A mixed excitation LPC vocoder operating at very low bit rate

J.S.Mao, S.C.Chan and K.L.Ho
Department of Electrical and Electronic Engineering
The University of Hong Kong
Pokfulam Road, Hong Kong

ABSTRACT

This paper presents a 1.2 kb/s mixed LPC vocoder based on MultiBand Excitation (MBE) model. The vocoder extracts the pitch by a robust and efficient tracking algorithm. The cut-off frequency, which is a fundamental parameter in mixed excitation system, is obtained by the v/uv decision of MBE analysis. In order to reduce the bit rate, the coder has used frame interpolation between neighboring frames. A fast and reliable linear interpolation algorithm is proposed. Informal listening tests indicate that for either clean speech or telephone speech, the synthesized speech sounds natural and intelligible, and the quality is better than that of 2.4 kb/s LPC-10e standard.

1 INTRODUCTION

In very low bit rate speech coding below 2.4 kb/s, frame parameters are restricted to as few as about 20-30 bits per frame. Therefore not all of the characteristic parameters can be preserved after coding. At this low bit rate, some vocoders such as CELP cannot work well, while other basic vocoders are still possible to provide intelligible speech quality. One of the examples is LPC vocoder, which can even operate at 800 BPS [1]. But the quality is not good enough when compared with 2400 BPS LPC-10. Another example is mixed LPC vocoder [2]. The output speech is a sum of voice excited source and a noise-like excited source. The voicing degree is controlled by the cut-off frequency detected in the speech spectrum. The buzziness of the output speech can be significantly reduced and the naturalness of the speech will be preserved by proper mixing of these two sources. Recently, an improved mixed excitation LPC vocoder which is operated at 2.4 kb/s is proposed by McCree and Barnwell III [3]. More bits are used in transmitting voicing information to mimic more characteristics of natural human speech. The quality is reported to be close to that of the 4.8 kb/s FS1016 CELP coder. Very low bit rate vocoders were also developed based on MultiBand Excitation vocoder [4][5], etc. Low bit rate speech coding below 1.2 kb/s using frame interpolation techniques has been reported in [6]. However, the large frame size (25ms/frame) degrades the performance of the LPC filter, and the computation complexity becomes much larger.

This paper proposes a new 1200 BPS mixed excitation LPC vocoder. It predicts the cut-off frequency based on MBE v/uv decision, and uses mixed excitation to synthesize the output speech. An improved linear frame interpolation between neighboring frames is also introduced, with small additional computation complexity compared with the 2400 BPS MBE vocoder.

This paper is organized as follows: in section 2 the robust and efficient pitch detection algorithm is discussed. In section 3, the voicing analysis and encoding will be described. Section 4 is devoted to the frame parameters interpolation algorithm. The results of the informal listening tests are given in section 5.

2 PITCH ESTIMATION

Pitch estimation is performed in two stages. The first stage is the detection of initial pitch. In this paper, we calculate the speech auto-correlation function $R_w(l)$ from the speech spectrum:

$$R_w(l) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|S_w(e^{j\omega})\right|^2 e^{j\omega l}d\omega \qquad (1)$$

where $S_w(e^{j\omega})$ is the spectrum of the windowed speech which passes through the inverse LPC filter and the l is the time delay. (1) can be efficiently realized by FFT and IFFT. The inverse LPC filter flattens the output spectrum and makes the pitch prediction more reliable. After normalized by $R_w(0)$, $R_w(l)$ is used to extract the initial pitch P which is the time delay corresponding to the maximum autocorrelation value. If $R_w(P)$ < 0.4, the frame is declared as unvoiced and pitch refinement is unnecessary. Otherwise, it is taken as voiced. To avoid pitch doubling and halving, further checking of pitch at sub-multiples and multiples is necessary [7]. Furthermore, one frame look-ahead and one frame look-back pitch tracking algorithms are employed. Figure 1 shows a pitch tracking path and its value $R_w(P)$ in a female speech sentence.

The second stage is the pitch refinement based on MBE principle. The search range is [P-5, P+5], and the search resolution is up to quarter sample. The candidate of fundamental frequency ω_0 which results in the minimum value of error function $E_r(\omega_0)$, is selected as the refined ω_0 [8]:

$$E_R(\omega_0) = \sum_{m=1}^{N}\left|S_w(\omega) - \hat{S}_w(m,\omega_0)\right|^2 \qquad (2)$$

0-7803-3777-8/97/$10.00 © 1997 IEEE

where $S_w(\omega)$ is the windowed original speech spectrum, $S_w(m,\omega_0)$ is the synthetic spectrum with fundamental frequency candidate ω_0, and N is the length of FFT. In this vocoder, ω_0 will be quantized in 7 bits .

3 VOICING ANALYSIS

The cut-off frequency is predicted by MBE harmonic v/uv decision [8]. For each harmonic, a normalized error between the original and the synthesized spectrum is calculated:

$$\varepsilon_k = \frac{\int_{am}^{bm} G(\omega)\left|\left|S_w(\omega)\right| - \left|A_m\right| \cdot \left|E_w(\omega)\right|\right|^2 d\omega}{\int_{am}^{bm} G(\omega)\left|S_w(\omega)\right|^2 d\omega} \qquad (3)$$

Where ε_k is the normalized error of the k^{th} harmonic. The interval (a_m, b_m) is an interval with width which is three times that of the fundamental frequency and is centered at the k^{th} harmonic. $G(\omega)$ is a frequency weighting function. $S_w(\omega)$ is the hamming windowed original spectrum. A_m is the synthesized spectral envelope at the k^{th} harmonic, and $E_w(\omega)$ is the excitation spectrum. If ε_k is below 0.5, this harmonic is declared as voiced, otherwise it is treated as unvoiced. After all of the harmonics are v/uv decided, the frequency of the first unvoiced harmonic is treated as the cut-off frequency F_c. Below F_c, the output speech is synthesized by a series of voiced harmonics, while the output speech above F_c is excited by noise source. The overall output is the sum of these two sources.

Since small changes in F_c do not seem to be perceptible, it is possible to quantize F_c with 3 bits. Suppose the sampling rate is 8 kHz, we use the uniform quantization from 0 Hz to 4 kHz with 8 levels, i.e. 500 Hz between intervals. We select the first level as full-unvoiced, and set the last level as full-voiced state. Through experiments we find that (3) provides a reliable v/uv detection and the prediction of the cut-off frequency F_c is reasonable. Figure 2 displays the extraction of F_c in a typical speech spectrum.

4 FRAME PARAMETERS INTERPOLATION

To reduce the bit rate to as low as 1.2 kb/s, we use both frame interpolation and frame rate reduction. Since short time speech is a non-stationary signal, analysis with larger frame size may not track the pitch very well. For the same reason, the 10th short- term LPC filter cannot model the speech accurately. So we choose 22.5 ms as the frame size in our vocoder. The bits per frame are cut down to 27 bits. In odd frames, all the parameters (LSFs, pitch, energy and v/uv decision) are quantized and transmitted. While in even frames, only pitch and v/uv decision are directly quantized and transmitted to the receiver. The LSF parameters and energy have to be interpolated from the previous and next frame parameters. The interpolation index is transmitted to the receiver instead of the LSFs and energy. In the receiver, the LSFs and energy are reconstructed by interpolation. The accumulative spectral distortion (dB) is frequently used as the performance measure:

$$SD = \sqrt{\frac{1}{F_s}\int_0^{F_s}\left|10\log_{10}\frac{P(f)}{\hat{P}(f)}\right|^2 df} \qquad (4)$$

Where F_s is the sampling rate, $P(f)$ and $\hat{P}(f)$ represent the power spectra of the original and the interpolated LPC filters, respectively. Here $\hat{P}(f)$ is linearly interpolated by the previous and the next LSFs vectors as follows:

$$lsf_j(i) = lsf_{j-1}(i) + \left[lsf_{j+1}(i) - lsf_{j-1}(i)\right]\frac{l}{K-1}$$

$$l = 0,1,2,\ldots,K-1 \qquad (5)$$

where K is an integer number which is a power of 2, $lsf_j(i)$ is the i^{th} LSF in the j^{th} frame. The index with minimum SD in LSFs interpolation is selected as the interpolation index $k_{\min}$, which is also used to reconstruct the energy E_j:

$$E_j = E_{j-1} + (E_{j+1} - E_{j-1})\frac{k_{\min}}{K-1} \qquad (6)$$

Since (4) has to be calculated K times per frame, it requires a lot of computation. Here, we replace (4) by a more efficient LSFs distance measure [9]:

$$Dist(f,\hat{f}) = \sum_{i=1}^{10}\left[g_i w_i (f_i - \hat{f}_i)\right]^2 \qquad (7)$$

where $f_i, \hat{f}_i$ are the i^{th} LSFs in the original and interpolated LSFs vectors, respectively, w_i is a weighting function derived from the LPC spectral envelope:

$$w_i = \left|P(f_i)\right|^{0.2} \qquad (8)$$

and g_i is the perceptual weight given by:

$$g_i = \begin{cases} 1.0 & 1 \le i \le 8 \\ 0.8 & i = 9 \\ 0.4 & i = 10 \end{cases} \qquad (9)$$

Thus, the LSFs which are near the spectral formants will have more weights than those that are in the spectral valleys. The perceivable distortion will therefore be greatly reduced. The LSFs interpolation has been tested in 8000 frames of clean speech and telephone speech, respectively. Table 1 compares the two distortion measures in frame interpolation:

database	method	SD (dB)	> 2 dB (%)
clean speech	original	1.13	6.94
	proposed	1.18	7.11
telephone speech	original	1.26	4.95
	proposed	1.32	5.25

Table 1. Spectral distortion in LSFs interpolation

From table 1, it can be concluded that there are no significant differences between the two algorithms, either in avg. SD (dB) or outliers (>2 dB). But the proposed measure is much easier to implement. The slight increase in the proposed measure is due to the fact that the linear interpolation cannot always match the original LPC spectral envelope when frame size is large. Figure 3 shows an example of LSFs linear interpolation. Table 2 shows the bit allocation of the proposed 1.2 kb/s mixed excitation LPC vocoder using frame interpolation.

	Bits/frame (22.5 ms)	
parameters	odd frame	even frame
10 LSFs	24	—
pitch	7	7
energy	5	—
k_{min}	—	5
v/uv	3	3
total	54 bits/ 2 frames=1.2 kb/s	

Table 2. Bit allocation of 1.2 kb/s mixed excitation LPC vocoder

5 INFORMAL LISTENING TESTS

The performance of the 1.2 kb/s mixed LPC vocoder is evaluated by informal listening tests together with the 2.4 kb/s LPC-10e vocoder. 50 male and female sentences are selected from clean and telephone speech database. The 1.2 kb/s and 2.4 kb/s synthesized speech are played twice, and the listeners are required to give their preferences. Table 3 gives the overall preferences of these two vocoders.

Preference (%)	1.2 kb/s mixed LPC	2.4 kb/s LPC-10e	not sure
clean speech	50	30	20
telephone	62	24	14

Table 3. Speech quality preference

It can be seen from table 3 that for clean speech, the preference on the 1.2 kb/s mixed LPC vocoder is slightly better than that of the 2.4 kb/s LPC-10e. While for telephone speech, the preference for the 1.2 kb/s vocoder is much higher than that of the 2.4 kb/s LPC-10e. It is also found that the encoded male speech sounds more intelligible than that of the female due to the larger frame size.

6 CONCLUSION

This paper presents a new mixed excitation LPC vocoder operating at 1.2 kb/s. It emphasizes on mixed LPC modeling and frame parameters interpolation in very low bit rate speech coding. An efficient and reliable pitch estimation and cut-off frequency detection based on MBE principles are discussed. A fast linear interpolation algorithm is also proposed, which has a performance close to the one using spectral distortion. Informal listening tests show that this vocoder performs better than the 2.4 kb/s LPC-10e, provides natural-sounding and intelligible speech. This vocoder is currently being implemented in a 40 MHz floating point TMS320C31, and requires about 17 MIPS.

7 ACKNOWLEDGMENT

The authors would like to thank the Oregon Graduate Institute's Center for Spoken Language Understanding for providing the materials of multilanguage telephone speech, and Mr. Kiu Li for technical support.

REFERENCES

[1] D. Y. Wong, B. H. Juang, and A. H. Gray Jr., "An 800 bit/s Vector Quantizatoin LPC vocoder," IEEE Trans. ASSP, Vol. 30, pp.770-780, October 1982

[2] 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.

[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] Tian Wang, Kun Tang , Chongxi Feng, "A High Quality MBE-LPC-FE Speech Coder at 2.4 kbps and 1.2 kbps," Proc. IEEE ICASSP-96, pp.208-211, May 1996.

[6] S. Yeldener, A. M. Kondoz and B. G. Evans, "Multiband linear prediction speech coding at very low bit rates," IEE Proc. -Vis. Image Signal Process., Vol. 141, No.5, pp.289 - 296, October 1994.

[7] W. Hess, *Pitch Determination of Speech Signals: algorithms and devices,* Berlin: Springer, 1983.

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

[9] 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.

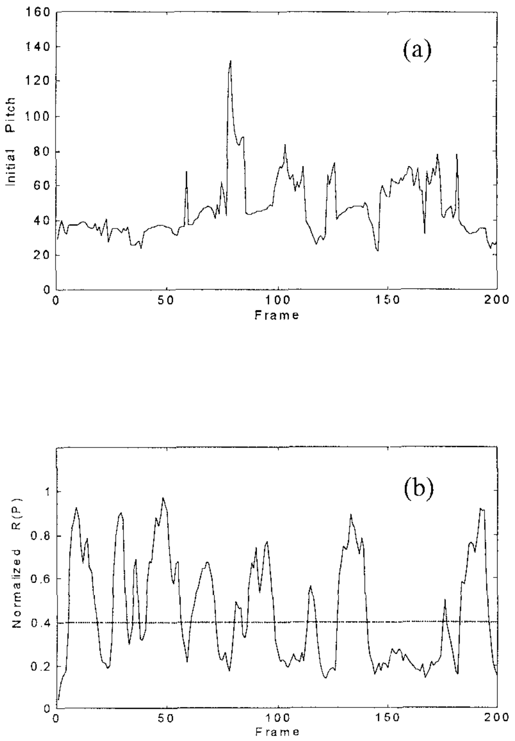

Figure 1 The variation of (a) pitch lag and (b) its auto-correlation value with time

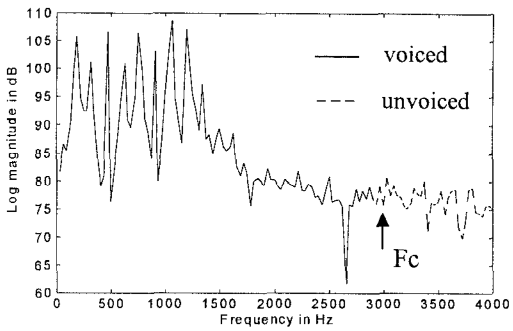

Figure 2 Cut-off frequency extraction

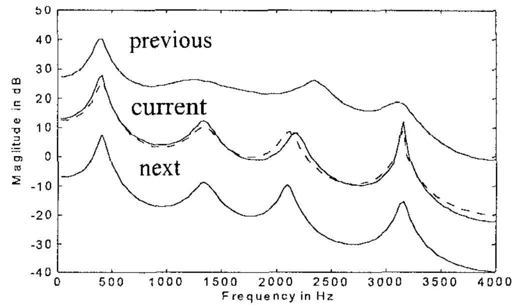

Figure 3 LPC spectral envelope interpolation
——— original ----· interpolated