On a Reduction of Pitch Searching Time by Preliminary Pitch in the CELP Vocoder

CELP 보코더에서 예비피치에 의한 피치검색 단축

Daesik Kim*, Myungjin Bae*, Jongjae Kim**, Kyungjin Byun**,
Kichun Han**, Hahyoung Yoo**

김 대 식*, 배 명 진*, 김 종 재**, 변 경 진**, 한 기 천**, 유 하 영**

※ 본 논문은 전자통신연구소의 1994년도 수탁과제 연구 지원에 의해 수행되었습니다.

ABSTRACT

Code Excited Linear Prediction(CELP) vocoder exhibits good performance at data rates below 4.8 kbps. The major drawback to CELP type coders is their large amount of computation. In this paper, we propose a new pitch search method that preserves the quality of the CELP vocoder with reduced complexity. The basic idea is to restrict the pitch searching range by estimating the preliminary pitches. Applying the proposed method to the CELP vocoder, we can get approximately 87% complexity reduction in the pitch search.

요 약

부호여기된 선형예측(CELP) 음성부호화기는 4.8 kbps 이하의 낮은 전송 비율에서도 좋은 성능을 갖는다. CELP형 부호기의 단점은 많은 계산량을 필요로 한다는 것이다. 본 논문에서, 우리는 복잡성을 줄이면서 CELP 보코더의 음질을 유지하는 새로운 피치 검색법을 제안하였다. 그 기본 개념은 음성 파형의 예비피치를 찾아내어 피치검색 범위를 제한하는 것이다. CELP 보코더에 제안한 방법을 적용하므로써, 피치검색에서 기존의 방법에 비해 약 87%의 복잡성이 감소되었다.

I. INTRODUCTION

The main methods of coding for transmission or storage of speech signals can be classified into following three types generally ; waveform coding, source coding, and hybrid coding method.

In the waveform coding method, the redundancies in speech waveforms are removed before they are transmitted through the transmission channel or stored in some storage medium. Examples of waveform coding methods are PCM, ADM, ADPCM, etc. In recent, due to the improvement of the manufacturing techniques and the development of the DSP (Digital Signal Processor) algorithms, the standard ADPCM chip has realized with a bit rate of 32 kbps [1]. Also, the waveform coding method can maintain the high quality and personality, because, in the processing procedure, two pieces of information, both the vocal tract filter information that represent the meaning of message and the excitation information that reflect the personality and the feeling of a person, are not specially separated but transmitted.

*숭실대학교 정보통신공학과
Soong Sil University Dept. Telecommunication Engineering
*전자통신연구소, IC 개발부
IC Technology Dept., ETRI
접수일자 : 1994년 8월 30일

The methods of source coding are very closely related to the speech production model. In speech signals these methods separate the excitation information and the filter information, and then each is coded. Examples of source coding methods are LPC, PARCOR, LSP, MBE, formant coding, etc. These algorithms have low transmission rates about 1 kbps.

The hybrid coding methods have the memory efficiency of source coding and the naturalness and articularity of waveform coding. In this method, the formant information is encoded generally by Linear Predictive Coding method (LPC), and according to the method of encoding the residual signal, these methods are classified into RELP, VSELP, MPLP, and CELP. Among these methods, Code Excited Linear Prediction(CELP) is adopted for the mobile communication recently.

The CELP method encodes the pitch period of speech signal with applying to pitch filter. The pitch searching method applied primarily to this pitch filter is the correlation method using pitch lag. The pitch lag and gain of the pitch filter in pitch searching method by correlation is determined by searching the maximum correlation of all pitch lags. But this pitch searching procedure must be performed about all possible pitch intervals, therefore it is difficult to implement it with a DSP chip, and it also needs much handling time.

For that reason, in this paper, we propose a new method to reduce the pitch searching time by preliminary pitch in the CELP vocoder.

Ⅱ. THE PRINCIPLE OF CELP VOCODER

The block diagram of CELP speech encoder is shown as Fig. 2-1. A $10th$ order LPC all pole structure usually is applied to the formant synthesis filter. The LPC coefficients are encoded by converting them to LSP coefficients for efficient quantization, and these are converted to LPC coefficients again when decoding. The LPC coefficients are encoded every 20 ms, and provided differently each subframe of 5 ms with interpolating. Also the excitation source parameter is encoded every subframe of 5 ms.

Both the encoder and the decoder make use of two excited source. The first excited source is a long-term(pitch) predictive state or an adaptive codebook. The second excited source is an excitation codebook. The codeword length is 128 samples in the low transmission rate. These two excited sources are multiplied by a conformational gain term, respectively, and then summed. This is a combined excitation sequence. The excited output of each subframe is applied to update the long-term filter state of the adaptive codebook to be made use of in the next subframe.

In CELP vocoder, because of applying vector

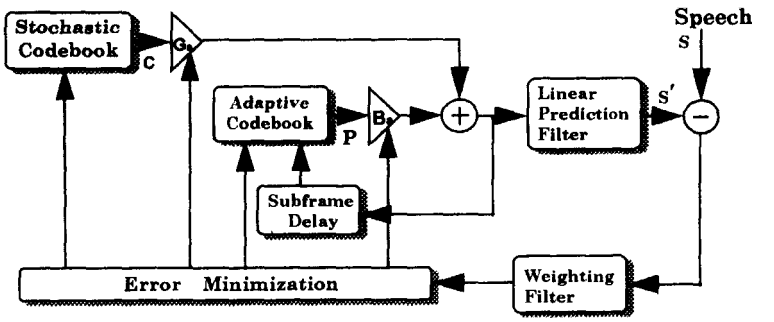

Fig. 2-1 A CELP speech coder

quantization to the residual signal that the formant information is strained, the data to be transmitted for representing the residual signal component is only the index of codebook and a gain. Consequently, the transmission rate can be lowered below 4.8 kbps and if these parameters are transmitted together with additional error correcting code, it may be or can be robust under transmission noise. Since it is analyzed repeatedly to maintain optimal toll quality by an analysis-by-synthesis method, the quality is excellent in the low transmission rate.

The CELP vocoder has complex structure as it must always compare and synthesize the speech signal. Specially, it requires a lot of computation, and wastes the most of computation time to find the input excitation parameters and the coefficients of pitch filter p(z).

The pitch searching is to obtain pitch period information corresponding to long-term correlation of speech signal. The pitch analysis of speech signal sampled in 8 kHz performs each 5 ms. The spectrum analysis can be performed successfully by an open circuit structure, but the pitch analysis must be done by a closed one. That is, the optimally pitch delay must be determined through repetitive comparison. The optimal pitch gain is computed by using the resulted pitch delay and quantized. It is important that this pitch searching procedure has a large effect on computation of CELP vocoder together with codebook search.

Ⅲ. PITCH SEARCHING METHOD

The pitch searching procedure is to determine the optimal pitch delay and gain by using a closed circuit structure. That is, this procedure computes achieves autocorrelation values with altering gradually time delay and regards time delay that has the maximum value of autocorrelation as pitch period.

So far the proposed methods to improve pitch search are self-excited structure [7], extended adaptive codebook structure [5], delta pitch search structure [6], 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 :

$$E(L) = \frac{\sum_{n=0}^{M-1} (s(n)s(n-L))}{\sum_{n=0}^{M-1} (s(n-L)s(n-L))} \qquad (3\text{-}1)$$

where M is subframe length and L is time delay. Therefore, the correlation is obtained the value near 100% in each pitch period, and the similarity differs according to amplitude variation and 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 (3-1) must be repeatedly performed about all pitch delays as much as possible. This requires many computation owing to perform multiplication and addition each M 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 the latest DSP chip, and this computation complexity is occupied the 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.

Ⅳ. REDUCTION OF PITCH SEARCHING TIME

To obtain the time lag which has maximum correlation, it is needed to search the pitch duration sequentially. As the sequential pitch searching method retains too much of the time in processing, firstly, we need to know the duration of high

correlation in pre-processing. By searching this duration for pitch search, the computation time is reduced.

The pitch of speech signal is defined the duration between the peaks or valleys. In the case of pitch detection by using the peaks, the autocorrelation is high at the lags corresponding to prominent peaks. Also, because pitch period is not existed in 2.5 ms, the preliminary pitch that will be applied to pitch search is obtained by decimating waveform as follows :

First, a frame of 19 samples is investigated with duration number i. At this time, with computing the maximum peak of *i-th* that is composed of 19 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 samples because of the effect of the phase variation of the third formant of speech signal, therefore this effect can be removed by performing decimation procedure after speech signal is filtered by Hanning :

$$s'(n-2) = \frac{s(n)+2s(n-1)+3s(n-2)+2s(n-3)+s(n-4)}{9} \tag{4-1}$$

where, the cutoff frequency of this filter is 2.67 kHz. To use the difference or distance between detected peaks and valleys as preliminary pitch, the autocorrelation in (3-1) must be performed when the difference between the first founded prominent peak(valley) as standard and the next peak(valley) exists only within interval as following, the autocorrelation in (3-1) must be performed :

$$T_p(2i) = p(i, 0) - T_{hp}, \quad \text{and}$$
$$T_v(2i+1) = v(i, 0) - T_{hv}, \; i = 1, 2, \ldots, 12 \tag{4-2}$$

where T_{hp} is the position of the first prominent peak and T_{hv} is the position of the first valley.

The detected preliminary pitch collection is applied to (3-1), $T_p(i)$ is acquired from maximum $E(T_p(i))$ determined by the pitch value L, and the gain of pitch filter is

$$b_i = E_{xy}/E_{yy}$$

$$E(L) = \frac{\sum_{n=0}^{L-1} (s(n)s(n-L))}{\sum_{n=0}^{L-1} (s(n-L)s(n-L))}. \tag{4-3}$$

The peak and valley are searched, respectively, every 19 samples by considering the interval between peaks and valleys. And if the preliminary pitch interval is found each, the pitch searching time is reduced much more than that of the full pitch search method as following :

$$T_R = \frac{2}{19} \times 105 = 11\%. \tag{4-4}$$

Where 5% in computation time is added by considering the time that performing decimation to find the preliminary pitch.

V. EXPERIMENTS & RESULTS

For the simulation, we used the IBM-PC/486DX II (50 MHz) interfaced with A/D converter for input and output of speech signals. The sampling frequency is 8 kHz and quantization level is 12 bit/samples. And, 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 the following sentences were spoken 5 times, respectively.

Sentence 1) /IN SOO NE KO MA GA CHUN
JAE SO NYUN WL JO A HAN DA/

Sentence 2) / JE SU NIM KE SEO CHUN JI CHANG JO WI KIO HUN WL MAL SUM HA SEOSS DA /

Sentence 3) / SOONG SIL DAE JUNG BO TONG SIN CONC HAK KWA UM SEONG SIN HO CHU RI YUN GU SIL /

Sentence 4) / GONG IL I SAM SA O RUK CHIL PAL GU /

Where the meaning of sentence 1 is "Insoo's young boy likes a genius kid", sentence 2 is "Jesus spoke of the lessons of the creation of the heavens and the earth", sentence 3 is "Speech signal processing team at the department of information and telecommunication, Soongsil University", and sentence 4 is "one two three four five six seven eight nine", spoken in Korean.

The implementation of pitch searching in CELP vocoder is performed with the C-language. In the block diagram of Fig. 5-1, the pre-processing as block of correlation searching and of preliminary pitch detection using its results is added to which are represented by in dotted line. Where $1/A(z)$ is the transfer function of formant filter, $A(z)/A(z/a)$ is the transfer function of perceptual weighting filter, and the ZIR is zero input filter response in previous state. $y_L(n)$ is the synthesized speech waveform by pitch lag L, E_{xy} is the cross correlation between the input speech and the synthesized speech, and E_{xx} is the autocorrelation of speech waveforms.

For performance test of pitch searching method, the procedure of computer simulation is divided into two part. Firstly, the sequential pitch search method is executed by incrementing the pitch lag L in period of pitch searching (from 20 to 147). Implemented result is shown in Fig. 5-2(e). In this case the distinct correlation is obtained at each pitch period.

The second part of processing is implemented by the proposed method. After speech signal passes through Hanning filter whose cut-off frequency is 2.67 kHz, we detect a peak and vally per 19 samples in a frame. We obtained the time difference between the first prominent standard peak and the next peak, and if the obtained the time interval is within a range from 20 to 147 samples, it is considered as the preliminary pitch. Also, similarly the preliminary pitch is extracted from the detected valleys. The results are shown in Fig. 5-2(b) and 5-2(c). Simultaneously the pitch search is not performed in period which has not preliminary pitch. The optimal pitch is determined by the pitch that has maximum prediction gain among the preliminary pitches. The correlation value set zero in these skip duration. This results are illustrated in Fig. 5-2(d). In this case the position

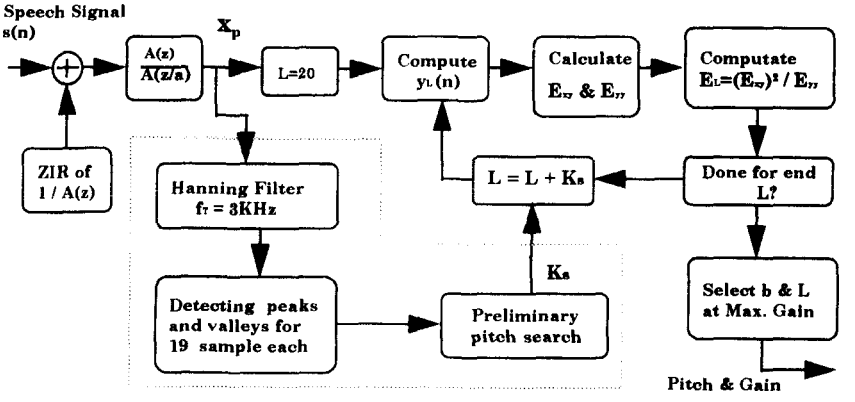

Fig. 5-1 The pitch search algorithm proposed in this paper.

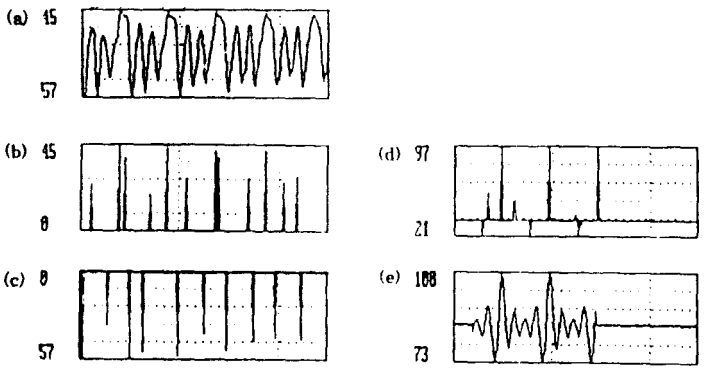

Fig. 5-2 A result for utterance 1.
(a) Speech Signal
(b) Waveform Decimated for Peaks
(c) Waveform Decimated for Valleys
(d) Pitch Filter Coefficients at Preliminary Pitches
(e) Pitch Filter Coefficients by Full Search

of maximum peak of correlation is correctly the same as the optimal pitch.

To obtain the difference of pitch search time between two procedure, the average searching time of 1 sec unit is obtained for above utterances. The sequential pitch search method is need to average 7.52 sec, but proposed method needs average 1.02 sec, resultingly pitch search time is reduced about 87%. As the estimated time value is different according to computer types, we have considered only relative time reduction rate in evaluation. But, the prediction gain of proposed method is degraded by 0.75 dB in average from 10.89 dB to 11.64 dB.

VI. CONCLUSION

The CELP vocoder provides high toll quality by using an analysis-by-synthesis that compares input speech signal with synthesized speech. But it is difficult to implement it in real-time with the existing DSP chip, because the computation time is very large. In CELP vocoder, the pitch searching time hold approximately the half of overall coding time. Accordingly, we proposed a new algorithm to reduce the pitch searching time by using a preliminary pitch.

The pitch period is the interval between peaks or valleys in speech waveform. Additionally, the pitch in the speech has a value generally above 2.5 ms. Thus, by using these properties, we have detected prominent peaks(valleys) in 2.375 ms before performing pitch detection, and then used these interval as preliminary pitch. As results, the pitch searching time is reduced by detecting the coefficient of pitch filter about preliminary pitches. With this proposed algorithm, the result of performing pitch search have been degraded average 0.75 dB than that of the the sequential pitch search, but the pitch searching time have bee reduced about 87%.

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. R. C. Rose and T. P. Barnwell, "Design a Performance of an Analysis-by-Synthsis Class of Predictive Speech Coders," IEEE Trans. Acoustics, Speech and Signal Processing, Vol. 38, No. 9, pp. 1489-1503, Sep. 1990.
4. A. Le 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.
5. J. Menez, C. Galand, M. Rosso, and F. Bottau, "Adaptive Code Excited Linear Predictive Coder (ACELPC)," Proc. Int. Conf. on Acoustics, Speech and Signal Processing, 1989.
6. Joseph P. Campbell, Jr., Vanoy C. Welch, and Thomas E. Tremain, "An Expandable Error Protected 4800 bps CELP Coder(U. S. Fedral Standard 4800 bps Voice Coder)," Proc. Int. Conf. on Acoustics, Speech and Signal Processing, 1989.
7. R. C. Rose and T. P. Barnwell III, "Quality Compression of Low Complexity 4800 bps Self Excited Code Excited Vocoders," Proc. Int. Conf. on Acoustics, Speech and Signal Processing, 1987.
8. Grant Davidson and Allen Gersho, "Complexity Reduction Methods for Vector Excitation Coding," Proc. Int. Conf. on Acoustics, Speech and Signal Processing, 1986.
9. M. R. Schroeder and B. S. Atal, "Code-Exited Linear Prediction (CELP) : High-Quality at Low Bit Rates," Proc. Int. Conf. on Acoustics, Speech and Signal Processing, pp. 25. 1.1-25. 1.4, 1985.
10. L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signal*, Prentice-Hall, 1978.
11. J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*, Springer Verlag, New York, 1976.
12. S. G. BAE, H. R. KIM, D. S. KIM, and M. J. BAE, "On a reduction of pitch searching time by preliminary pitch in the CELP vocoder," WESTPRAC-V, pp. 1104-1111, Vol. 2, Aug. 1994.

▲ **Daesik Kim** : Vol.13, No.1E 참고
▲ **Myungjin Bae** : Vol.13, No.1E 참고
▲ **Jongjae Kim** : Vol.13, No.3 참고
▲ **Kyungjin Byun** : Vol.13, No.3 참고
▲ **Kichun Han** : Vol.13, No.3 참고
▲ **Hahyoung Yoo** : Vol.13, No.3 참고