ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

HIGH QUALITY HARMONIC EXCITATION LINEAR PREDICTIVE SPEECH CODING AT 2 KB/S

Changchun Bao[1] *Jason Lukasiak*[2] *Christian Ritz*[2]

[1]Speech and Audio Signal Processing Lab, Beijing University of Technology, Beijing 100022
baochch@bjut.edu.cn

[2] Whisper Lab, TITR, University of Wollongong, NSW 2522, Australia
jasonl@elec.uow.edu.au chritz@elec.uow.edu.au

ABSTRACT

This paper• presents a high quality harmonic excited linear predictive (HE-LPC) speech coder operating at 2 kb/s based on a harmonic excitation model with two bands. The system incorporates novel features such as: combined pitch detection, residual harmonic matching voicing determination, extraction and interpolation of residual harmonic magnitudes. Subjective listening tests indicate that this coder has same quality as that of Federal Standard MELP coder at 2.4 kb/s whatever training database is from Chinese or English.

1. INTRODUCTION

Most current speech coders operating below the bit rates of 4kb/s fall into one of two categories: the linear prediction based techniques such as, the Mixed Excitation Prediction (MELP) [1], the Harmonic & Stochastic Excitation (HSX) algorithm [2], and the harmonic excitation LPC (HE-LPC) [3], and frequency domain techniques such as Waveform Interpolation (WI) [4] and the Multi-Band Excitation (MBE) [5]. These vocoders are capable of producing good quality speech between 2.4 kb/s and 4 kb/s. Below 2.4 kb/s, however, these coders suffer from distortions introduced by coarse quantization of model parameters due to the limited number of bits.

Harmonic Excitation Linear Speech Coding (HE-LPC) has the potential of producing good quality speech at very low bit rates (2.4kb/s and below). This coding scheme uses the advantage of both time domain (LPC based) and frequency domain techniques to improve the speech quality. In this paper, we are reporting on a 2 kb/s HE-LPC coder that provides same intelligible and natural quality speech as MELP vocoder operating at 2.4 kb/s.

2. HE-LPC SPEECH CODER

The simplified block diagram of the HE-LPC coder is shown in figure 1.

2.1 LP analysis and quantization

Similar to [3], a 10th order Linear Predictive (LP) analysis is performed for each 200 samples frame using a Hamming window of length is 240 samples at 8 kHz sampling rate. The resulting LP coefficients are converted to the LSF domain. A vector of 10 LSFs is divided into three sub-vectors of dimensions 3, 3 and 4 and quantized with 30 bits using split vector quantization (SVQ) technique [6].

In order to preserve the ordering property in the quantized LSFs, a sequential search is conducted in this paper. First of all, the second and the third codebook entries based on their first element are ordered from minimum to maximum, respectively. Secondly, the first LSF sub-vector is quantized. Thirdly, the second LSF sub-vector is quantized through searching those entries that its first element is larger than the third element in the first quantized sub-vector. Finally, the third LSF sub-vector is quantized through searching those entries that its first element is larger than the third element in the second quantized sub-vector. Hence, the stability of the LP filter can be guaranteed.

For our SVQ of the LSF at 30 bits/frame, the SD is 0.72dB, the percentages of frames with SD larger than 2dB but less than 3dB is 0.26% and the percentages of frame with SD larger 4 dB is 0%. These results indicate our SVQ scheme of the LSFs achieves transparent quality [6].

2.2 Pitch detection

Pitch detection is based on the normalized cross-correlation coefficient ρ defined in reference [3]:

$$\rho = \frac{\sum_{n=0}^{N-1} s_p(n)s_p(n-\tau)}{\sqrt{\sum_{n=0}^{N-1} {s_p}^2(n)\sum_{n=0}^{N-1} {s_p}^2(n-\tau)}} \quad (1)$$

The normalized cross-correlation is calculated independently over two overlapping windows. The first window comprises the entire current frame. The second window comprises the second half of the current frame and the first half of the look-ahead frame. In equation (1), τ is an integer value representing the delay between 20 and 128 samples, N is the length of the frame and $s_p(n)$ is the speech signal that is pre-processed by DC removing, low-pass filtering and numerical filtering.

Similar to reference [7], after finding the optimal delay for each window, we can use the following thresholds and logic defined in (2) to combine the optimal delays from the two windows to obtain a more reliable delay estimate for the current frame. If (τ_1, ρ_1), and (τ_2, ρ_2) are the optimal delays and the corresponding normalized cross-correlation coefficients found for the two overlapped windows, respectively, the final delay

•This work was supported by the National Natural Science Foundation of China under grant 60372063, and the Natural Science Foundation of Beijing under grant 4042009.

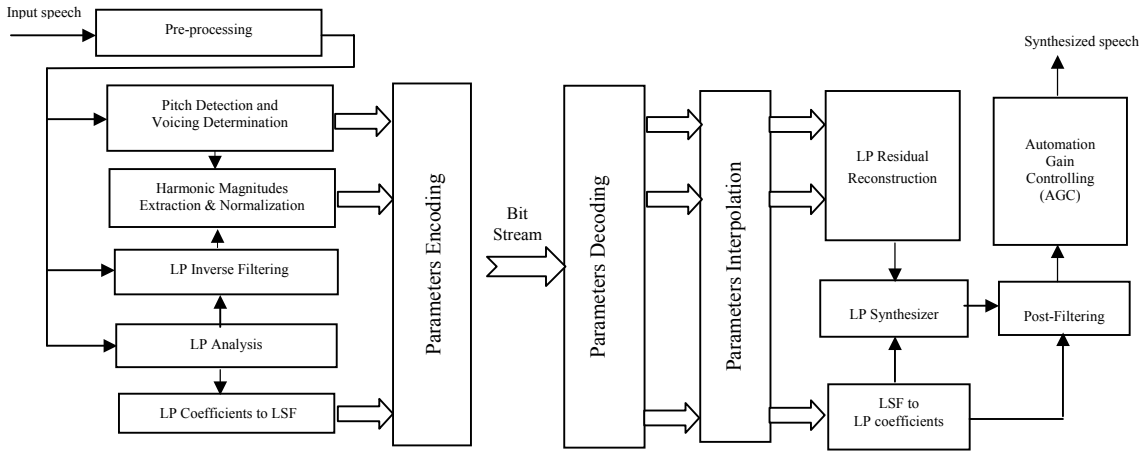

Figure 1: Simplified block diagram of HE-LPC speech coder

estimate τ_{opt} is obtained by

$$\tau_{opt} = \begin{cases} \tau_1, & \rho_1 > \rho_2 + 0.4 \ \ and \ \ |\tau_1 - \tau_2| > 15 \\ (\tau_1 + \tau_2)/2, & \rho_1 > \rho_2 + 0.4 \ \ and \ \ |\tau_1 - \tau_2| \le 15 \\ \tau_1, & else \end{cases} \quad (2)$$

where, ρ_1 and ρ_2 are considered as the confidence values indicating how reliable the corresponding pitch estimates are. For example, if ρ_1 is much larger than ρ_2, it indicates that the τ_1 estimate is more reliable than τ_2. Using this reliability information, preferences can be given to the look-ahead frame estimates (τ_2, ρ_2). The performance of the coder at voicing onsets is improved. The pitch delay is directly quantized with 7 bits.

2.3 Voicing decision

In HE-LPC coder, the cut-off frequency that separates the residual into periodic (below) and stochastic (above) is determined by the frequency that maximizes the voicing probability p_v. Below this cut-off frequency, the speech is declared as voiced while the harmonic above this frequency is declared as unvoiced. In this paper, p_v is estimated based on the energy of the low-pass filtered speech, E_{lpf}, normalized cross-correlation coefficient ρ that is defined in equation (1), and the harmonic matching coefficient ρ_H defined as:

$$\rho_H = \frac{\sum_{n=0}^{N-1} e(n)\, e_H(n)}{\sqrt{\sum_{n=0}^{N-1} e^2(n) \sum_{n=0}^{N-1} e_H{}^2(n)}} \quad (3)$$

Where $e(n)$ is linear predictive residual, and $e_H(n)$ is harmonic-based linear predictive residual synthesized by:

$$e_H(n)_{0 \le n \le N-1} = 2.0 * \sum_{m=1}^{K} \left(C_m \cos\left(\frac{2\pi m n}{\tau_{opt}} \right) + D_m \sin\left(\frac{2\pi m n}{\tau_{opt}} \right) \right) \quad (4)$$

In equation (4), C_m and D_m are the Discrete Fourier series coefficients corresponding to harmonic peaks in residual domain. The notation K indicates the number of harmonics and $K = \lfloor \tau_{opt} \rfloor / 2$. Considering the transition frames and unvoiced frame, ρ_H is determined as:

$$\rho_H = \begin{cases} \rho_{H_1}, & |\rho_{H_1} - \rho_{H_0}| > 0.4 \\ 0.5 * (\rho_{H_0} + \rho_{H_1}), & |\rho_{H_1} - \rho_{H_0}| \le 0.4 \end{cases} \quad (5)$$

Where ρ_{H_1} and ρ_{H_0} indicate current and previous harmonic matching coefficients. Once energy E_{lpf}, normalized cross-correlation coefficient ρ and the harmonic matching coefficient ρ_H are determined, voicing probability p_v is obtained by the flow chart given in Fig. 2.

In Fig. 2, T_e is the energy threshold that depends on the dynamic range of the input speech that has been pre-processed, T_ρ is the threshold of the normalized cross-correlation coefficient ρ, T_{H_1}, T_{H_2} and T_{H_3} are the thresholds of the harmonic matching coefficient ρ_H. T_ρ, T_{H_1}, T_{H_2} and T_{H_3} are determined by subjective listening tests. In order to reduce buzz in synthesized, optimized pitch τ_{opt} is set to 80 samples at 8khz sampling rates when $p_v = 0$. Voicing probability p_v is quantized with 2 bits.

2.4 Extraction and Quantization of Harmonics

The quantized LP coefficients are used to find the LP residual required for determination of the excitation harmonic amplitudes. A Hamming window with $N_w = 300$ samples is used for estimating residual harmonic amplitudes. The center of the window lies in the middle of current frame and covers 50 samples in the previous frame and 50 samples in the future frame. The reason for using a relative long window is to better represent harmonic structure for male speakers with low pitch.

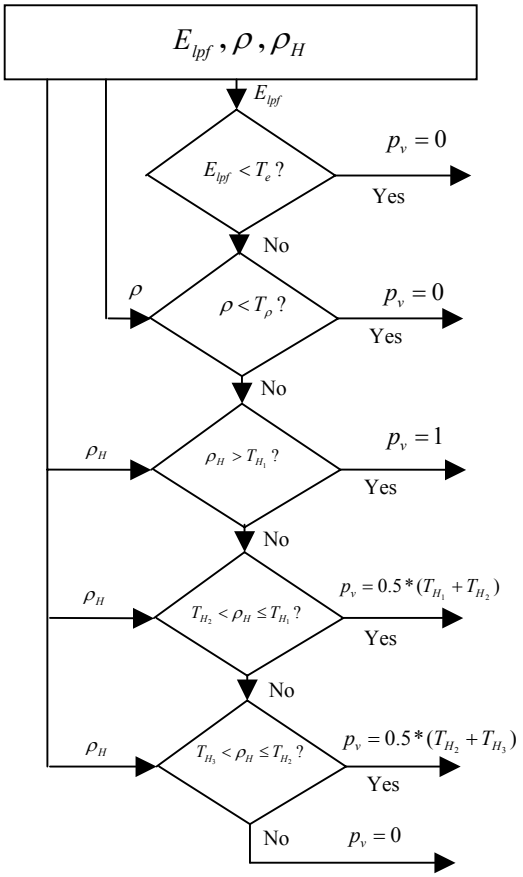

Figure 2: Flow chart of voicing decision

By padding zeros to the windowed residual signal, a 512-point FFT is used. First, harmonic peak estimation of the LP residual is conducted with the optimal pitch τ_{opt}. Second, the Discrete Fourier series coefficients, C_m and D_m, corresponding to harmonic peaks are derived from FFT coefficients as follows

$$\begin{cases} C_m = 2 * R(i)/N_w \\ D_m = -2 * I(i)/N_w \end{cases}, i = HP(m), m = 1, 2, \cdots, K \qquad (6)$$

where $R(i)$ and $I(i)$ indicate the real part and image part of FFT, $HP(m)$ indicates m^{th} harmonic peak position corresponding to i^{th} index of FFT from 1 to 256. Third, the harmonic amplitudes are normalized by its power as:

$$\begin{cases} \hat{C}_m = C_m / \sqrt{E_n} \\ \hat{D}_m = D_m / \sqrt{E_n} \end{cases}, \quad m = 1, 2, \cdots, K \qquad (7)$$

where E_n is the harmonics power that is calculated as:

$$E_n = 0.5 * \sum_{m=1}^{K} \left(C_m^2 + D_m^2 \right) \qquad (8)$$

The harmonic magnitudes with unit power are calculated as:

$$A_m = \sqrt{\hat{C}_m^{\,2} + \hat{D}_m^{\,2}}, \quad m = 1, 2, \cdots, K \qquad (9)$$

The main motivation of this normalization is to separate the power and shape of harmonic amplitudes so that they can be quantized separately to achieve higher coding efficiency.

Since the dimension of harmonic amplitude varies with the pitch, the variable dimension vector quantizer (VDVQ) will be utilized for spectrum quantization. Fortunately, the normalized harmonic amplitudes in residual domain are nearly flat. We can truncate harmonic amplitudes to get a fixed dimension vector. In the receiver, the truncated harmonic amplitudes are estimated as the average of transmitted harmonic magnitudes. With this method, the LP harmonic amplitudes vector can be reduced to as short as 10 dimension. Informal subjective listening test indicates that the speech quality does not degrade by using this truncation before harmonic magnitudes quantization. This 10 dimension harmonic codebook is vector quantized with 6 bits. The power is scalar quantized with 5 bits.

2.5 Parameters Interpolation

At the receiving end, the recovered LSFs between successive frames are linearly interpolated into the 4 sub-frames to ensure a smooth transition. The voicing probability is directly linearly interpolated between the previous frame and the current frame. In order to prevent pitch doubling and halving, the pitch is interpolated similarly to [4] and [7] as:

$$P(n)_{\tau_{opt2} \ge \tau_{opt1}} = \begin{cases} \eta\tau_{opt1} + n(\tau_{opt2} - \tau_{opt1})/(\eta(N-1)), & 0 \le n < N/2 \\ \eta\tau_{opt1} + n(\tau_{opt2} - \tau_{opt1})/(N-1), & N/2 \le n < N-1 \end{cases}, \ \eta = \left\lfloor \frac{\tau_{opt2}}{\tau_{opt1}} \right\rceil \qquad (10)$$

$$P(n)_{\tau_{opt2} < \tau_{opt1}} = \begin{cases} \tau_{opt1} + n(\eta\tau_{opt2} - \tau_{opt1})/(N-1), & 0 \le n < N/2 \\ \tau_{opt1} + n(\eta\tau_{opt2} - \tau_{opt1})/(\eta(N-1)), & N/2 \le n < N-1 \end{cases}, \ \eta = \left\lfloor \frac{\tau_{opt1}}{\tau_{opt2}} \right\rceil \qquad (11)$$

In equations (10) and (11), $P(n)$ is interpolated instant pitch, τ_{opt1} and τ_{opt2} are the pitch values for previous frame and current frame, N is the length of the frame, η is the ratio between τ_{opt1} and τ_{opt2} rounded to the nearest integer, which can be considered as an indicator showing whether the pitch has doubled or halved. The harmonic magnitudes shaped by power are linearly interpolated through padding zero to the shorter harmonics or inserting zero between the shorter harmonics. This kind of interpolation is based on following two group harmonic magnitudes with same dimension:

$$A_1(m)_{\substack{\tau_{opt2} \ge \tau_{opt1} \\ 1 \le m \le K_2}} = \begin{cases} 0, & m \ne \eta * i \\ F_1(i), & m = \eta * i \end{cases}, \ \eta = \left\lfloor \frac{\tau_{opt2}}{\tau_{opt1}} \right\rceil, \ i = 1, 2, \cdots, K_1 \qquad (12\text{-}1)$$

$$A_2(m)_{\substack{\tau_{opt2} \ge \tau_{opt1} \\ 1 \le m \le K_2}} = F_2(i), \ i = m = 1, 2, \cdots, K_2 \qquad (12\text{-}2)$$

or

$$\underset{\substack{\tau_{opt2}<\tau_{opt1}\\ 1\le m\le K_1}}{A_1(m)} = F_1(i),\ i=m=1,2,\cdots,K_1 \tag{13-1}$$

$$\underset{\substack{\tau_{opt2}<\tau_{opt1}\\ 1\le m\le K_1}}{A_2(m)} = \begin{cases} 0, & m \neq \eta * i \\ F_2(i), & m = \eta * i \end{cases},\ \eta = \left\lfloor \frac{\tau_{opt1}}{\tau_{opt2}} \right\rfloor,\ i=1,2,\cdots,K_2 \tag{13-2}$$

where $F_1(i)$ and $F_2(i)$ are the recovered harmonic magnitudes from previous frame and current frame, K_1 and K_2 are the harmonic number for previous frame and current frame, and $A_1(m)$ and $A_2(m)$ are the transformed harmonic magnitudes by padding or inserting zeros based on $F_1(i)$ and $F_2(i)$, respectively.

2.6 Speech Synthesis

In the HE-LPC coder, the excitation signal $e(n)$ that is specified by pitch, its harmonic magnitudes and voicing probability is given by the following sinusoidal model:

$$e(n) = \sum_{m=1}^{K(n)} A_m(n)\cos(m\phi(n) + \theta_m(n)) \tag{14}$$

where

$$\theta_m(n) = \begin{cases} D(m), & m \le K(n)p_v(n,m) \\ U(m), & m > K(n)p_v(n,m) \end{cases} \tag{15}$$

$K(n)$, $p_v(n,m)$, $A_m(n)$ and $\theta_m(n)$ are the interpolated number of harmonics, interpolated voicing probability, the interpolated m^{th} harmonic amplitude and the estimated m^{th} harmonic phase in the instant of n, respectively. A fixed spectrum phase spectrum $D(m)$ from a voiced segment generated by a high-pitched male speaker is used below the cut-off frequency and a random phase $U(m)$ distributed uniformly on the interval $[-\pi, \pi]$ is used above the cut-off frequency. The phase track $\phi(n)$ is computed by incrementally summing the area under the frequency track curve $F(n) = 1/P(n)$. The reconstructed residual signal $e(n)$ is used to excite the LP synthesis filter to obtain synthesized speech. Then, a traditional post-filter as used in CELP coder is added to enhance speech.

3. 2 KB /S CODER BIT ALLOCATION

For operation at 2kb/s, a frame length of 25ms (200 samples at 8 kHz sampling rates) is used. Therefore, 50 bits/frame are available for coding the model parameters as listed in Table 1.

Table 1: Bit allocation for 2 kb/s HE-LPC coder

Parameters	Bits/Frame	Bit Rate (b/s)
LSF	30	1200
Pitch	7	280
Gain	5	200
Voicing	2	80
1~10 Harmonics	6	240
Total	50	2000

4. SUBJECTIVE TEST RESULTS

To evaluate the performance of the 2kb/s HE-LPC coder, we have conducted an informal subjective A/B test. The codebooks in HE-LPC coder are trained by Chinese and English database, respectively. Sixteen listeners compared the 2kb/s HE-LPC coder with the 2.4kb/s MELP vocoder. Sixteen sentences in Chinese spoken by 8 male and 8 female speakers were used for listening test. The test results are listed in Table 2 corresponding to Chinese codebook training database and in Table 3 corresponding to English codebook training database. , From these two tables, we can see that the subjective quality of the 2 kb/s HE-LPC is identical to the Federal Standard 2.4 kb/s MELP vocoder

Table 2: A/B test results for Chinese training database

Test	2 kb/s HE-LPC	2.4 kb/s MELP	No preference
Male	26.92%	27.88%	45.19%
Female	25.89%	24.11%	50%
Total	26.41%	25.99%	47.60%

.

Table 3: A/B test results for English training database

Test	2 kb/s HE-LPC	2.4 kb/s MELP	No preference
Male	32.03%	33.59%	34.38%
Female	28.91%	26.56%	44.53%
Total	30.47%	30.08%	39.46%

5. REFERENCES

[1] A. V. McCree et al., "A 2.4 kb/s MELP Coder Candidate for The New U. S. Federal Standard," IEEE ICASSP-96, pp. 200-203.

[2] C. Laflamme, R. Salami, R. Matmti and J-P. Adoul, "Harmonic-Stochastic Excitation (HSX) Speech Coding below 4kbits/s," IEEE ICASSP-96, pp. 204-207.

[3] Changchun BAO, "Harmonic excitation LPC (HE-LPC) speech coding at 2.3 kb/s." IEEE ICASSP-03, pp. I-784~I-787.

[4] W. B. Kleijn, and J. Haagen, "Waveform Interpolation for Coding and Synthesis, " *in Speech Coding Synthesis by W. B. Kleijn and K. K. Paliwal, Elsevier Science B. V.*, Chapter 5, pp. 175-207, 1995.

[5] D. Griffin, and J. S. Lim, "Multiband Excitation Vocoder," IEEE Trans. ASSP, Vol.36, No. 8, pp. 1223-1235, 1988.

[6] Kuldip K. Paliwal and B. S. Atal, "Efficient Vector Quantizaion of LPC paramters at 24 bit/frame," IEEE Trans. On Speech, and Audio Processing, Vol. 1, No. 1, pp. 3-14, 1993.

[7] Eddie L. T. Choy, Waveform Interpolation Speech Coding at 4 kb/s. M. S. Thesis, McGill University, 1998.