MBE SYNTHESIS OF SPEECH CODED IN LPC FORMAT

K. F. Lam and C. F. Chan

Department of Electronic Engineering
City Polytechnic of Hong Kong
Tat Chee Avenue
HONG KONG

ABSTRACT

A method to produce high quality speech from signals coded in linear predictive coding (LPC) format is proposed. In this method, speech signals coded in LPC format is synthesized via multiband excitation (MBE) model. Since MBE model demands more parameters than LPC model can provide, techniques had been developed to regenerate the necessarily information for MBE synthesis from LPC coded speech. A V/UV regeneration scheme for MBE synthesis is proposed to extract the V/UV decisions from spectral envelopes by a long term statistical training approach. Informal listening shows that the method provides a significant improvement of speech quality over conventional LPC synthesis. Simulation shows that the synthetic speech produced by the proposed method is comparable to the speech synthesized using MBE.

1. INTRODUCTION

LPC [1] has been widely used in low bit rate speech coding and synthesis. The parameters for a short time speech spectrum coded in LPC format includes a set of reflection coefficients, a voiced/unvoiced (V/UV) flag and pitch value if the spectrum is declared as voiced.

MBE model is a speech production model developed by Griffin [2]. The parameters for a short time spectrum coded in MBE format are spectral amplitude of each harmonic band, pitch value and a set of V/UV decisions for the spectrum. Due to the improved V/UV excitation, MBE can produce high quality speech.

Although LPC is well developed with many practical applications, it is basically a low quality synthesis technique due to its simple V/UV excitation model. Applying MBE synthesis to the LPC coded speech is a cost effective way to improve the speech quality.

In order to apply MBE synthesis to LPC data, we had to convert the LPC parameters into the MBE parameters. The spectral amplitude of each harmonic band can be simply obtained by sampling the LPC spectral envelope which is represented by the reflection coefficients. However, there are two MBE parameters, the V/UV decision for each harmonic band and the pitch value for unvoiced frames, needed to be recovered from LPC parameters.

1.1. Regeneration of V/UV decisions

A MBE synthesizer requires a set of V/UV decisions for generating the excitation. It is obviously necessary to regenerate a set of V/UV decisions from the existing information coded in LPC. We observed that the spectral envelope and the V/UV allocation among the spectrum are highly correlated. This assumption can be verified by examining some typical speech spectra. Fig.1(a) and Fig.1(b) show a typical voiced and unvoiced spectra, respectively. We noticed that strong formants normally appear in voiced spectra. The harmonic bands around these formants are most likely to be voiced. However, in unvoiced spectra, there do not normally have strong formants and the harmonic bands are most likely to be unvoiced. This correlation can be easily extracted by a long term statistical training technique and the V/UV decisions can then be regenerated using this correlation.

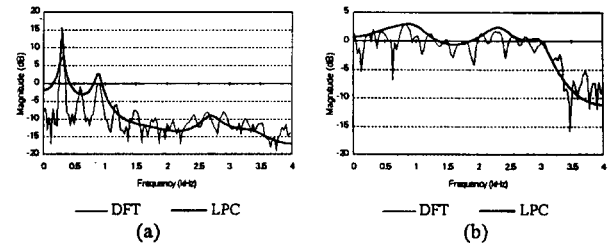


Fig.1(a) Voiced speech spectrum (b) Unvoiced speech spectrum

Since voiced and unvoiced frames have very different spectral envelope shape (Fig.1), it is reasonable to form two codebooks for these two kinds of vectors using VQ. Each codebook contains spectral information represented by the reflection coefficients of LPC-10 format. It is found from observation that unvoiced frames are normally with flatter spectra (Fig.1(b)), a codebook with smaller size can be used for VQ classification. The two spectral codebooks for voiced frames and unvoiced frames we generated are with 64 codewords and 32 codewords, respectively. Each spectral codevector is assigned with a set of V/UV decisions retrieved from the long term statistical training. The V/UV flag defined in LPC is used to determine which of the codebooks, voiced or unvoiced codebook, is to be used.

1.2. Recovery of pitch value for unvoiced frames

Unlike LPC synthesizers, MBE synthesizers need the pitch value for each speech spectrum, while LPC provides this information only when a spectrum is declared as voiced. By making an assumption that the pitch contour

0-7803-2431-5/95 $4.00 © 1995 IEEE

evolves slowly with time, the pitch value of an unvoiced spectrum can be set to the pitch value of the previous voiced spectrum.

Fig.2 shows the system proposed to synthesize LPC coded speech using a MBE synthesizer.

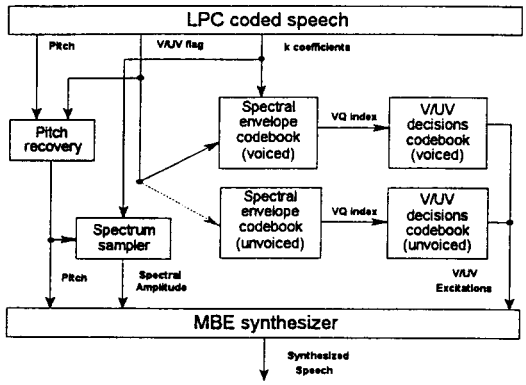


Fig.2 Block diagram of the V/UV decisions recovery system

2. TRAINING OF THE V/UV DECISIONS CODEBOOK BY MBE ANALYSIS

In order to retrieve the V/UV decisions from the spectral envelopes, we firstly determine the degree of V/UV mixture of a spectrum. Since the V/UV error defined in the original MBE analysis is not linearly proportional to the degree of V/UV mixture, a linearization function is added to the error. A set of V/UV mixture function is then defined according to the linearized V/UV error. By comparing the averaged V/UV mixture function to a threshold value, a set of V/UV decisions is obtained. The set of retrieved V/UV decisions is then stored in either of the voiced or unvoiced V/UV decisions codebook depending on its classification.

2.1. Weighted V/UV error

In the original MBE analysis, the V/UV error of each harmonic band is defined as

$$\xi_m = \frac{\sum_{\omega=a_m}^{b_m}\left[|S(\omega)| - A_m|E(\omega)|\right]^2}{(1-M_O B)\sum_{\omega=a_m}^{b_m}|S(\omega)|^2} \qquad (1)$$

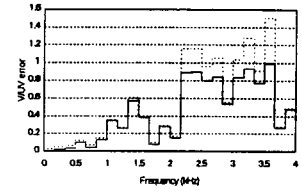
where $1/(1-M_O B)$ is an unbiasing factor, $S(\omega)$ is the input speech spectrum, $E(\omega)$ is the excitation spectrum obtained by assuming all the harmonic bands are voiced, A_m is the band magnitude, B is a constant found out by experiments and M_O is the total number of harmonic bands. The interval $[a_m, b_m]$ is the interval of the mth harmonic of the pitch frequency.

If the unbiasing factor is absent, the error will be smaller for longer pitch periods because the spectral envelope is sampled more densely for longer pitch periods [4]. However, the unbiasing factor is a non-linear function with respect to pitch period and it causes the value of the V/UV error of an individual band, ξ_m, not linearly proportional to the degree of V/UV mixture of the harmonic band. Since the V/UV error is used to generate the averaged mixture function, $\bar{\alpha}(\omega)$, which will be described later in the paper, non-linearity in the error will bias the function so that it cannot represent the degree of V/UV mixture correctly. The error is then necessary to be linearized.

One method to linearize the V/UV error is to multiply the error with a non-linear function. In the original MBE analysis, there is a V/UV threshold, T_{vuv}, used for determining V/UV classification. It is observed that the non-linearity for the error greater than T_{vuv}, i.e. unvoiced case, and the error smaller than the T_{vuv}, i.e. voiced case, are different. In the unvoiced case, the V/UV error is found exponentially proportional to the degree of V/UV mixture. However, in the voiced case, the V/UV error is quite linearly proportional to the degree of V/UV mixture. The linearized V/UV error is then defined as

$$\varphi_m = \begin{cases} T_{vuv} - F_v(T_{vuv} - \xi_m), & \xi_m \le T_{vuv} \quad voiced \\ T_{vuv} + F_{uv}(\xi_m - T_{vuv}), & \xi_m > T_{vuv} \quad unvoiced \end{cases} \qquad (2)$$

where F_v is a scaling factor, F_{uv} is a non-linear function. It is found $F_{uv}=1/\mathrm{Exp}(2(\xi_m - T_{vuv}))$ and $F_v=1.1$ work well in practice. Fig.3 shows the V/UV error, ξ_m, and the linearized V/UV error, φ_m, for a short time spectrum with 24 harmonic bands. The V/UV threshold value, T_{vuv}, used is 0.35.



----- ξ_m —— φ_m

Fig.3 The V/UV error, ξ_m, and the weighted V/UV error, φ_m, for a short time spectrum with 24 harmonic bands

2.2. V/UV mixture function

Since we have made an assumption that the spectral envelope and the V/UV allocation among a short time speech spectrum are highly correlated and verified this is true, a long term statistical training technique can be employed to extract this correlation from a set of training vectors.

We define the V/UV mixture function for a short time speech spectrum which indicates the degree of V/UV mixture of the spectrum as a function of frequency as

$$\alpha(\omega)=\begin{cases}1-\varphi_1 & a_1 \le \omega \le b_1\\ 1-\varphi_2 & a_2 \le \omega \le b_2\\ \cdots\cdots & \\ 1-\varphi_M & a_M \le \omega \le b_M\end{cases} \quad (3)$$

MBE analysis is applied to the set of vectors and each vector will obtain its own V/UV mixture function, $\alpha(\omega)$. Each training vector is then mapped to a corresponding spectral codevector described earlier using VQ. By averaging all the $\alpha(\omega)$ of training vectors mapped to a codevector, an averaged V/UV mixture function for this codevector, $\bar{\alpha}(\omega)$, is obtained.

Fig.4 shows two typical averaged V/UV mixture functions, $\bar{\alpha}(\omega)$, for codevectors in the voiced codebook and the unvoiced codebook. In the voiced spectrum (Fig.4(a)), the value of the averaged V/UV mixture in the low frequency region is high. This agrees with our observation stated earlier in the paper that, in voiced spectra, the harmonic bands around the strong formants are most likely to be voiced. In the unvoiced spectrum (Fig.4(b)), the value of the averaged V/UV mixture is rather low. This also agrees with the observation that the harmonic bands in unvoiced spectra are most likely to be unvoiced.

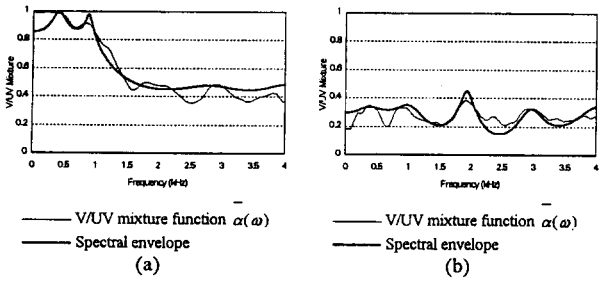


Fig.4(a) An averaged mixture function, $\bar{\alpha}(\omega)$, for a voiced codevector, (b) An averaged mixture function, $\bar{\alpha}(\omega)$, for an unvoiced codevector

2.3. Threshold value for V/UV mixture function

By comparing the averaged mixture function to a threshold value, T_{mix}, a binary series of V/UV decisions with respect to ω for a codevector, $VUV(\omega)$, can be obtained. If $\bar{\alpha}(\omega) \ge T_{mix}$, the V/UV decision at this frequency is declared as voiced, i.e. $VUV(\omega)=1$, otherwise it is declared as unvoiced, i.e. $VUV(\omega)=0$. In order to declare a harmonic band as voiced or unvoiced, we can simply compare the voiced and unvoiced portions of the V/UV decisions in this frequency interval. If $\frac{1}{b_m-a_m}\sum_{\omega=a_m}^{b_m} VUV(\omega) \ge 0.5$, the harmonic band is declared as voiced, otherwise it is declared as unvoiced. The V/UV decision for each harmonic band in a spectrum can be retrieved.

In order to determine the optimal threshold value objectively, we minimize the percentage of wrongly classified harmonic bands for each spectral codevector which is defined as

$$E_{vuv}=\frac{1}{T}\sum_{i=1}^{T}\left[\frac{n_i}{M_i}\right]\times 100\% \quad (4)$$

where M_i is the total number of harmonic bands in a short time speech spectrum, T is the total number of training vectors mapped to the spectral codevector and n_i is the number of wrongly classified harmonic bands which is determined by comparing the set of V/UV decisions generated by original MBE analysis of the spectrum and the set of V/UV decisions retrieved by the proposed method.

Obviously, n_i is a function of the threshold, i.e. $n_i(T_{mix})$, we can simply adjust the value of T_{mix} from the range [0,1] until the percentage error reaches its minimum value.

3. PERFORMANCE COMPARISON

An objective measurement of performance is used to compare the proposed method and the conventional LPC method. The percentage of wrongly classified harmonic bands, ε_{vuv}, for the two methods are measured by applying them to a large set of speech. Table 1 shows the percentage errors of speech inside and outside the training set for generating the codebooks.

	Inside training set		Outside training set	
	LPC	proposed method	LPC	proposed method
ε_{vuv}	29.53	8.81	37.47	11.06

Table 1 Percentage errors of wrongly classified harmonic band

The results show that the proposed V/UV recovery method provides a significant improvement to the conventional LPC coded speech.

4. MBE SYNTHESIS

In order to synthesize LPC speech using MBE model, we firstly derive the spectrum magnitudes of each harmonic band by sampling the LPC spectral envelope which is represented by 10 k-coefficients. A set of V/UV decisions with respect to ω can be obtained by mapping the spectral envelope to the spectral codebook, in which each codeword has a set of V/UV decisions assigned to it, using the VQ searching technique. The V/UV decision of each harmonic band is easily retrieved by comparing the voiced and unvoiced portions of the frequency band. If the frame is declared as voiced, we can have the pitch value directly

from LPC code. If the frame is declared as unvoiced, we use the pitch value provided by the previous voiced frame. We synthesize the voiced harmonic bands in time domain by summing up the outputs of oscillators oscillating at the harmonic frequencies. The magnitudes of sine waves generated from the oscillators are set to the spectrum magnitudes. The unvoiced portion of the spectrum is synthesized in frequency domain. A white noise sequence is windowed using a Hamming window and is transformed to frequency domain using FFT. It is then normalized to have unity energy and multiplied by each unvoiced band magnitude in frequency domain. The unvoiced spectrum is then transformed back to time domain and added to voiced speech using a weighted overlap-add method. Since the phase information is absent in LPC code, we used a quadratic phase interpolation [3] scheme for smoothing the voiced speech in adjacent frames.

5. SIMULATION AND RESULTS

Speech signal used for the synthesis is coded in TI5220 LPC code format. The number of bits used for coding the LPC parameters are shown in Table 2.

	Gain	Pitch	Repeat Frame	k-coefficients
no. of bits	4	6	1	39

Table 2 Bit allocation of TI5220 format

Speech signal used for training V/UV codebooks and spectral codebooks is sampled at 8kHz and windowed by 256 points Hamming windows. The duration of each frame is 25ms. The signal is then transformed to frequency domain using FFT. The spectral envelope stored in the spectral codebooks are represented using 10 k-coefficients. The V/UV decisions series for each codevector is stored using 254 bits. Fig.5(a) shows the original spectrum overlayed with the synthetic spectrum produced by the proposed method. Synthesis is also performed with a MBE vocoder and a LPC vocoder for comparison and the synthetic spectra are shown in Fig.5(b) and Fig.5(c). The bit allocation for the LPC vocoder is in TI5220 format. The bit allocation for the MBE vocoder is shown in Table 3.

	Gain	Pitch	V/UV	LSP coefficients
no. of bits	6	8	16	34

Table 3 Bit allocation of TI5220 format

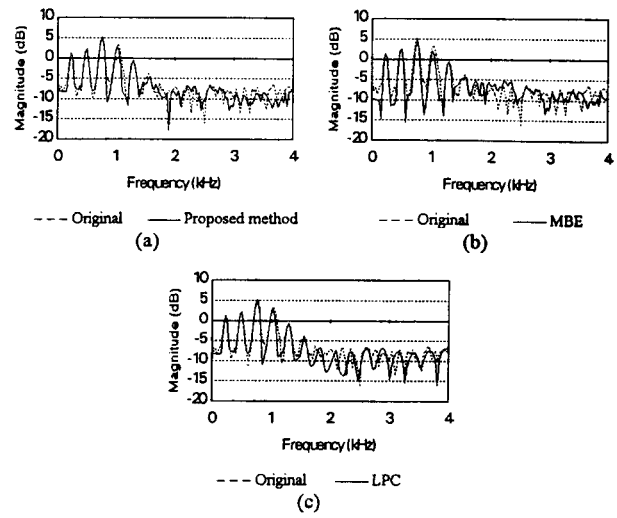


Fig.5(a) New method (b) MBE (c) LPC

It is obviously that, from the simulation result, the synthetic spectrum produced using the proposed method is comparable to speech produced using conventional MBE vocoder. From the informal listening test, the synthetic speech using the proposed method shows a significantly improvement comparing to the speech produced by original LPC synthesizer. The speech sounds much more natural and smooth. It does not have any noticeable reverberant.

6. CONCLUSION

In this paper, we presented a method to synthesize LPC speech using MBE model. This improved the quality of the conventional LPC synthesis method. We also presented a method to recover the V/UV decisions for MBE from LPC spectral envelopes using long statistical training technique.

7. REFERENCES

[1] J.Makhoul, "Linear Prediction: A tutorial review", Proc. IEEE, pp561-580, 1975.

[2] D.W.Griffin and J.S.Lim, "Multiband Excitation Vocoder", IEEE Trans. on Acoustics, Speech, and Signal Processing, pp1223-1235, 1988.

[3] H.Yang, S.N.Koh and P.Sivaprakasapillai, "Quadratic Phase Interpolation for Voiced Speech Synthesis in MBE Model", Electronics Letters, pp856-857, 13th May 1993.

[4] D.W.Griffin, "Multiband Excitation Vocoder", Ph.D. dissertation, M.I.T., Cambridge, MA, 1987.