A New Fast Pitch Search Algorithm using Line Spectrum Frequency in the CELP Vocoder

CELP보코더에서 Line Spectrum Frequency를 이용한 고속 피치검색

MyungJin BAE*, SangMok SOHN*, HahYoung YOO** Kyung Jin BYUN**
배 명 진*, 손 상 목*, 유 하 영**, 변 경 진**

※논문은 전자통신연구소의 1996년도 수탁과제 연구 지원에 의해 수행되었습니다.

ABSTRACT

Code Excited Linear Prediction(CELP) vocoder exhibits good performance at data rates below 8 kbps. The major drawback of CELP type coders is a large amount of computation. In this paper, we propose a new pitch searching method that preserves the quality of the CELP vocoder reducing computational complexity. The basic idea is that grasps preliminary pitches using the first formant of speech signal and performs pitch search only about the preliminary pitches. As applying the proposed method to the CELP vocoder, we can reduce complexity by 64% in the pitch search.

요 약

부호여기된 선형예측(CELP) 음성부호화기는 4.8kbps이하의 낮은 전송 비율에서도 좋은 성능을 갖는다. CELP형 부호기의 단점은 많은 계산량을 필요로 한다는 것이다. 본 논문에서, 우리는 복잡성을 줄이면서 CELP보코더의 음질을 유지하는 새로운 피치검색법을 제안하였다. 이 방법은 CELP보코더의 포만트 필터단에서 찾은 제 1 포만트를 이용하여 예비피치를 찾고, 피치검색을 예비피치 구간에서만 수행하는 것이다. 제안한 방법을 CELP보코더에 적용하므로써, 기존의 방법에 비해 약 64%의 복잡성이 감소되었다.

Ⅰ. INTRODUCTION

Linear predictive speech coders have dominated speech coding applications for the past two decades. A common characteristic of these coders is that open-loop methods are used for the analysis of the spectrum filter and the excitation signal. With these open-loop methods, no performance measure is defined directly between the original speech and the reconstructed speech. The analysis-by-synthesis method, or closed-loop analysis method, has also been successfully applied to several speech coding techniques such as multipulse excited LPC and code excited linear prediction(CELP). With a perceptually weighting distortion measure, the analysis part of these speech coding schemes can be optimized by minimizing the chosen distortion measure between original speech and reconstructed speech.

Fig. 1-1 is a schematic diagram of the CELP vocoder. The excitation signal is formed by filtering the selected random sequence through the selected pitch synthesizer. For the closed-loop excitation analysis, a suboptimal sequential procedure is used. This procedure first assumes zero input to the pitch synthesizer and employs the closed loop pitch synthesizer analysis method to compute the pitch lag and the pitch filter coefficients. Pitch synthesizer fixed, then a closed loop method is used to find the best excitation random sequence, C_i, and compute the corresponding gain, G. To save computation, we use the first order pitch synthesizer.

For speech coding at 8 kbit/s, the sampling rate is 8 kHz and the frame size is 160 samples. The spec-

*Soongsil University
**ETRI-VLSI Lab
접수일자: 1996년 3월 7일

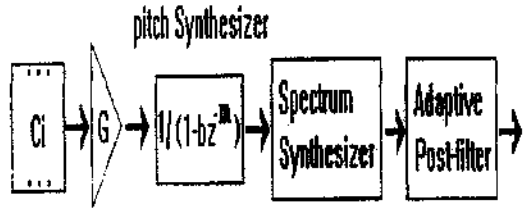

Fig. 1-1 A block diagram of the CELP vocoder.

trum filter uses the 26-bit coding scheme and seven bits are used to specify 128 random sequences for the excitation codebook. Seven bits are allocated for the pitch period, m, with a range from 20 to 147 samples. Three bits each are allocated for the gain, G, and the pitch synthesizer coefficient, b, respectively. The excitation information is updated four times per frame.

The closed-loop analysis of the random excitation and the pitch synthesizer require extremely high computational complexity. For real-time implementation using a fixed point DSP chips, substantial reduction of the computational complexity is essential. By using these complexity reduction techniques, real time implementation of the CELP coder becomes practical.

Ⅱ. PITCH SEARCHING METHOD

The pitch searching procedure is to determine the optimal pitch delay and gain by using a closed loop structure. That is, this procedure computes autocorrelation values with altering time delay gradually and regards time delay that has the maximum value of autocorrelation as pitch period. Conventional methods until now to improve pitch search are self-excited structure[2], extended adaptive codebook structure[3], delta pitch search structure[4], etc. These methods reduce the pitch searching time by considering the correlation between adjacent pitch periods.

In pitch searching, the normalized correlation E(.) of residual signal s(n) according to time delay is computed as follows:

$$E(L)=\frac{\sum_{n=0}^{N-1}(s(n)\,s(n-L))}{\sum_{n=0}^{N-1}(s(n-L)\,s(n-L))} \qquad (2.1)$$

where N is subframe length and L is time delay. Therefore, the correlation is obtained by the value near 100% in each pitch period, and the similarity differs according to the amplitude variation and the periodicity of waveform. When the time delay conforms to the constant times of periodicity of speech waveform, the autocorrelation has a maximum value.

To obtain the most desirable time delay in pitch searching, the correlation equation in Eq.(2-1) must be repeatedly performed about all pitch delays as much as possible. This requires many computation owing to perform multiplication and addition each N time, every time delay L (from 20 to 147). For this reason, the pitch searching time of CELP vocoder needs over 5 MIPS when implementing with integer type DSP chip. This computation complexity is occupied half of overall complexity. And, as far as it has no effect on pitch search error, we need the technique to reduce only pitch searching time.

Ⅲ. Fast Pitch Searching Method by the LSF

The pitch searching is to obtain the pitch gain and pitch lag when synthesized speech signal is similar to original speech[1], that is, the correlation in time delay of pitch lag is maximum. To obtain the time lag which has maximum correlation, it needs to search sequentially the range of real pitch. Because this sequential pitch searching method consumes too much time, we will perform pitch searching only about preliminary pitch which has minimum period component of the voiced signal.

According to the excitation source, speech signal can be classified into voiced, unvoiced and plosive. For the source of unvoiced speech is the random noise generator, it has no periodicity. But, because it has the formant at near 3kHz, the average zero crossing rate(ZCR) of unvoiced speech is higher than that of voiced signal. The voiced speech is produced by forcing air through the glottis with the tension of vocal cords adjusted so that they vibrate in a relaxation oscillation, thereby producing quasi-periodic pulses which excite the vocal tract.

In the voiced signal, the energy of the first formant(F_1) is higher about 10dB than that of the side formants. Therefore voiced signal in time domain is dominated to the effect of F_1. In one pitch interval, the inverse of average zero crossing interval(ZCI) is equal to $2F_1$. So the formants can be damping oscillation during the pitch interval in the time domain.

When we sample voiced signal with 8㎑, the inverse of fundamental frequency, F_0^{-1}, is the value of between 20 and 200 samples and the inverse of first formant frequency, F_1^{-1}, is the value of between 10.6 and 32 samples. So, we perform pitch searching on the representative in 20 samples. But, F_1 is higher than or equals to F_0, therefore we get the average period of F_1, and use it as interval for getting main preliminary pitch.

Many methods that estimate average period of F_1, in one frame have been proposed. We choose the line spectrum pair(LSP) method to estimate F_1, since LSP coefficients obtained by the formant filter section. The LPC coefficients transform to the LSP frequency. $A(z)$ is defined such as:

$$A(z) = 1 - a_1 z^{-1} - \dots - a_{10} z^{-10} \tag{3.1}$$

where $a_i (1 \le i \le 10)$ are LPC coefficients. $P_A(z)$ and $Q_A(z)$ are defined as follows:

$$\begin{aligned} A_A(z) &= A(z) + z^{-11} A(z^{-1}) \\ &= 1 + p_1 z^{-1} + \dots + p_5 z^{-5} + p_5 z^{-6} \\ &\quad + \dots + p_1 z^{-10} + z^{-11} \end{aligned} \tag{3.2}$$

$$\begin{aligned} Q_A(z) &= A(z) - z^{-11} A(z^{-1}) \\ &= 1 + q_1 z^{-1} + \dots + q_5 z^{-5} + q_5 z^{-6} \\ &\quad + \dots + q_1 z^{-10} + z^{-11} \end{aligned} \tag{3.3}$$

where $p_i = -a_i - a_{11-i}$ $\quad 1 \le i \le 5$

$q_i = -a_i - a_{11-i}$ $\quad 1 \le i \le 5$

The LSP frequencies are the ten roots which exist between w=0 and w=0.5 in the following two equations:

$$\begin{aligned} P'(w) &= \cos 5(2\pi w) + p_1' \cos 4(2\pi w) \\ &\quad + \dots + p_4' \cos(2\pi w) + p_5'/2 \end{aligned} \tag{3.4}$$

$$\begin{aligned} Q'(w) &= \cos 5(2\pi w) + q_1' \cos 4(2\pi w) \\ &\quad + \dots + q_4' \cos(2\pi w) + q_5'/2 \end{aligned} \tag{3.5}$$

where the p′ and q′ are computed recursively as follows from the p and q values defined above.

$$\begin{aligned} &p_0' = q_0' = 1 \\ &p_i' = p_i - p_{i-1}' \qquad 1 \le i \le 5 \\ &q_i' = q_i + q_{i-1}' \qquad 1 \le i \le 5 \end{aligned} \tag{3.6}$$

A property of the LSP frequencies is that if the LPC filter is stable, the roots of the two functions alternate: the smallest root, w_1, is the lowest root of $P'(w)$, the next smallest root, w_2, is the lowest root of $Q'(w)$, etc. Thus, w_1, w_3, w_5, w_7 and w_9, are the roots of $P'(w)$, and w_2, w_4, w_6, w_8 and w_{10} are the roots of $Q'(w)$. Line spectrum pair obtains p/2 order real root of equation instead of p order double root which is more easy than formant frequency.

Line spectrum pair frequency(LSF) computes for driving the formant filter at CELP vocoder. LSF displays well the resonance peak of speech formant. Especially the first root of $Q'(w)$ means the resonance frequency of first formant. Thus, in this paper, the average period of the first formant, F_1^{-1}, is obtained by LSP as follows:

$$F_1^{-1} = Fix(f_s / w_2) \tag{3.7}$$

where $Fix(.)$ is an integer function of variable and f_s is the sampling frequency.

In order to getting the preliminary pitch of given speech, we calculate average period of the first formant. If the period was longer than the minimum pitch period, 20 samples, then this value is to be the decimation interval, $D_1 = F_1^{-1}$, for searching the preliminary period. But if the value was shorter than or equal to minimum pitch, decimation interval, D_1, is 20.

First, a frame of D_1 samples is invested with duration number i. At this time, with computing the maximum peak of i-th composed D_1 samples, the magnitude and the position value of it are stored in peak buffer $p(i, 1)$ and $p(i, 0)$, respectively. Likewise, with measuring the minimum valley, the magnitude and the position value of it are stored in valley buffer $v(i, 1)$ and $v(i, 0)$, separately.

In this way, if the peak and the valley are found, the preliminary pitch may have error of a few sample because of the effect of the phase variation of the third formant of speech signal. Therefore, this effect of the higher formant can be removed by performing above decimation procedure after speech signal is filtered by LPF, the cutoff frequency of the filter is 2.67kHz. To use the detected peak and valley as preliminary pitch, when the difference between the first found prominent peak(valley) as standard and the next peak (valley) exist only in interval as following, the autocorrelation Eq.(3-1) must

be performed:

$$T_p(2i) = p(i, 0) - T_{hp} \quad \text{and}$$
$$T_v(2i+1) = v(i, 0) - T_{hv}, \quad i = 1, 2, \ldots, 12 \tag{3.8}$$

where T_{hp} and T_{hv} are the position of the first prominent peak and the first prominent valley, respectively. The detected preliminary pitch collection, $T_p(i)$, is applied to computation of $E(L) = Exy/Eyy$, and then we choose a maximum $E(T_p(i))$ as such as the pitch lag, L, of pitch filter. Then, the coefficient of pitch filter is

$$b_i = E_{xy}/E_{yy} \tag{3.9}$$
$$= \frac{\sum_{n=0}^{L-1} (s(n)\, s(n-L))}{\sum_{n=0}^{L-1} (s(n-L)\, s(n-L))}$$

The peak and valley are searched one per D_1 samples by considering separately the interval of peaks and valleys. Fig. 4-1 is to present this algorithm mentioned above.

Ⅳ. EXPERIMENTAL RESULTS

For the simulation, we used the IBM-PC/486DX2 (66MHz) interfaced with A/D converter for input and output of speech signal. The sampling frequency is 8 kHz and quantization level is 16 bit/samples. On each utterance, the frame length is 160 samples and the subframe length with 40 samples is processed. The speech data composed of 3 Korean speaker's utterances(a female 20 years old, a male 22 years old, and a male 28 years old) and 20 sentences were spoken by 5 times.

The implementation of pitch searching in CELP vocoder is performed with the C language. For performance test of pitch searching method, the procedure of computer simulation is divided into two part. Firstly, the sequential pitch search method is processed by increasing the pitch lag L in full pitch searching range(from 20 to 147).

The second part of processing is implemented by the proposed method, first, the decimation interval, D_1, is determined by separating the components of speech signal and then the preliminary pitches are searched. This method performs the separation with phoneme component which approximately obtains the first formant by detecting the LSF. The period of the first formant obtained like this is applied decimation interval to obtain the preliminary pitches. We can get the preliminary pitches by decimation of residual signal for each decimation interval. The decimation method is carried out by searching peak and valley of the residual signal.

To obtain the difference of pitch search time between two procedures, the average searching time every 1 sec utterance is obtained by pitch computation. The sequential pitch search method is need to average 7.52 sec, but proposed method needs average 2.71 sec, resultingly pitch search time is reduced to average 64%. As the estimated time value is different according to computer types, we have considered only relative time reduction rate in the evaluation. But, the prediction gain of proposed method is degraded by average 0.83dB degradation comparing to the full pitch search in clean speech as shown in table 4-1.

Table. 4-1 Pitch filter gain according to the SNR.

Mehod \ SNR(dB)	Clean	20	6	0
Full search	11.65	11.33	8.70	7.61
Proposed search	10.82	10.48	7.92	6.87
Degradation	0.83	0.85	0.78	0.74

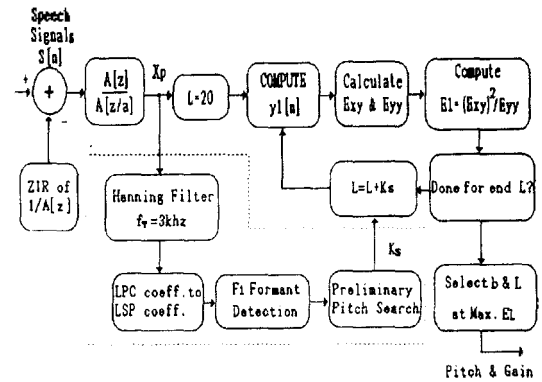

Fig. 4-1 The pitch search algorithm proposed in this paper.

Ⅴ. CONCLUSION

The CELP vocoder provides high quality at low bit rate by using an analysis by synthesis that compares input speech signal with synthesized speech. But it is difficult to implement in real-time with the

finite word length DSP chip, because the computation time is very long. In CELP vocoder, the pitch searching time takes a half of overall processing time. Therefore, we proposed a new algorithm to reduce the pitch searching time by using preliminary pitches with little distortion.

In this paper, we pre-grasp the period of preliminary pitches by the line spectrum frequency and then carried out final pitch search only about these preliminaries. The pitches of speech signal are detected generally above 2.5 ms and are equal to or longer than the period of the first formant. Average period of the first formant is estimated by w_2 frequency of the LSP obtained from the formant filter section. With this proposed algorithm, the result of performing pitch search have been degraded average 0.83 dB than that of the sequential pitch search, but the pitch searching time have been reduced by 64%.

REFERENCES

1. A.N. Ince, *Digital Speech processing* (speech coding, synthesis, and recognition), Kluwer Academic Publishers, 1992.
2. W.B. Kleijn et al., "Fast Methods for the CELP Speech Coding Algorithm," IEEE:Trans., Acoustics, Speech and Signal Processing, Vol.38, No.8, pp.1330-1341, Aug. 1990.
3. A.L. Guyader, D. Massaloux, and J.P. Petit, "Robust and Fast Code Excited Linear Predictive Coding of Speech Signals," Proc. Int. Conf. on Acoustics, Speech and Signal Processing, 1989.
4. I. Gerson and M. Jassuik, "Techniques for improving the Performance of CELP type Speech Coders," Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp.205-208, 1991.
5. U. Balss, U. Kipper, H. Rciriger and D. Wolf, "Improving the Speech Quality of CELP Coders by Optimizing the Long Term Delay Determination," EUROSPEECH-'92, pp.59-62, 1992.
6. M.J. Bae, D.S. Kim, H.Y. Jeon and S.G. Ann, "On a new predictor for the waveform coding of speech signal by using the dual autocorrelation and the sigma-delta technique," IEEE Proc. of ISCAS'94, Vol.6, No.3, pp. 261-264, June 1994.
7. M.J. Bae and J.H. Lee, "On a Waveform Coding Method using the Nonuniform Sampling for Speech Signals Separated to High-Low Band," Proc. of GLOBECOM'95, Vol.3, pp.1599-extra page, Nov. 1995.

▲배 명 진(Myung Jin Bae)
현재 : 숭실대학교 정보통신공학과 교수
(제14권 1E호 참고)

▲손 상 목(Sang Mok Sohn) 1970년 1월 11일생

1996년 2월 : 숭실대학교 공과대학 정보통신공학과 졸업(공학사)
1996년 : 숭실대학교 전기공학과 석사과정 재학중

▲유 하 영(Hah Young Yoo)
현재 : 한국전자통신연구소 반도체개발단 책임연구원
(제14권 1E호 참고)

▲변 경 진(Kyung Jin Byun)
현재 : 한국전자통신연구소 반도체개발단 선임연구원
(제14권 1E호 참고)