$

The average highband spectral distortion is slightly higher this is due to the fact that we are using only 256 unique classes in MFCC based GMM.

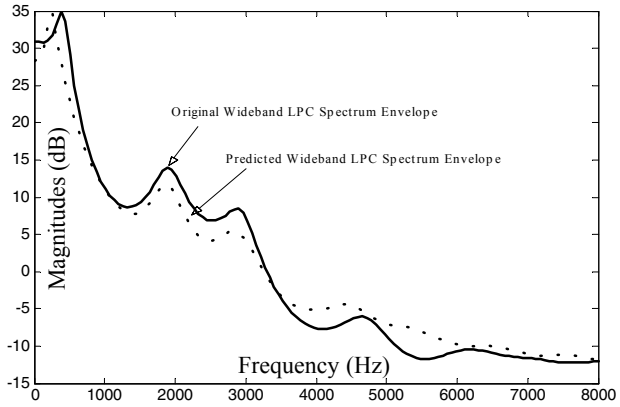

Figure 2. Original and predicted WSE

6. Highband Parameter Estimation

The highband parameters are estimated from the predicted wideband spectrum envelope. The parameters that we need to re-synthesize a wideband speech from its corresponding narrowband version are consists of *wideband pitch*, *highband harmonic magnitudes*, *wideband spectrum envelope* (original lowband spectrum envelope linked smoothly with estimated highband spectrum envelope) and *wideband gain*. The wideband pitch is used to sample the wideband spectrum envelope at pitch harmonics to get the highband harmonic magnitudes. The wideband pitch is obtained from the estimated lowband pitch during the lowband harmonic plus noise analysis of CELP reproduced speech. The wideband pitch is obtained from the equation 6.2.

$$\omega_{o_{lowband}} = \frac{2 \times \pi \times f_o}{f_s} \quad 6.1$$

$$\omega_{o_{wideband}} = \frac{\omega_{o_{lowband}}}{2} \quad 6.2$$

The wideband pitch is then used to sample the estimated wideband spectrum envelope at pitch harmonics to obtain the highband harmonic magnitudes. The highband harmonic phases are simply ignored being of little importance.

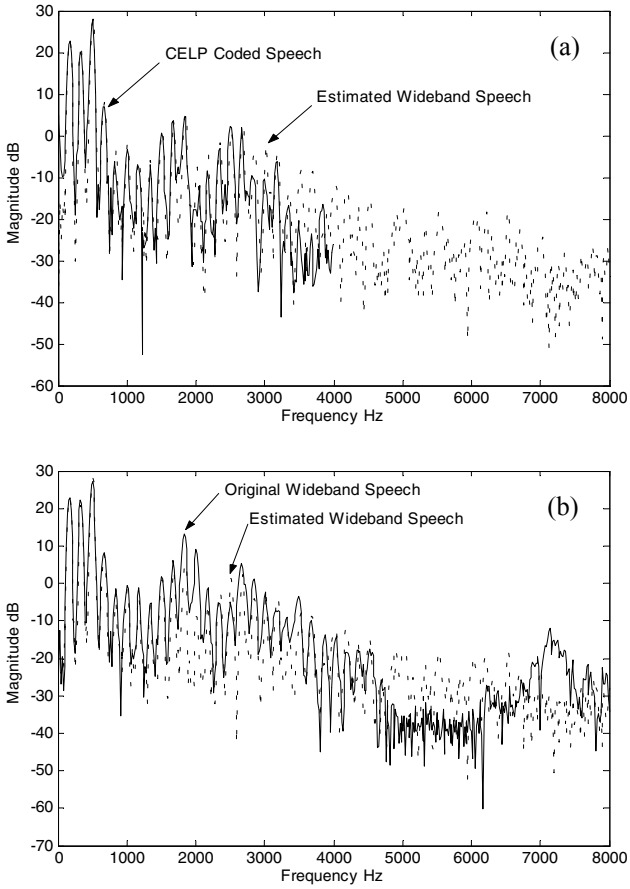

Figure 3. Magnitude spectrums (a) CELP coded and estimated speech (b) Original and estimated speech

7. Wideband Speech Re-synthesis

The wideband speech is synthesized by selecting the frame length double to that of the lowband analysis frame length. As in harmonic analysis each frame is classified into a number of voiced and unvoiced harmonics, therefore the harmonics declared as voiced during the analysis are generated from the sinusoidal oscillators. Frequency tracks between the adjacent frames are also determined to make the synthesis speech smooth and continuous. The unvoiced spectrum is obtained by filtering the Gaussian white noise with unit variance from the synthesis filter, whose coefficients are extracted from autocorrelation data, which is obtained from a weighted LPC spectrum. Fig.3 shows the magnitude spectrums of CELP reproduced speech, estimated wideband speech and original wideband speech signal. The pair-wise listening tests also performed on the estimated wideband speech and CELP coded narrowband speech. Non-expert listeners were invited to participate the experiments. The synthesized wideband speech was played back at 16 kHz while the narrowband CELP coded speech was played back at 8kHz. The ITU-R7 point comparative scale for grading was used during the pair-wise listening tests. The results of the pair-wise listening experiment are shown in fig.4 with mean and 95% confidence intervals.

8. Conclusion

An enhancement system is proposed to improve the quality of narrowband CELP coded speech via lowband harmonic plus noise analysis and wideband extension by using an MFCC based Gaussian Mixture Model. The quality of CELP coded speech after the harmonic analysis and wideband extension has been improved significantly. The estimated wideband speech is preferred when compared to the CELP coded speech and pleasant to listen however it presents some musical sounds the on going research is targeted to reduce these effects.

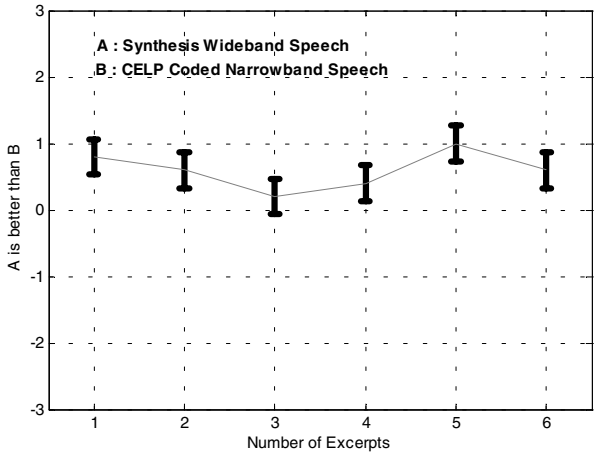

Figure 4 Pair-wise listening experiment result

9. References

[1] Schoreder, M.R. and Atal, B.S.: 'Code excited linear prediction (CELP): high-quality speech at Very low bit rates', ICASSP, pp. 937– 940,1985

[2] Griffin, D.W., and Lim, J.S.: 'Multi-band excitation vocoder', IEEE Trans. ASSP, Vol. ASSP-36, No.8, pp. 1223-1235, August 1988.

[3] R. McAulay and T. Quatieri "Speech analysis/synthesis based on sinusoidal representation" IEEE Transaction on Accoustic Speech and Signal Processing, vol.34, pp.744-754, August 1986.

[4] P. Kun-Youl and K. Hyung Soon, "Narrowband To Wideband Conversion of Speech Using GMM based Transformation", ICASSP 2000, Vol.III, PP 1843-1846.

[5] Davis, S. B., Mermelstein, P., "Comparison of Parametric Representations for Monosyllabic Work Recognition in Continuously Spoken Sentences " IEEE Transaction on Acoustics, speech and Signal Processing, vol. ASSP-28, no.4 pp: 357-366, August 1980.

[6] R. F. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment" in Proc. IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, 1993, pp.125-128.

[7] Blimes. J., "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models" International Computer Science Institute Tech. Rep. ICSI TR-97-021; 1997.