HARMONIC EXCITATION LPC (HE-LPC) SPEECH CODING AT 2.3 KB/S

Changchun Bao

Electronic Information and Control Engineering College
Beijing University of Technology, 100022, P. R. China
baochch@bjpu.edu.cn

ABSTRACT

This paper presents an algorithm for encoding speech signal at 2.3kb/s based on a uniform harmonic modeling of the excitation signal. The algorithm uses the robust pitch detection and efficient voicing analysis to split the LPC excitation into two bands. The lower band is related to the voiced parts of speech, while the upper band represents unvoiced speech. A fixed phase spectrum from a voiced segment generated by a male speaker is added into the uniform harmonic modeling of the excitation signal. This kind of fixed phase reduced the buzz effectively and produced soft natural speech. A short-term post-filter is utilized at the decoder to enhance the quality of synthesized speech. Subjective test in Chinese showed that the 2.3 kb/s HE-LPC coder performance is better than that of federal standard 2.4 kb/s MELP coder.

1. INTRODUCTION

In recent years, high quality speech coding at 2.4 kb/s and below is one of the most interesting topics in speech coding fields because many applications and services in telecommunications and secure speech communication need these low bit rates speech coding urgently. In the past few years, some successful speech coding algorithms, for example, the Waveform Interpolation (WI) algorithm [1], the Multi-Band Excitation (MBE) algorithm [2], the Mixed Excitation Prediction (MELP) [3] algorithm, the Harmonic & Stochastic Excitation (HSX) algorithm [4], and the Split-Band LPC (SB-LPC) algorithm [5], produced more intelligible and more natural speech quality than traditional binary excitation LPC10 vocoder. The common features of these algorithms are that the harmonic and stochastic components of speech signal (or residual signal) are modeled and synthesized separately.

In this paper, we propose a uniform harmonic model, which makes the algorithm insensitive to the errors of unvoiced/voiced decision, to describe the harmonic and stochastic components of linear predictive residual. The robust pitch detection and efficient voicing decision algorithm are used to split the LPC excitation into two bands. For the lower band that is related to the voiced speech, a fixed phase spectrum from a male speaker is added into the uniform harmonic modeling of the excitation signal to remove buzz. In addition, an efficient VQ algorithm for LSFs is used in this coder.

This paper is organized as follows. In section 2, we present the proposed encoding scheme that includes linear prediction and quantization, pitch detection and voicing decision. In section 3, we describe the decoding scheme including harmonic model of excitation signal and bit allocation. In section 4, the subjective test results are given.

2. HE-LPC SPEECH ENCODER

The block diagram of the encoder is shown in Figure 1. In 2.3 kb/s implementation, 20 ms speech frame are used. In addition, it needs one frame of look-ahead for LP analysis and pitch detection. The different parts of the encoding algorithm are described in following subsections.

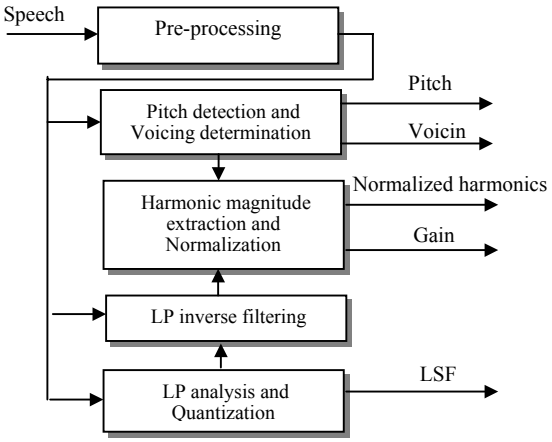

Figure 1: Block diagram of HE-LPC encoder

2.1 LP analysis and quantization

A 10^{th} order Linear Predictive (LP) analysis is performed for each frame using Hamming window which length is

240 samples at the 8kHz sampling rate. The center of window lies at the right boundary of current frame. In other words, the window covers 120 samples in the current frame and 120 samples in the future frame. The auto-correlation method is performed on the windowed speech to generate the filter coefficients**.** A 30 Hz bandwidth expansion is applied by multiplying the LP coefficients with 0.998^k, $k = 1,\cdots,10$. The resulting coefficients are converted to the LSF domain and quantized with the following Prediction-Splitting VQ (PSVQ) approach [6].

The LSF predictive vector in current frame is obtained by the following predictor equation:

$$\widetilde{lsf}_i^{(n)} = a_i \vec{lsf}_{i-1}^{(n)} + b_i \hat{lsf}_i^{(n-1)} \text{ , } i = 1, 2, \cdots, 10 \quad (1)$$

where

$$\vec{lsf}_i^{(n)} = lsf_i^{(n)} - \bar{lsf}_i^{(n)} \text{, } i = 1, 2, \cdots, 10 \quad (2)$$

is the i^{th} unbiased LSF parameter in n^{th} frame, $lsf_i^{(n)}$ is the i^{th} LSF parameter in n^{th} fame, $\bar{lsf}_i^{(n)}$ is the mean of i^{th} LSF parameter, and $\hat{lsf}_i^{(n-1)}$ is the i^{th} unbiased LSF parameter quantized in $(n-1)^{th}$ frame. In equation (1), a_i and b_i are called predictive coefficients and interpolation coefficients, respectively. The coefficients, a_i and b_i can be found out by minimizing following square predictive error:

$$\varepsilon_i = \sum_{n=1}^{M_F} \left[\vec{lsf}_i^{(n)} - \widetilde{lsf}_i^{(n)} \right]^2 \quad (3)$$

where M_F is the total frame number of training data. By setting $\partial \varepsilon_i / \partial a_i = 0$ and $\partial \varepsilon_i / \partial b_i = 0$, the coefficients a_i and b_i can be estimated.

In encoding, the prediction residual vector

$$r_i^{(n)} = \vec{lsf}_i^{(n)} - \widetilde{lsf}_i^{(n)} \text{, } i = 1, 2, \cdots, 10 \quad (4)$$

is split into two vector. The first vector contains first four LSF residuals and the second vector contains remaining 6 LSF residuals. These two residual vectors are trained independently with 10 bits by using LBG method [7] and weighted distortion measure [8]. For PSVQ of the LSF at 20 bits/frame, the SD is 0.94dB, the percentages of frames with SD larger than 2dB but less than 3dB is 1.998% and the percentages of frame with SD larger 4 dB is 0%. These results indicate our PSVQ scheme of the LSF achieve the transparent quality.

The quantized LSF between successive frame are linearly interpolated into the 4 sub-frames to ensure a smooth transition. Each of the interpolated set of LSFs is then converted back into LPC coefficients before they are used for LP inverse filter to compute the LP residual signal in the current frame. In addition, the last interpolated LP coefficients of the current frame will be used to compute 40 residual samples in the future frame. These future residual samples will be used for harmonic magnitude extraction of the current residual frame.

2.2 Pitch detection and voicing decision

Pitch detection is finished once per frame in speech domain. First, the mean of speech signal is removed, and then low-pass filtered using an 800 Hz IIR filter and a 9th order numerical filter. This numerical filter removed the effect of the first formant greatly, and made the pitch detection more reliable. For each frame of the data processed, the pitch detector carries out the calculations independently over three 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, and the third window comprises the entire look-ahead frame. Thus the pitch detector has the 20ms algorithm delay.

Next, the computations of normalized cross-correlation coefficient over all desired delay values are carried out separately in each window. This normalized cross-correlation coefficient, denoted as ρ, is defined as:

$$\rho = \max \left\{ 0, \min \left\{ \frac{\sum_{n=0}^{N-1} s(n)s(n-\tau)}{\sqrt{\sum_{n=0}^{N-1} s^2(n) \sum_{n=0}^{N-1} s^2(n-\tau)}}, 1.0 \right\} \right\} \quad (5)$$

where τ is an integer value representing the delay between 20 and 120 samples, N is the length of the frame and $s(n)$ is the speech signal that is removed mean and low-pass filtered. If the delay value τ corresponds to the true pitch period of the signal or and integer multiplier of that, the corresponding ρ value would be close to 1.0. In contrast, ρ tends to be considerably less than 1.0 for all delays if signal displays no periodic character (unvoiced speech). Therefore, in order to find the true pitch, the delay τ that yield the maximum ρ is searched. This delay will be referred to as the optimal delay.

The delay τ for each window is divided into three ranges, (80,120), (40,79) and (20,39), and the smaller values is used to avoid choosing pitch multiples with a post-processing.

After finding the optimal delay for each window, we can use the following thresholds and logic to combine the optimal delays from the three windows to obtain a more reliable delay estimate for the current frame. If we let (τ_1, ρ_1), (τ_2, ρ_2) and (τ_3, ρ_3) are the optimal delays

and the corresponding normalized cross-correlation coefficients found for the three overlapping windows, respectively, the final delay estimate $\hat{\tau}_{opt}$ is obtained by

$t_1 = \rho_2/\rho_1$, $t_2 = \rho_2/\rho_3$, $t_3 = \rho_1/\rho_2$, $t_4 = \rho_3/\rho_2$

If (t_1 >1.8 and t_2 >1.8 OR t_3 >1.8 and t_4 >1.8)

$\hat{\tau}_{opt} = (\tau_1 + \tau_3)/2$, $\rho = (\rho_1 + \rho_3)/2$

Else

$\hat{\tau}_{opt} = \tau_2$, $\rho = \rho_2$

Note that the value of $\hat{\tau}_{opt}$ is integer. Hence, the pitch detector described above gives only integer pitch value. Indeed, integer pitch with a resolution of one sample for 8 kHz sampling rate is sufficient for the HE-LPC coder implementation. In our coder implementation, the pitch delay is quantized with 7 bit.

In HE-LPC coder, voicing probability p_v that the speech signal is divided into two bands is estimated for each frame no matter what type of speech it is. This p_v is used to determine the cut-off frequency between the periodic and stochastic band. 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} , and normalized cross-correlation coefficients, ρ . If E_{lpf} is less than a given threshold, p_v is directly determined as zero and the pitch frequency is set to 100 Hz. Otherwise, p_v is determined based on the value of ρ . In order to save bits, ρ is quantized with 2 bit based on subjective listening test.

2.3 Quantization of LP Residual Harmonic

The quantized LP coefficients are used to find the LP residual required for determination of the excitation harmonic amplitudes. LP residual signal is transformed into the frequency domain using a 256 point FFT. The harmonic amplitude estimation of LP residual is similar to the SEEVOC [9], but we use an optimal pitch instead of an average pitch. After the harmonic amplitudes are extracted, they are then normalized by power. 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. Consequently, the spectrum has a variable dimension. In general, the harmonic amplitudes are described with an appropriate variable dimension vector quantizer (VDVQ). In our implementation, the pitch is allowed to vary from 20 to 120 resulting in 10 to 60 harmonic amplitudes. If VDVQ techniques are selected, the computational complexity and storage requirement are very high. Fortunately, the LP harmonic amplitudes normalized are nearly flat and tend to unity. We can truncate harmonic amplitudes to get a fixed dimension vector. In the receiver, the truncated harmonic amplitudes are set unity. With this method, the LP harmonic amplitudes vector can be reduced to as short as 10 dimension. This codebook is trained with LBG algorithm and the bit consumption is 9 bits. The power is scalar quantized in logarithmic domain with 8 bits.

3. HE-LPC SPEECH DECODER

The block diagram of the decoder model is shown in Figure 2. The received mode parameters represent the signal at the end of the frame being synthesized and these parameters are interpolated with the ones at the end of the past throughout the synthesis procedure.

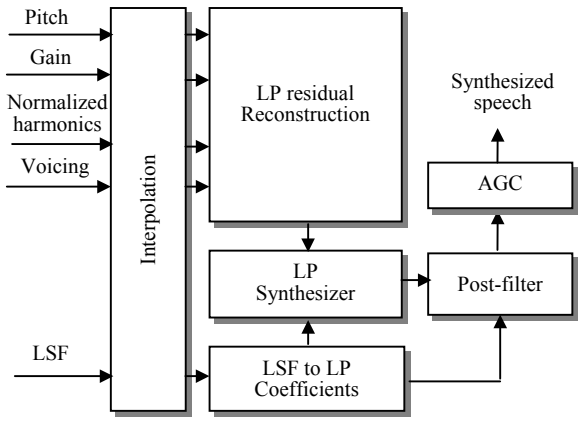

Figure 2. Block diagram of the decoder

3.1 Harmonic Model of Excitation Signal

In the HE-LPC coder, the voiced and unvoiced speech will be described with a uniform harmonic model. No matter what type of speech signal, reconstructed speech is obtained with following excitation:

$$e(n) = \sum_{k=1}^{L(n)} A_k(n)\cos(k\phi(n) + \theta_k(n)) \qquad (6)$$

where

$$\theta_k(n) = \begin{cases} D(k), & k \le L(n)p_v(n,k) \\ U[-\pi,\pi], & k > L(n)p_v(n,k) \end{cases} \qquad (7)$$

$L(n)$, $p_v(n,k)$ and $A_k(n)$ are the number of harmonic, voiced probability and the k^{th} harmonic amplitude at the sample point n, respectively. The phase $D(k)$ is a fixed spectrum. This fixed phase spectrum is drawn from a voiced segment generated by a high-pitched male speaker who can offer more harmonics than a low-pitch speaker.

With the linear interpolation, the pitch value $P(n)$ at every point can be obtained. Thus, the phase track $\phi(n)$ can be computed by incrementally summing the area under the frequency track curve $F(n)$. The relationship between $F(n)$ and $P(n)$ can be expressed as:

$$P(n) = 1/F(n) \tag{8}$$

The phase contour at each sample point can be updated on a per-sample basis by

$$\phi(n) = \phi(n-1) + \int_{n-1}^{n} [2\pi/P(m)]dm \tag{9}$$

where $\phi(n)$ and $\phi(n-1)$ are the current and the previous phase values. The integral corresponds to the incremental area between the interval $n-1$ and n.

3.2 Speech Synthesis and Bit Allocation

The reconstructed residual signal is used to excite the LP synthesis filter to obtain the final speech signal. The reconstructed speech is then post-filtered with a formant post-filter $A(z/\beta)/A(z/\alpha)$ and a tilt compensation filter $1-\mu z^{-1}$ [10], where $\beta = 0.5$, $\alpha = 0.8$ and $\mu = 0.5$. At the output of the post-filter, an automatic gain control procedure is used to ensure the energy of the output signal to be close to the original speech energy. The bit allocation scheme of HE-LPC coder at 2.3 kb/s is given in table 1.

Parameters	Bits/Frame	Bits/Second
LSF	20	1000
Pitch	7	350
Gain	8	400
Voicing	2	100
1~10 Harmonics	9	450
Total	46	2300

Table 1. Bit allocation for 2.3 kb/s HE-LPC coder

4. SUBJECTIVE TEST RESULTS

To evaluate the performance of the 2.3 kb/s HE-LPC algorithm, we have conducted an informal subjective A/B test in Chinese. Eleven listeners compared the 2.3 kb/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. The test results, listed in table 2, indicate that the subjective quality of the 2.3 kb/s HE-LPC is better than federal standard 2.4 kb/s MELP vocoder. The HE-LPC preference is higher for female than for male speakers.

Test	2.3 kb/s HE-LPC	2.4 kb/s MELP	No preference
Female	42.05%	19.32%	38.64%
Male	27.27%	28.41%	44.32%
Total	34.66%	23.86%	41.48%

Table 2. A/B test results

5. REFERENCES

[1] 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.

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

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

[4] 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.

[5] I. Atkinson, S. Yeldener, A. M. Kondoz, "High Quality Split-Band LPC Vocoder Operating at Low Bit Rates, "IEEE ICASSP-97, pp. 1559-1562, 1997.

[6] Bao Changchun and Dai Yisong, "One-step Interpolation Predictive Vector Quantization of LSP Parameters," The Journal of China University of Posts and Telecommunications, Vo.3, No. 1, pp. 21-26, 1996.

[7] Y. Linde, A. Buzo, R. M. Gray, "An algorithm for Vector Quantizer Design," IEEE Trans. Commun. , Com-28, (1): 84-95, 1980.

[8] 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, Jan. 1993.

[9] D. B. Paul, "The Spectral Envelope Estimation Vocoder," IEEE Trans. On Acoust., Speech and Signal Proc., ASSP-29, 1981, pp. 786-794.

[10] J. H. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," IEEE Trans. On Speech Audio Process., vol. 3. No. 1, pp. 59-71, 1995.