ENHANCING QUALITY OF CELP CODED SPEECH VIA WIDEBAND EXTENSION BY USING VOICING GMM INTERPOLATION AND HNM RE-SYNTHESIS

Dar Ghulam Raza and Cheung-Fat Chan

Department of Electronic Engineering, City University of Hong Kong
83, Tat Chee Avenue Kowloon Hong Kong
raza.dar@plink.cityu.edu.hk, eefchan@cityu.edu.hk

ABSTRACT

This paper presents a procedure to improve the quality of narrowband (0-4khz) CELP coded speech. The procedure is based to refine the pitch periodicity and reinsert the high frequency components (4-8khz) in the narrowband CELP decoded speech. The narrowband CELP decoded speech is first analyzed with Harmonic+Noise analyzer and Lowband information are extracted. By exploiting the Lowband spectrum envelope and V/UV information, the highband (4-8khz) spectrum envelope is recovered statistically by using a voiced/unvoiced gaussian mixture model with interpolation. Lowband information along with the estimated highband information is then fed to the Harmonic+Noise synthesizer to re-synthesize a wideband speech. The objective and subjective tests are performed to evaluate the quality of the re-synthesis wideband (0-8khz) speech. The results of the above experiments show that the re-synthesis wideband speech is pleasant to listen with crispy characteristics and preferred over the CELP coded speech.

1. INTRODUCTION

The communication industry is using extensively CELP (Code Excited Linear Predictive) coders that can code or compress a speech signal up to 4.8kbs [1]. But beside its ability to operate at such a low bit rates, CELP coders present somewhat degraded quality of speech. The two distinct artifacts present in CELP coded speech are known as **hoarse** and **muffing** characteristics. The muffing characteristics are due to the lack of the high frequency component in the narrowband speech signal. CELP coders were originally designed to operate with the existing telephone networks with the bandwidth limited from 0.3-3.4khz. Reinserting the high frequency component (4-8khz) in CELP coded speech reduces the muffing characteristics. The re-generation of the high frequency component introduces natural characteristics in a narrowband speech signal. This also leads to better fricative differentiation and thus higher intelligibility. The hoarse characteristics are inherent in CELP coders due to the stochastic excitation signal, which is selected from a codebook in which each codevector is generated from a Gaussian random numbers generator with unit variance. Refining the pitch periodicity of CELP coded speech reduces the hoarse characteristic. The CELP decoded speech is then re-synthesized to refine the pitch periodicity with Harmonic+Noise coder, which is very well known to produce high quality of synthesis speech [2]. As CELP coders are being used widely in industry i.e. FS 1016 coder, QCELP coder of North American Cellular (CDMA) IS96. It is therefore required to further improve the quality of CELP coders.

In section 2, an overview of the enhancement system is presented. Section 3, describes the lowband analysis of CELP decoded speech with Harmonic+Noise analysis. In Section 4, a strategy is discussed to estimate the wideband spectrum envelope and highband information by using a voiced/unvoiced gaussian mixture model. Section 5 describes the re-synthesis of wideband speech and section 6 presents the experiment details. Finally in section 7, conclusion is presented.

2. ENHANCEMENT SYSTEM

An enhancement system is designed to improve the quality of CELP coded speech as depicted in figure 1. The narrowband CELP decoded speech is first analyzed by a Harmonic+Noise analyzer and lowband information are extracted which includes fundamental frequency or pitch (ω_o), gain, 10LSP (Line Spectrum Pairs) [3], harmonic magnitudes, phases and voiced/unvoiced information for each harmonic. The lowband spectrum envelope is then obtained from the 10LSP parameters and fed to a voiced or unvoiced gaussian mixture model to obtain a wideband spectrum envelope. Sampling the wideband spectrum envelope at pitch harmonics and normalizing the highband magnitudes, the highband information (harmonic magnitudes, V/UV) are obtained. Finally all the information from lowband and highband are fed to the Harmonic+Noise synthesizer, which re-synthesize a wideband speech signal (0-8khz).

3. LOWBAND HARMONIC ANALYSIS

In Harmonic+Noise modeling, the speech signal is assumed to be composed of a deterministic component which is also known as quasi-periodic or voiced part of speech and stochastic component also known as non-periodic or unvoiced part of speech. The voiced part is modeled as sum of harmonics of the pitch or fundamental frequency (ω_o) and is given mathematically in equation (1).

$$s_v(n) = \sum_{l=1}^{L} a_l \cos(nl\,\omega_o + \phi_l) \qquad (1)$$

Where a_l and ϕ_l are the amplitude and phases of the l^{th} harmonic and L is the number of harmonics present in voiced speech. The unvoiced part is modeled as random noise. During the lowband harmonic analysis, the lowband information are estimated which includes the fundamental frequency, gain, harmonic magnitudes, phases, frequencies, 10LSP parameters and voiced/unvoiced information for each harmonic. A slow varying short time spectrum envelope is obtained from the 10LSP parameters.

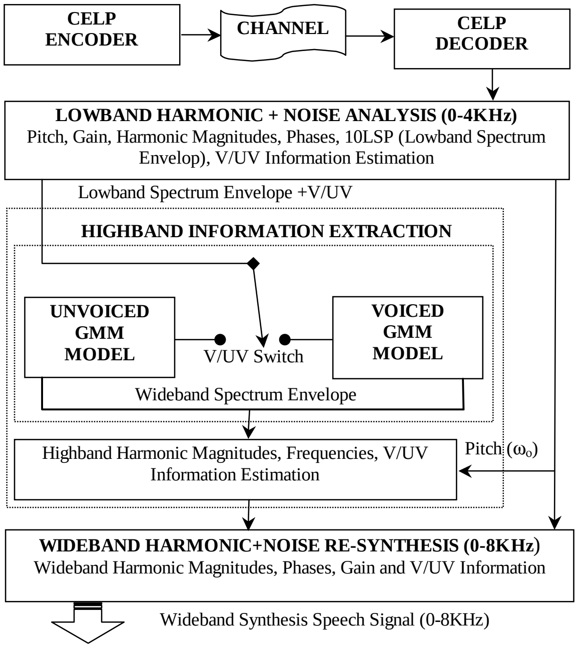
CELP ENCODER
CHANNEL
CELP DECODER
LOWBAND HARMONIC + NOISE ANALYSIS (0-4KHz)
Pitch, Gain, Harmonic Magnitudes, Phases, 10LSP (Lowband Spectrum Envelop), V/UV Information Estimation
Lowband Spectrum Envelope +V/UV
HIGHBAND INFORMATION EXTRACTION
UNVOICED GMM MODEL
V/UV Switch
VOICED GMM MODEL
Wideband Spectrum Envelope
Highband Harmonic Magnitudes, Frequencies, V/UV Information Estimation
Pitch (ωo)
WIDEBAND HARMONIC+NOISE RE-SYNTHESIS (0-8KHz)
Wideband Harmonic Magnitudes, Phases, Gain and V/UV Information
Wideband Synthesis Speech Signal (0-8KHz)

Figure 1. Enhancement System

4. HIGHBAND INFORMATION ESTIMATION

The highband information is estimated from the highband spectrum envelope obtained from a voicing gaussian mixture model with interpolation. We designed two separate gaussian mixture model to represent the voiced and unvoiced distribution of speech spectrum envelopes. During the lowband harmonic analysis, 10LSP parameters are extracted. A lowband spectrum envelope is then obtained from the 10LSP parameters. The lowband spectral vector is then fed to the voicing gaussian mixture model and the MAP (Maximum A posterior Probabilities) for all the classes are obtained. The highband spectral vector is then recovered by interpolating between the three gaussian classes having the highest MAP. The predicted highband spectrum envelope is then used to estimate the highband harmonic magnitudes, and voiced/unvoiced information.

4.1 GMM Model

A finite mixture of Gaussian densities can be expressed mathematically as given in (2.1) and (2.2)

$$p(x) = \sum_{j=1}^{q} \pi_j \, p(x;\theta_j) \qquad (2.1)$$

$$\sum_{j=1}^{q} \pi_j = 1 \qquad (2.2)$$

Where **q** is the number of normal component densities in a gaussian mixture model and π_j are the mixing proportions or component weights of each component 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 (2.3)

$$\theta_j = \left(\pi_j, \mu_j, \Sigma_j\right) \qquad j = 1,2\cdots\cdots q \qquad (2.3)$$

Where π_j, μ_j and Σ_j are the component weight, mean vector and covariance matrix of the j_{th} component density of a Gaussian mixture model. The probability of an input vector x for the j_{th} component density can be determined by the following equation (2.4)

$$p(x;\theta_j) = \frac{1}{\left(\sqrt{2\pi}\right)^n \left|\Sigma_j\right|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)\right] \qquad (2.4)$$

To use the above Gaussian mixture model, first we need to estimate its parameters. There is remarkable variety of estimation methods such as methods of moments, maximum likelihood, minimum chi-square, least squares and Bayesian approaches. We used the maximum likelihood method, which is the most popular method to estimate the finite mixture parameters. Maximum likelihood uses the well-known EM (Expectation and Maximization) algorithm iteratively to estimate the parameters.

4.2 Training of GMM

The mean vectors $\boldsymbol{\mu}$, covariance matrices $\boldsymbol{\Sigma}$ and component weights $\boldsymbol{\pi}$ for each component density are estimated during the training of a Gaussian mixture model. A phonetically well-balanced wideband (0-8khz) speech corpus, contributed by many male and female speakers was collected. 18LSP (Line Spectrum Pairs) parameters were extracted to get the parametric representation of the wideband speech corpus. 18LSP parameters then plotted to 128-point spectral vectors. The wideband spectral vectors were then classified into two groups containing voiced and unvoiced spectral vectors. To train a voiced GMM model with 128 normal component densities, we use LBG algorithm with split initialization to classify 128-dimensional voiced spectral vectors into 128 clusters. Then sample mean vectors (128x1), covariance matrix (128x128) and component weights are obtained from each cluster. These parameter values are then used as initial values for EM algorithm. The EM algorithm has two steps; the E-step or expectation step uses equation (3) to classify the input vectors.

$$y_i \in C_k \Leftrightarrow k = \underset{j}{\mathrm{argmin}}[\log \pi_j - \log|\Sigma_j| + (y_i - \mu_j)^T \Sigma_j^{-1}(y_i - \mu_j)] \qquad (3)$$

$$\pi_k = \frac{N_k}{N}, \qquad \mu_k = \frac{1}{N_k}\sum_{y_i \in C_k} y_i \qquad (4.1)$$

$$\Sigma_k = \frac{1}{N_k}\sum_{y_i \in C_k}(y_i - \mu_k)(y_i - \mu_k)^T \qquad (4.2)$$

Where $\mathbf{y_i}$ is the input vector and $\mathbf{C}_k$ is the k^{th} component density. The M-step updates the parameters according to equations (4.1) and (4.2). Where N_k is the number of spectral vectors belong to cluster *k* and *N* is the total number of spectral vectors in training data.

4.3 Recovering Highband Spectrum Envelope

The GMM model is used as a tool of recovering statistically the highband spectrum envelope. Let ***x*** be the lowband spectral vector and ***y*** be the highband spectral vectors then the joint density of wideband vector ***z*** = (***x***, ***y***)^T is modeled as a mixture of q component of (2 x n)-variate Gaussian function [5].

$$p(z\,|\,\theta_j)=\sum_{j=1}^{q}\frac{\pi_j}{(2\pi)^n\left|\Sigma_j\right|^{1/2}}\exp\left[-\frac{1}{2}(z-\mu_j)^T\,\Sigma_j^{-1}(z-\mu_j)\right] \qquad (5.1)$$

$$\sum_{j=1}^{q}\pi_j=1,\quad \pi_j\geq 0 \qquad (5.2)$$

Then the highband spectrum envelope is obtained by interpolating between the three component densities having the highest MAP for lowband vector ***x*** as given in (6)

$$p(x)=\frac{\dfrac{\pi_j}{(2\pi)^{n/2}\left|\Sigma_j\right|^{1/2}}\exp\left[-\dfrac{1}{2}(x-\mu_j)^T\,\Sigma_j^{-1}(x-\mu_j)\right]}{\displaystyle\sum_{j=1}^{q}\dfrac{\pi_j}{(2\pi)^{n/2}\left|\Sigma_j\right|^{1/2}}\exp\left[-\dfrac{1}{2}(x-\mu_j)^T\,\Sigma_j^{-1}(x-\mu_j)\right]} \qquad (6)$$

Where p(***x***) = p(θ_j | ***x***) is the posterior probability of the input lowband vector x for the jth class. Finally equation (7) is used to obtain the interpolated version of the highband spectrum envelope. Where ***w*** is the number of the component densities used for interpolation and in this case it was set equal to three.

$$p(y\,|\,x)=\frac{1}{w}\sum_{i=1}^{w}\underset{i}{\arg\max}\,[\,p_i(\theta_j\,|\,x)] \qquad (7)$$

After the estimation of the highband spectrum envelope the highband magnitudes at the pitch harmonics are estimated and the highband energies are normalized. The lowband and highband magnitudes along with voiced/unvoiced information are fed to the wideband synthesizer to re-synthesize an improved quality of wideband speech.

The efficiency of our trained voiced and unvoiced Gaussian mixture model is also evaluated with different number of classes and covariance matrices, based on the diagonal covariance matrices for each class and a single global covariance matrix for all classes. We found that a single global covariance matrix gives better performance. Full covariance matrix for each class was not considered due to the higher memory and computation reasons. Voiced Gaussian mixture model was designed with 128 voiced classes while the unvoiced gaussian mixture model was designed with 64 unvoiced classes.

$$SD_{ave}=\left[\frac{1}{N}\sum_{n=1}^{N}\frac{1}{\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\left(10\log\left|H_n(\omega)\right|-10\log\left|H'_n(\omega)\right|\right)^2 d\omega\right]^{1/2} \qquad (8)$$

The average spectral distortion in lowband and highband between the original wideband and predicted wideband spectrum envelope is calculated by using (8) and the results are given in table (1). The average spectral distortion in lowband is a bit higher for both models but it is reasonable in the higher bands. Increasing the number of classes in GMM model can reduce the average spectral distortion in the lowband but this will result in memory and computation overhead too. The performance of the designed GMM models was also evaluated out of the training data and we found that the average distortions and outliers were slightly higher than that of over the original training data.

Voiced GMM 128-Classes	$SD_{AVERAGE}$ dB	Outlier > 3dB
Lowband (0-4khz)	4.52	--
Highband (4-8khz)	1.89	7.23

Unvoiced GMM 64-Classes	$SD_{AVERAGE}$ dB	Outlier > 4dB
Lowband (0-4khz)	8.34	--
Highband (4-8khz)	3.16	9.23

Table 1. Performance Evaluation of GMM

5. WIDEBAND SPEECH SYNTHESIS

Wideband speech is synthesized after recovering the highband information. The size of the synthesis window used was twice of that the analysis window. Suppose if the size of the analysis window was 30ms (240 samples at 8khz sampling frequency), then we used synthesize frame of size 30ms (480 samples at 16khz sampling frequency). As in harmonic analysis each frame is classified into a number of voiced and unvoiced harmonics, so the harmonics declared as voiced during the analysis are generated from the sinusoidal oscillators with quadratic phase interpolation using measured phases. The highband harmonic phases are not considered. 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.

6. EXPERIMENT DETAILS

First of all, a wideband speech corpus (0-8khz) sampled at 16khz with 16 bit/sample, mono and contributed by a large number of speakers was collected. The narrowband speech was obtained from the wideband speech corpus by filtering it with a linear phase FIR filter with cutoff frequency at 4khz and decimated by a factor of two. 18LSP parameters were extracted from the wideband speech and used for training a Gaussian mixture model. During the lowband harmonic analysis voiced/unvoiced, pitch, harmonic magnitudes, phases and spectrum envelopes were extracted from the lowband speech. Narrowband spectrum envelope was then used to predict the highband spectrum envelope by a voicing Gaussian mixture model. Predicted highband spectrum envelope is then used to estimate highband harmonic magnitudes and the voiced/unvoiced information. To determine the capability of our enhancement system to estimate the highband speech, the experiment is first performed on the original lowband speech. Figure 3(a) shows a short time magnitude spectrum of original wideband and CELP coded narrowband speech. The figure 3(b) shows a short time magnitude spectrum of original and estimated wideband speech and it is clear that the recovered highband harmonic structure is very close to the original wideband speech.

7. CONCLUSTION

An enhancement system is proposed to improve the quality of narrowband CELP coded speech via lowband Harmonic+Noise analysis and wideband extension by using a voicing Gaussian Mixture Model. The quality of CELP coded speech after the harmonic analysis and wideband extension has been improved significantly. The wideband speech is pleasant to listen but it shows some hissing artifacts, the on going research is focused to reduce these artifacts, which encounters in higher frequency spectrum.

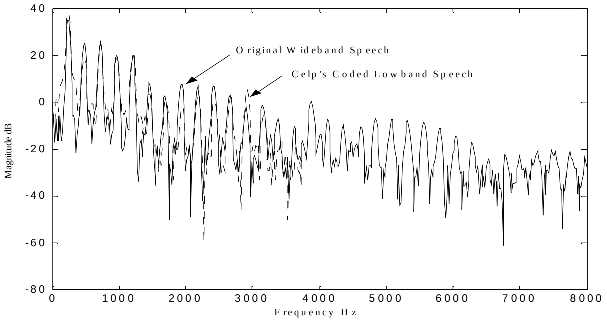

Figure 3(a). Short Time Magnitude Spectrum of CELP's Coded and Original Wideband Speech

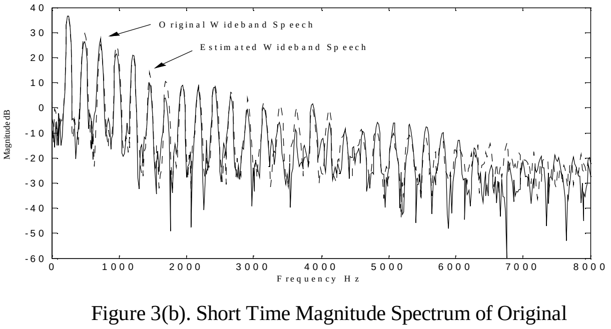

Figure 3(b). Short Time Magnitude Spectrum of Original and Estimated Wideband Speech

8. 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] Itakura, F. (1975) "Line Spectrum representation of Linear Predictor Coefficients of Speech Signal", Trans. Committee on speech Research Acoust. Soc. Jap, S75-34.

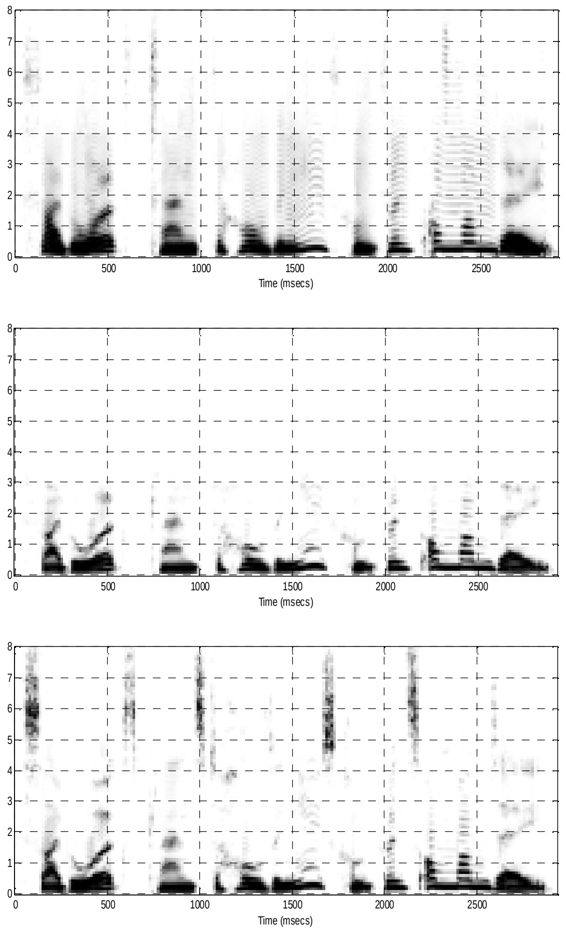

Figure (4). Spectrograms of Estimated Wideband, Original Narrowband and Original wideband Speech Signal.

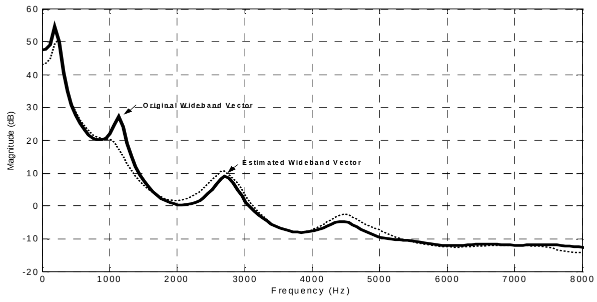

Figure 2. A Voiced Original and Estimated Wideband Spectral Envelopes