WIDEBAND RE-SYNTHESIS OF NARROWBAND CELP-CODED SPEECH USING MULTIBAND EXCITATION MODEL

Cheung-Fat Chan and Wai-Kwong Hui

Department of Electronic Engineering
City University of Hong Kong
83, Tat Chee Avenue Kowloon, HONG KONG
EMail: eecfchan@cityu.edu.hk

ABSTRACT

In this paper, a method for improving the quality of narrowband CELP-coded speech is present. The approach is to reduce the hoarse voice in CELP-coded speech by enhancing the pitch periodicity in the reproduction signal and also to reduce the muffing characteristics of narrowband speech by regenerating the highband components of speech spectra from the reproduction signal. In the proposed method, multiband excitation (MBE) analysis is performed on the reproduction speech signal from a CELP decoder and the pitch periodicity is enhanced by re-synthesizing the speech signal using a harmonic synthesizer according to the MBE model. The highband magnitude spectra are regenerated by matching to lowband spectra using a trained wideband spectral codebook. Information about the voiced/unvoiced (V/UV) excitation in the highband are derived from a training procedure and then stored alongside with the wideband spectral codebook so that they can be recovered by indexing to the codebook using the matched lowband index. Simulation results indicate that the quality of the wideband re-synthesized speech is significantly improved over the narrowband CELP-coded speech.

1. INTRODUCTION

Code excited linear predictive (CELP) coding is one of the widely used low-bit-rate speech coding techniques[1]. It is well-known that speech produced by CELP coders suffers from some quality degradation which are generally described as muffing with hoarse and noisy characteristics. The muffing characteristics are mainly due to the lack of high frequency components in the reproduction signal because these low-bit-rate speech coders were designed to operate at narrowband (0-4 kHz) with 8 kHz sampling frequency. Basically, the CELP coding mechanism employs a block-by-block analysis procedure where the perceptually weighted mean-squared error between the input speech and the synthetic speech is minimized. In the decoder, speech is generated by exciting a synthesis filter with an excitation signal which is constructed as a weighted sum of signals from a long-term adaptive codebook and from a fixed stochastic codebook. The purpose of the long-term adaptive codebook is to introduce the pitch periodicity in the synthetic signal during voiced speech. However, even with the use of high resolution pitch analysis based on fractional tap delay[2], the pitch periodicity introduced is still not sufficient and, also because of the use of noisy stochastic excitation, the background noise level is generally high. Therefore CELP-coded speech is always described as hoarse with noisy characteristics. Because there is a wide installation base of CELP coders in the commercial world, for examples, the Federal Standard FS1016 coder[3], the VSELP coder of EIA/TIA IS54[4] and the full-rate GSM coder[5], there is an urgent need to further improve the quality of CELP-coded speech while keeping their encoding format intact. This paper will describe a method to improve the quality of CELP-coded speech by regenerating the high-frequency components (4-8kHz) at the decoder and also by reducing the coding noise inherent in voiced harmonic bands. The techniques investigated are based on wideband re-synthesis of CELP-coded speech through the use of multiband excitation (MBE) model.

2. ENHANCEMENT BY MBE RE-SYNTHESIS

MBE model can produce speech of high quality because it allows the flexibility of mixing voiced and unvoiced energies in the frequency domain[6]. In MBE model, speech spectrum is divided into a number of signal bands which are centered on the pitch harmonics. Each band can be individually declared as voiced or unvoiced. The MBE model parameters consist of a set of band magnitudes and phases, a set of binary voiced/unvoiced decisions, and a pitch frequency. Specifically, in MBE analysis, a short-time speech spectrum $S(k)$ is matched to a pitch-dependent synthetic harmonic spectrum $E(\tau,k)$ by minimizing an error function $\xi(\tau)$ defined in (1) with respect to the searching variable τ for the pitch period.

$$\xi(\tau)=\frac{\sum_{m=1}^{M(\tau)}\sum_{k=a_m(\tau)}^{b_m(\tau)}\left[|S(k)|-A_m(\tau)|E(\tau,k)|\right]^2}{(1-\tau B)\sum_{m=1}^{M(\tau)}\sum_{k=a_m(\tau)}^{b_m(\tau)}|S(k)|^2} \qquad (1)$$

where $M(\tau)$ is the total number of harmonic bands in the speech spectrum, $a_m(\tau)$ and $b_m(\tau)$ are, respectively, the lower and upper boundaries of the m^{th} harmonic band, B is a weighting factor for biasing the pitch dependent error, and $A_m(\tau)$ is the band magnitude calculated as:

$$A_m(\tau)=\frac{\sum_{k=a_m(\tau)}^{b_m(\tau)}|S(k)||E(\tau,k)|}{\sum_{k=a_m(\tau)}^{b_m(\tau)}|E(\tau,k)|^2} \qquad (2)$$

The optimum pitch period τ_0 is selected such that the lowest matching error is obtained, i.e., $\tau_0 = \arg\min_{\forall\tau}[\xi(\tau)]$. After the determination of pitch period, the optimum band magnitudes are obtained as $A_m \equiv A_m(\tau_0)$ and the optimum matching error in each band is calculated as:

$$\varepsilon_m = \frac{\sum_{k=a_m(\tau_0)}^{b_m(\tau_0)} \left[|S(k)| - A_m(\tau_0)|E(\tau_0,k)|\right]^2}{\sum_{k=a_m(\tau_0)}^{b_m(\tau_0)} |S(k)|^2} \quad (3)$$

The voiced/unvoiced decision for each band is then determined by comparing the band error ε_m to a predefined threshold level. If the band error is smaller than the threshold, a voiced band is detected otherwise an unvoiced band is detected.

For the purpose of wdieband enhancement based on MBE re-synthesis, the lowband information are obtained from narrowband CELP-coded speech using MBE analysis and passed to the wdieband MBE synthesizer for synthesis. The lowband information include the V/UV decisions, the band magnitudes and phases of the voiced bands, and the signal spectrum declared as unvoiced. The enhancement system, therefore, needs to estimate the highband information from all the information available in the lowband. These include the V/UV decisions, the band magnitudes and phases of the voiced bands, and the band magnitudes of the unvoiced bands. Fig. 1 shows the block diagram of the proposed enhancement system.

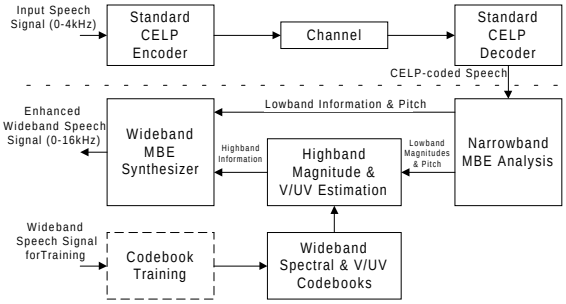

Fig. 1: Block Diagram of the Enhancement System

2.1. Noise Reduction in CELP-Coded Speech

As mentioned previously, the background noise in CELP-coded speech is high because noisy stochastic excitation is used. This background noise in between the pitch harmonics in voiced bands can be easily observed by comparing the spectral plots of the original speech (solid curve) and the CELP-coded speech (doted curve) shown in Fig. 2. The coder used in this simulation is the 4.8kbps FS1016 CELP coder which has a 9-bits fractional adaptive pitch codebook[3]. The spectral plots shown in Fig. 2 clearly indicate that even with high-resolution pitch analysis, CELP coders still produce noisy pitch harmonics.

It is well known that speech produced by MBE coders is perceptually less nosier than those produced by CELP-based coders because a smooth harmonic excitation rather than a noisy stochastic excitation is used. In this research, a simple technique to reduce the coding noise was proposed. The idea is to perform the MBE analysis on the reproduction speech and then to replace the voiced portions of the reproduction speech spectrum by the corresponding voiced harmonic spectrum synthesized by the MBE synthesizer. However, the unvoiced portions of the reproduction speech spectrum will be passed to the MBE synthesizer without modification. The effect of this process is to "clean" up the background noise in between the voiced harmonic bands. Since MBE analysis is performed on the reproduction speech which has weaker pitch harmonics, robust high-resolution pitch detector must be used in the analysis to achieve an accurate estimate of pitch. A frequency-domain pitch estimator based on closed-loop minimization as described previously is sufficient for this purpose. However, since the pitch harmonics at low energy regions of reproduction speech spectrum are heavily corrupted by coding noise, it is necessary to improve the ability to discriminate voiced and unvoiced energies in these regions by allowing small offsets to the harmonic band positions in the synthetic harmonic spectrum during matching. From experimental studies, it is also found that the threshold level for U/UV classification has to be increased from 0.3 normally used in MBE coder to about 0.5. Nevertheless, pitch tracking is generally not necessary because incorrect declaration of voiced bands as unvoiced bands will only affect the noise cleaning ability of the proposed algorithm since all unvoiced energies are allowed to pass through untouched. Note that, unlike in MBE coders, the band magnitudes and phases derived are not quantized in the enhancement system.

The MBE synthesizer used in this experiment is based on a harmonic model where the voiced harmonics are synthesized using sinusoidal oscillators with quadratic phase interpolation using the measured phases[7]. The unvoiced spectrum in the reproduction speech is converted back to time domain via inverse FFT and added to the re-synthesized voiced signal to obtain the output signal. With this enhancement in voiced spectra, the re-synthesized speech was shown by listening tests to be less noisier than the ordinary CELP-coded speech. Fig. 2 also shows a spectral plot of the re-synthesized speech (dashed curve) which clearly indicates the capability of this MBE re-synthesis technique for cleaning-up the coding noise. Pair-wise comparison tests have also confirmed that the CELP-coded speech with pitch enhancement is preferred over the CELP-coded speech without enhancement.

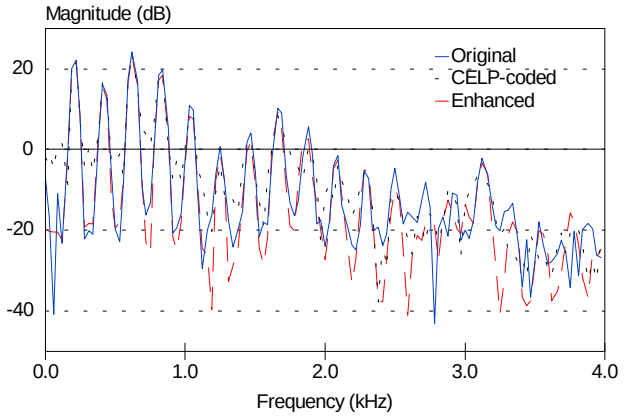

Fig. 1: Spectra of Original, Coded and Enhanced Speech

2.2. Estimation of Highband Envelope from

Lowband Information

It has been observed by many researchers that the high-frequency spectra of speech signal are somehow correlated with their low-frequency counter parts. In this research, a straightforward classification technique was developed to extract the information related to this correlation. This information is then utilized for estimating the highband envelope from the lowband information which is available in the synthesizer. Specifically, in the proposed method, a 20 minutes wideband (16 kHz sampling frequency) speech database contributed by many speakers was designed and used to train a spectral codebook. The training is performed by using the well-known generalized Lloyd algorithm with splitting as initialization. The resulting codebook consists of 1024 line spectral pair (LSP) codevectors, each has a dimension of 18. Conventional autocorrelation LPC analysis with 32ms Hamming window was employed and the weighted mean-squared LSP error was used as distortion measure during VQ codebook generation.

During the estimation of highband envelope in the synthesis stage, the band magnitudes A_m obtained from MBE analysis of narrowband CELP-coded speech are matched against the sampled band magnitudes $\hat{A}_m$ of the wideband LPC spectrum characterized by the 18-dimension LSP vector in the codebook. The spectral distortion used in this comparison is defined as:

$$d = \sum_{m=1}^{M} \left| A_m - \hat{A}_m \right|^2 \tag{4}$$

Specifically, $\hat{A}_m$ are derived from sampling the LPC spectrum at the pitch harmonics, i.e., $\hat{A}_m = g\left|H(m\omega_0)\right|$, where $H(\omega)$ is the LPC spectrum, ω_0 is the (digital) pitch frequency determined as $\omega_0 = \frac{\pi}{\tau_0 f_s}$ with f_s being the signal sampling frequency (8 kHz), and the gain is computed as:

$$g = \frac{\sum_{m=1}^{M} A_m \left|H(m\omega_0)\right|}{\sum_{m=1}^{M} \left|H(m\omega_0)\right|^2} \tag{5}$$

The codevector that achieved the smallest spectral distortion is then selected for estimating the highband envelope. Since MBE model is used for synthesis, the band magnitudes in the high-frequency portions of speech spectrum are generated by sampling the highband portions of LPC power spectrum derived from the selected wideband LSP codevector. Fig. 3(a) shows a set of band magnitudes obtained from MBE analysis on narrowband CELP-coded speech and Fig. 3(b) shows the sampled band magnitudes of the corresponding wideband LPC spectrum. The performance of this highband estimation scheme was shown to be 14dB using the SNR defined over the N highband magnitudes as:

$$SNR = 10\log_{10}\left[\frac{\sum_{n=1}^{N} A_n^2}{\sum_{n=1}^{N} \left|A_n - \hat{A}_n\right|^2}\right] \tag{6}$$

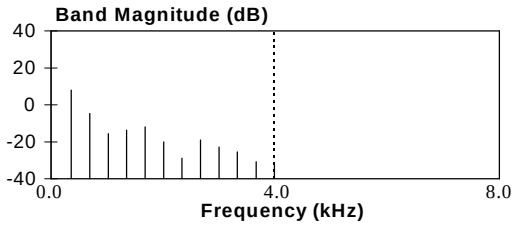

Fig. 3(a): Band Magnitudes from Narrowband CELP-Coded Speech Spectrum

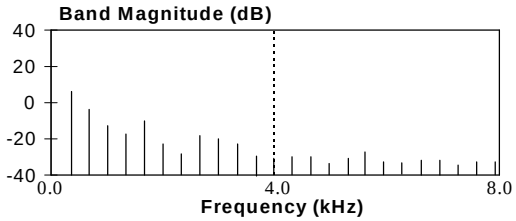

Fig. 3(b): Sampled Band Magnitudes from the Matched Wideband LPC Spectrum

2.3. Estimation of V/UV Information for the Highband

In MBE model, each harmonic band can be declared as either voiced or unvoiced. Obviously, it is necessary to estimate this voiced/unvoiced (V/UV) information in the highband. Many wideband enhancement methods typically assume that the high-frequency spectra consist of entirely unvoiced signals which is generally not correct[8]. In this paper, a method to estimate the highband V/UV information from the wideband spectrum envelope is proposed. This method is based on an assumption that the spectrum envelope and the distribution of V/UV energies in speech signals are highly correlated. This assumption can be readily shown to be valid by observing that high-energy formant regions in speech spectra contain mostly voiced energies while low-energy high-frequency regions of speech spectra largely contain unvoiced energies. In the proposed method, the same wideband speech database was used to design a separate codebook containing V/UV mixture functions as codewords. Each of these V/UV codewords is derived from a clustered set of V/UV mixture functions during the codebook training by taking their statistical average[9]. This V/UV codebook has a one-to-one correspondence to the wideband spectrum codebook derived earlier. The V/UV mixture function derived from the training database actually is a smoothed version of the band error function obtained during MBE analysis. The band error function is an indication of the distribution of V/UV energies in the speech spectrum, i.e., for voiced speech, the error is small, while for unvoiced speech, the error is large. Since only the V/UV information in the highband are needed in the synthesis stage, the storage of the V/UV codebook can be reduced to halve. During speech synthesis when the best wideband codeword in the spectrum codebook is determined, the corresponding V/UV mixture function from the V/UV codebook is also extracted. This V/UV mixture function is then utilized to assign the voiced and unvoiced signals in the highband.

2.4. Wideband MBE Synthesis

With all the necessary information for the highband available, synthesis of the wideband signals using the MBE model is rather

easy. Note that, for ordinary CELP decoders, if a LPC frame size of, say, 20 ms is employed, 160 samples of narrowband speech are needed to be synthesized. For synthesis of wideband signal, 320 samples will be needed for a LPC frame size of 20 ms. In this work, a conventional MBE synthesizer which generates voiced speech in time domain and unvoiced speech in frequency domain is employed. For voiced speech, the phase information for the lowband band magnitudes are extracted from the CELP-coded speech and utilized to control the phases of harmonic oscillators at frame boundaries. Note that all band magnitudes are linear interpolated between frames and their phase functions are quadratic functions. Since the phase information for the voiced bands in the highband are unavailable, phase functions for voiced bands in the highband are made to be slowly evolving functions. Unvoiced speech is synthesized in frequency domain. The unvoiced spectrum in the lowband are extracted from the original CELP-coded speech spectrum unmodified. The unvoiced spectrum in the highband are constructed by multiplying the estimated highband spectrum envelope with a unity energy white noise spectrum. The wideband unvoiced spectrum is then converted to time domain by using a 512-point IFFT and 320 samples of unvoiced signal are then obtained with an overlap-add method. Finally, the voiced and unvoiced signals are added to obtain the wideband synthetic speech signal.

3. SIMULATION RESULTS

In this simulation, a CELP coder was used to test the proposed wideband re-synthesis algorithm. All speech signals were initially band-limited and sampled at 16 kHz (this signal is denoted as wideband input signal). Afterward, the wideband input signal is band-limited to 4 kHz using a lowpass digital Butterworth filter and then sub-sampled by a factor of two to obtain a narrowband input signal. In this evaluation, several narrowband input speech sentences were firstly encoded and then decoded by a CELP coder conformed to the US Gov. FS1016 standard. The re-synthesis algorithm was applied directly on the reproduction speech from the CELP decoder. The wideband re-synthesized speech was then played back through a D/A converter operating at 16 kHz sampling rate. Fig. 4(a) shows the spectrogram of the wideband input signal and Fig. 4(b) shows the spectrogram of the enhanced wideband synthetic signal from the corresponding narrowband CELP-coded speech. Table I shows the MOS achieved for the narrowband CELP-coded speech and the wideband re-synthesized speech using the proposed algorithm. In all sentences tested, the MOS of the wideband re-synthesized speech were higher than those CELP-coded speech without enhancement by an average margin of about 0.6. Listeners in this test felt that the wideband re-synthesized speech is clean with crispy high-frequency characteristics and they all overwhelmingly agreed that the re-synthesized speech is more pleasant to listen to than the ordinary narrowband CELP-coded speech.

	Narrowband CELP-Coded Speech	Wideband Re-synthesized Speech
Sentence A (male speaker)	3.12	3.68
Sentence B (female speaker)	3.32	3.91

Table I Improvements in MOS

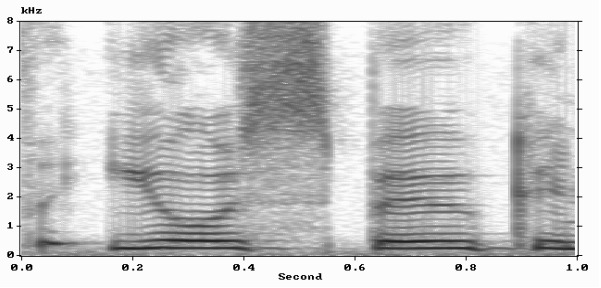

Fig. 4(a): Spectrogram of Original Speech

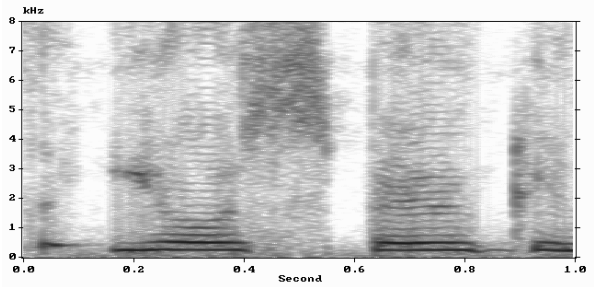

Fig. 4(b): Spectrogram of Enhanced Speech

4. CONCLUSION

The proposed method based on wideband MBE re-synthesis of narrowband CELP-coded speech was shown to be capable of improving the quality of CELP-coded speech without altering their encoding format. This was achieved by reducing the coding noise in between the pitch harmonics of voiced speech and regenerating the highband information from the lowband information available in the decoder.

5. 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. Kroon, P., and Atal, B.S., "Pitch Predictors with High Temporal Resolution," ICASSP, pp.661-664, 1990.
3. Campell, J.R., Jr., Tremain, T.E., and Welch, V.C., "The Federal Standard (FED-STD) 1016 4800 bps CELP Voice Coder," Digital Signal Processing I, pp.145-155, 1991.
4. Gerson, I.A., and Jasiuk, M.A., "Vector Sum Excited Linear Predictive (VSELP) Speech Coding at 8 kbps," ICASSP, pp.461-464, 1990.
5. Gerson, I.A., and Jasiuk, M.A., "A 5600 bps VSELP Speech Coder Candidate for Half-Rate GSM," IEEE Workshop on Speech Coding for Telecommunications, pp.43-44, 1993
6. Griffin, D.W., and Lim, J.S., "Multi-band Excitation Vocoder," IEEE Trans. On Acoustics, Speech, and Signal Processing, Vol. ASSP-36, No. 8, pp.1223-1235, August, 1988.
7. Digital Voice Systems, "IMMARSAT M Voice Codec, Version 2," IMMARSAT-M Specification, IMMARSAT, Feb. 1991.
8. Carl, H. and Heute, U., "Bandwidth enhancement of narrow-band speech signals", EUSIPCO VII. pp.1178-1181, 1994.
9. Chan, C.F., "High-Quality Synthesis of LPC Speech Using Multiband Excitation Model," European Conference on Speech Communication and Technology, pp.535-538, 1993.