A 2000 BPS LPC Vocoder based on Multiband Excitation*

J. S. Mao, Z. Chen, S.C.Chan, T.S.Ng and K.L.Ho
Department of Electrical and Electronic Engineering
The University of Hong Kong
Pokfulam Road, Hong Kong
jsmao@eee.hku.hk

ABSTRACT

This paper presents an improved mixed LPC vocoder at 2000 bps using Multi-Band Excitation analysis by synthesis algorithm. The new vocoder determines the voiced/unvoiced characteristics harmonic by harmonic in a frame, and finds the first voiced/unvoiced transition as the cut-off frequency, which is more accurate and efficient than traditional detection of cut-off frequency. The synthetic speech below the cut-off frequency is excited by a series of voiced harmonics, while the signal above the cut-off frequency is simulated by noise source. The final output speech is the sum of these two outputs. To increase the naturalness and clearness of the synthesized speech, this model applies phase prediction and spectral enhancement in the synthesizer. It is also possible to reduce the bit rate to 1200 bps. Informal listening test indicates that the output speech possesses higher intelligibility and quality than that of the 2.4 kbps LPC-10e standard, and is comparable to the 4.8 kbps FS1016 CELP vocoder.

1 INTRODUCTION

The concept of mixed LPC vocoder is first proposed by Makhoul et al [1]. In this model, the speech spectrum is divided into two regions, with the pulse source exciting the low-frequency region and the noise source exciting the high-frequency region. The synthesized speech is a mixture of the two parts. The cut-off frequency, which determines the voicing degree, is extracted from the signal spectrum by peak-picking algorithm. Although the mixed LPC vocoder is capable of producing more natural-sounding speech, efficiently eliminating most of buzz and recovering part of the fullness of natural speech, the errors in detecting the cut-off frequency may lead to excessive noise in synthesized speech. Moreover, the lack of phase determination in the voiced excitation cannot ensure high quality output speech.

More recently, McCree and Barnwell have presented a new mixed excitation LPC vocoder for low bit rate speech coding [2,3]. They preserve the low bit rate of the fully parametric model and introduce more free parameters to the excitation signal so that the synthesizer can mimic more characteristics of natural human speech. For example, there are two types of excitation pulses: periodic and aperiodic pulses, which are distinguished by *strongly voiced* and *weakly voiced decision*, respectively. The performance of the improved vocoder is close to that of the U.S. government standard 4.8 kbps CELP coder.

On the other hand, a novel 8 kbps Multi-Band Excitation (MBE) vocoder was proposed by Griffin and Lim in 1988 [4]. In the MBE vocoder, the short time speech spectrum is divided into nine or twelve bands. Each band is declared as either voiced or unvoiced according to the following principles: if the frequency band contains primarily periodic energy, it will be treated as voiced harmonic, otherwise it will be regarded as an unvoiced band. The synthesized speech is the sum of these multi-band signals. The voiced signals are generated by combining the harmonics of the fundamental frequency and the unvoiced signals are approximated with noise source. Later, different models have been proposed with better efficiency and robustness to channel errors. This includes the INMARSAT-M Codec [5], the 2400 bps MBE vocoder by Meuse [6], and the simplified v/uv division vocoder at 3000 bps by Nishiguchi [7]. We have made use of the simplified v/uv decision [7] in this work, but the neighboring information of a given harmonic is also used to determine the v/uv decision.

In very low bit rate speech coding below 2 kbps, bits required for v/uv decision and the LSP quantization have to be reduced. In this paper, an improved mixed LPC vocoder operating at 2000 bps is proposed based on the traditional MBE model. However, the cut-off frequency is estimated with more accuracy in the harmonic v/uv decision, and adaptive spectral enhancement techniques are used to improve the intelligibility and quality of the synthesized speech. Furthermore, it is possible to reduce the bit rate down to 1200 bps by this model.

The layout of this paper is organized as follows: Section 2 describes the structure of the mixed LPC vocoder and the analysis algorithm. Section 3 and section 4 are devoted to the quantization techniques and spectral enhancement of the synthesizer, respectively.

* This work is supported by the Hong Kong Research Grants Council and the Block Grant of The University of Hong Kong.

0-7803-3583-X/97 $10.00 ©1997 IEEE

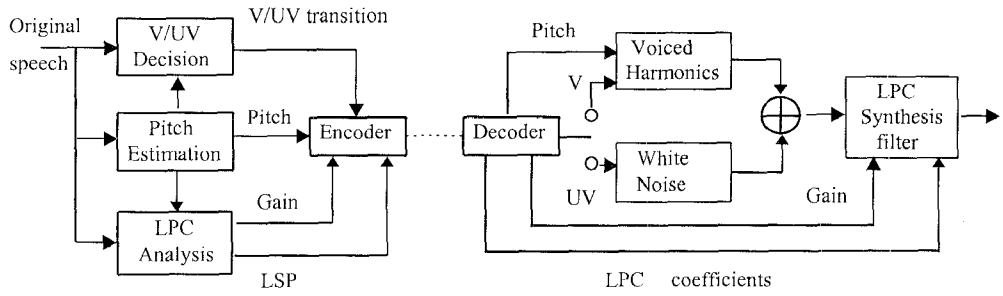


Figure 1. Structure of improved mixed LPC vocoder

Performance evaluation of the coder is discussed in section 5. Finally, the discussion and conclusion are given in section 6.

2 ANALYSIS

2.1 THE ARCHITECTURE OF IMPROVED MIXED LPC VOCODER

The structure of the improved mixed LPC vocoder is shown in figure 1. The sampling rate and the frame update rate of the new vocoder are respectively 8000 Hz and 50 Hz. The encoder consists of three parts: pitch estimation, LPC parameters modeling and voiced/unvoiced decision. In the decoder, output speech is synthesized by passing the sum of the voiced and the unvoiced excitation through the LPC synthesis filter.

2.2 PITCH ESTIMATION

The accuracy of the pitch determines the naturalness of the synthesized speech. In our coder, pitch detection is performed in two stages. The first stage is initial pitch detection, which uses an auto-correlation periodicity detector. After filtering by the inverse LPC filter, the residual signals are auto-correlated. The initial pitch is chosen as the time delay with the highest correlation value. To prevent pitch doubling, the smallest sub multiple of this initial pitch which has a correlation value greater than 50% of the highest value is used [8]. The second stage is pitch refinement. A fine search is performed in frequency domain centered at the initial pitch to obtain full pitch resolution up to one quarter sample. In order to get a smooth pitch tracking among neighboring frames, a look-back and look-ahead pitch tracking system is used. The estimation errors are almost corrected after this pitch smoothing.

2.3 V/UV TRANSITION DETECTION

In the new mixed LPC vocoder, the task of v/uv decision is to find the position of the first voiced/unvoiced transition, which corresponds to the cut-off frequency. Below this cut-off frequency, the speech is declared as voiced while the harmonics above this frequency are declared as unvoiced. The decision is done by calculating the normalized error between the original and the synthesized spectrum [4]:

$$\varepsilon_k = \frac{\int_{am}^{bm} G(w)\left[\left|S\ w(w)\right| - \left|Am\right| \cdot \left|Ew(w)\right|\right]^2 dw}{\int_{am}^{bm} G(w)\left|Sw(w)\right|^2 dw} \qquad (1)$$

Where ε_k is the normalized error of the kth harmonic. The interval (a_m, b_m) is an interval with width which is three times of the fundamental frequency and is centered at the kth harmonic. G(w) is a frequency weighting function. $S_w(w)$ is the hamming windowed original spectrum. Am is the synthesized spectral envelope at kth harmonic, and $E_w(w)$ is the excitation spectrum. To reduce the error in computing ε_k of the kth harmonic, we use both the kth and neighboring harmonics to determine ε_k. If ε_k is below an adaptive threshold [5], the kth harmonic is declared as voiced, otherwise it is unvoiced. After the v/uv decision is completed, the first v/uv transition location is chosen as the cut-off frequency. Under some circumstances, there are only one or two harmonics in the first unvoiced region, it is not suitable to regard the first v/uv transition as the cut-off frequency. Thus the second voiced to unvoiced transition will replace the first one. This will help to reduce the hoarseness of the synthesized speech. Figure 2 shows the mixed v/uv bands in a frame.

2.4 LPC PARAMETERS ANALYSIS

The short-term prediction is performed in frequency domain using the all-pole model:

$$H(w) = \frac{G}{A(w)} = \frac{G}{1 - \sum_{k=1}^{P} a_k \cdot e^{-jkw}} \qquad (2)$$

Where G is the LPC gain, P is the filter order (Here P=10), and A(w) is the inverse LPC filter transfer function. By minimizing the error between original spectrum $S_w(w)$ and the LPC spectrum envelope $H(w)$, the following equations are deduced:

$$\sum_{k=1}^{P} a_k \cdot R_{|i-k|} = -R_i \qquad (3)$$

$$R_k = \frac{1}{L} \sum_{l=0}^{L} \left|Sw(lw_0)\right|^2 \cdot \cos(klw_0) \qquad (4)$$

Where a_k is the kth LPC coefficients, L is the number of harmonics, w_0 is the fundamental frequency.

In order to compensate for the inaccurate modeling of the spectrum envelope with high pitch, we use linear interpolation in the log spectral domain [9]:

$$\log_{10}^{|Q(w)|} = \log_{10}^{|Sw(w)|} + \left(\frac{w - w_k}{w_{k+1} - w_k}\right)\left[\log_{10}^{|Sw(w)|} - \log_{10}^{|Sw(w)|}\right] \quad (5)$$

Here w_k is the frequency of the kth harmonic, Q(w) is the interpolated spectrum, and $w_k \le w \le w_{k+1}$.

3 QUANTIZATION

To ensure the stability of the LPC filter and have an efficient numerical computation, we transfer LPC parameters into Line Spectral frequency Parameters (LSP). In very low bit rate speech coding, the quantization of the LSPs is restricted to as few as 24 bits per frame. Therefore we use a 10th order LPC filter and the split VQ scheme to represent the spectral envelope. The LSPs are divided into two parts. The first part consists of the first 4 components in the LSP vector and the second contains the remaining 6 components. Each part is represented by a codebook with 4096 levels [10]. In pitch quantization, we choose a uniform quantizer with 128 levels, the quantized pitch ranges from 20 to 119. The LPC gain is quantized with 5 bits, while the v/uv transition is quantized with 4 bits. Table 1 describes the bit allocation:

Parameters	Bits/Frame	Bit Rate(bps)
LSPs	24	1200
Pitch	7	350
Gain	5	250
v/uv transition	4	200
Total	40	2000

Table 1. Bit allocation for the 2000bps mixed LPC vocoder

4 SPECTRAL ENHANCEMENT

4.1 LSP PARAMETERS INTERPOLATION

In speech synthesis, the frame-by-frame update of the LPC parameters controls the degree of accuracy with which the LPC filter can model the speech spectrum. When transitions that are perceptually important, occur in a frame, the model will fail to track the actual spectral shape accurately. This will cause perceivable distortion such as tremble sound especially in long and slow pronunciation. So LSP parameters interpolation between current frame and previous frame is necessary. In our coder, the current frame is divided into four sub frames, and linear interpolation is performed between the decoded LSP vector, $\widetilde{P}_n$ and the previous LSP vector, $\widetilde{P}_{n-1}$, for each sub frame. Four interpolated LSP vectors $\{\widetilde{P}_i\}$ i=0~3, are converted to LPC vectors, $\{\widetilde{a}_i\}$ i=0~3. The quantized LPC synthesis filter, $Ai(z)$(i=0~3), is used for synthesizing the decoded speech signal.

$$\widetilde{P}_{ni} = \begin{cases} 0.75\widetilde{P}_{n-1} + 0.25\widetilde{P}_n \\ 0.5\widetilde{P}_{n-1} + 0.5\widetilde{P}_n \\ 0.25\widetilde{P}_{n-1} + 0.75\widetilde{P}_n \\ \widetilde{P}_n \end{cases} \quad (6)$$

$$\widetilde{A}i(z) = \frac{1}{1 - \sum_{j=1}^{10} \widetilde{a}_{ij} Z^{-j}} \quad (7)$$

4.2 POSTFILTERING

To improve the perceptual quality of the synthesized speech, a postfilter is used in our codec. Since speech formant peaks are much more important than the formant valleys, the postfiltering strategy is to preserve the formant information by keeping the noise in formant valleys as low as possible. The postfilter that we used is constructed in frequency domain and is given by [9]:

$$H_p(e^{jw}) = \left(\frac{Hw(e^{jw})}{H\max}\right)^{\beta} \qquad 0 \le \beta \le 1 \quad (9)$$

where $H_p(e^{jw})$ is the postfilter function. $Hw(e^{jw})$ is the weighted synthetic spectral envelope, and $H\max$ is the maximum value of $|Hw(e^{jw})|$, β is the power (typical value is 0.2), and

$$Hw(e^{jw}) = H(e^{jw})W(e^{jw}) \quad (10)$$

$$W(e^{jw}) = \frac{1}{H(e^{jw}/\gamma)} \qquad 0 \le \gamma \le 1 \quad (11)$$

Here $H(e^{jw})$ is the LPC spectral envelope, $W(e^{jw})$ is the weighting function, and γ is the weighting coefficient which is typically 0.5. The final synthesized spectral function is:

$$\hat{H}(e^{jw}) = H_p(e^{jw})H(e^{jw}) \quad (12)$$

After postfiltering, the formant peaks are made narrow and the depth of formant valleys is increased. Thus it reduces the effects of noise, and improves the speech quality substantially. Figure 3 shows the LPC spectral envelope before and after postfiltering. Figure 4 shows the spectrum of the original and the synthesized speech.

5 PERFORMANCE EVALUATION

The performance of the proposed 2000 bps mixed LPC vocoder is evaluated by an informal listening test, which involves 20 listeners, with 20 sentences spoken by male and female speakers. The 2400 bps LPC-10e standard vocoder and 4800 bps FS1016 CELP are used as reference. Each listener will listen to a sentence twice for a specified vocoder. After listening to the sentence in three different coding systems, the listeners are asked to give their preference. The statistical result shows that average preference for the LPC-10e is 10%, while the preference for the improved mixed LPC vocoder is 40%, and that of the FS1016 4800 bps CELP is 50%. This

indicates that performance of the improved mixed LPC vocoder is much better than the 2400 bps LPC-10e and comparable to the 4800 bps FS1016 CELP.

6 DISCUSSION AND CONCLUSION

In this paper, an improved mixed LPC vocoder using multi-band excitation model is presented. Informal listening test indicates that the output speech maintains high quality and intelligibility which are much better than that of the 2400 bps LPC-10e vocoder, and is comparable to the 4800 bps FS1016 CELP. Another advantage of this LPC vocoder is that the bit rate can further be reduced to 1200 bps. One possible bit allocation is as follows: 20~22 bits for LSP parameters, 7 bits for pitch, 2 bits for v/uv transition, 5 bits for gain, with 25ms~30ms frame size. This research scheme is currently ongoing.

REFERENCES

[1] J. Makhoul , R. Viswanathan, R. Schwartz, and A. W. F. Huggins, " A Mixed -Source Model for Speech Compression and Synthesis," J. Acoust. Soc. Amer., Vol.64, pp.1577-1581, Dec 1978.

[2] Alan V. McCree and Thomas P. Barnwell III, "A Mixed Excitation LPC Vocoder Model for Low Bit Rate Speech Coding," IEEE Trans. Speech and Audio Processing, Vol.3, No.4, pp.242-250, July 1995.

[3] Alan V. McCree, Kwan Truong, E. Bryan George, Thomas P. Barnwell III and Vishu Viswanathan, "A 2.4 kbit/s MELP coder candidate for the new U.S. Federal Standard," Proc. IEEE ICASSP-96, pp.200-203, May 1996.

[4] Daniel W. Griffin and Jae S. Lim, "Multiband Excitation Vocoder," IEEE Trans. Acoust., Speech and Signal Processing, Vol.Assp-36, pp.1223-1235, August 1988.

[5] DVSI, "Inmarsat-M Voice Codec, Version 3.0," Inmarsat-M Specification, Inmarsat, August 1991.

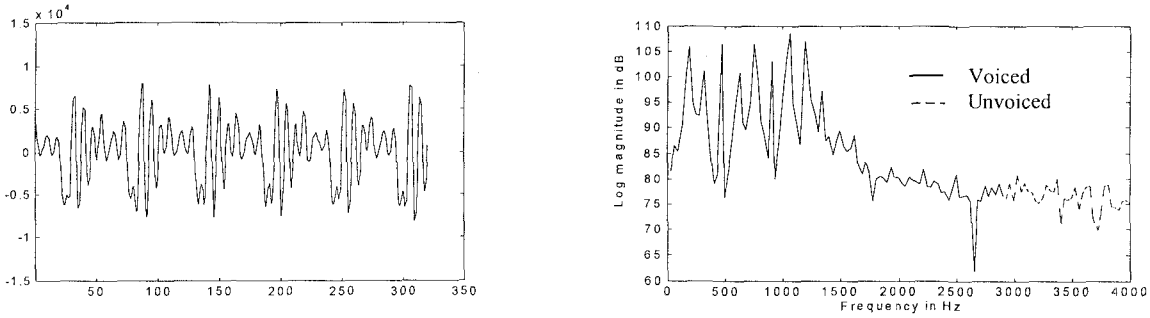
[6] P. C. Meuse, "A 2400 BPS Multi-Band Excitation Vocoder," Proc. IEEE ICASSP-90, pp.9-12, April 1990.

[7] M. Nishiguchi, J. Matsumoto, R. Wakatsuki and S. Ono, "Vector Quantized MBE with Simplified V/UV Division at 3.0 KBPS," Proc. IEEE ICASSP-93, pp.151-154, April 1993.

[8] Alan V. McCree, "A new LPC vocoder model for low bit rate speech coding," Ph.D Dissertation, Georgia Institute of Technology, 1992.

[9] A. M. Kondoz, "Digital Speech Coding for Low Bit Rate Communication Systems," Chichester: J. Wiley, pp.256-268, 1994.

[10] K. K. Paliwal and B. S. Atal, "Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame," IEEE Trans. Speech and Audio Processing, Vol.1, No.1, pp.3-14, Jan. 1993.



(a) (b)

Figure 2. Sustained vowel /a:/ (a) waveform (b)Fourier spectrum

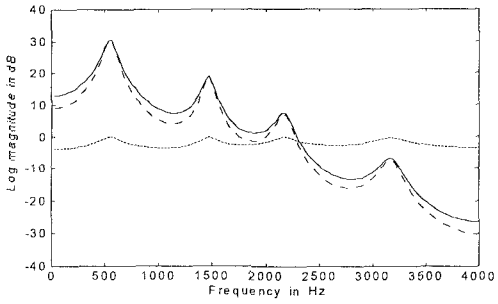


Figure 3. Postfiltering

—— Original LPC envelope

····· Postfilter response

---- Postfiltered envelope

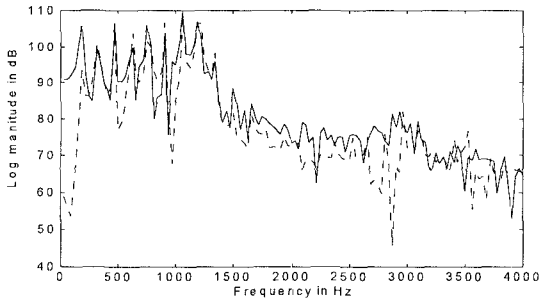


Figure 4. Spectrum comparison

—— Original ---- Synthesized