CELP 보코더에서 전처리에 의한 피치검색 시간의 단축

On A Reduction of Pitch Searching Time by Preprocessing in the CELP Vocoder

김대식*, 배명진*, 김종재**, 변경진**, 한기천**, 유하영**
(Daesik Kim*, Myungjin Bae*, Jongjae Kim**, Kyungjin Byun**,
Kichun Han**, Hahyoung Yoo**)

※본 논문은 전자통신연구소의 1994년도 수탁과제 연구 지원에의해 수행되었습니다.

ABSTRACT

Code Excited Linear Prediction(CELP) speech coders exhibit good performance at data rates below 4.8 kbps. This major drawback of CELP type coders is required much computation. In this paper, we propose a new pitch search method that preserves the quality of the CELP vocoder with reducing complexity. In the pitch searching, we detect the segments of high correlation by a simple preprocessing, and then carry out the pitch searching only for the segments obtained by the preprocessing. By using the proposed method, we can get approximately 77% 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. 이것은 음성 파형의 자기상관계를 간단한 전처리관계식에 의해 사전에 파악하여 필요한 구간에 대해서만 피치검색을 수행하는 방법이다. 제안한 방법은 피치검색에서 기존의 방법에 비해 약 77%의 복잡성이 감소되었다.

Ⅰ. INTRODUCTION

The problem of coding speech signals for transmission or storage has long been a subject of interest in speech research. Generally, speech coding algorithms can be classified into following three types ; waveform coding method, source coding method, and hybrid coding method.

In the waveform coding method, the repetitive redundancy in speech waveforms is removed before it is transmitted through the transmission channel or stored in some storage medium ; PCM, ADM and ADPCM belong to this type. Thanks to

*숭실대학교 정보통신공학과
Soong Sil University Dept. of Telecommunication Engineering
**전자통신연구소, IC 개발부
IC Technology Dept., ETRI

접수일자 : 1993년 2월 28일

the improvement of the manufacturing techniques and algorithms of DSP(Digital Signal Processor), the ADPCM chip has realized with a bit rate of 32 kbps[2]. Also, the waveform coding method can maintain the high quality and personality, because in the processing procedure both the vocal tract filter informations that represent the meaning of message and the excitation informations that reflect the personality and feeling of a person are not separated in two parts.

The source coding method is based on the speech production model. After speech signals are separated the excitation information into the filter information in speech signals, those of two are encoded. The methods that belong to this category are LPC, PARCOR, LSP, MBE, and formant coding. These algorithms are very efficient in channel capacity because they have a low transmission rate below 1 kbps.

The hybrid coding method has the memory efficiency of source coding and the naturalness and intelligibility of waveform coding. In this method, the formant information is coded gencerally by linear predictive coding(LPC) method, and the method coding the residual signal is RELP, VSELP, MPLP, and CELP. Among these methods, recently the coding technique adopted for the mobile communication is code excited linear prediction(CELP) method.

The CELP method codes the pitch period component of speech signal with applying pitch filter. Pitch searching method applied primarily to this pitch filter is correlation method according to pitch lag. Correlation pitch searching method is to decide optimal correlation as the delay and gain of pitch filter in searching correlation of two signal about all pitch delay that pitch exists. But, because this pitch searching procedure must search repeatedly about all pitch interval, it is difficult to implement with existing DSP chip and has many handling time.

For that reason, in this paper, employing the preprocessing technique beforehand grasping autocorrelation property of speech waveform, we propose a new method that can reduce the processing time of pitch search to about 77% in the CELP vocoder.

Ⅱ. THE PRINCIPLE OF CELP VOCODER

Fig. 2-1 is a block diagram of CELP speech coder. Formant synthesis filter usually is applied to the 10th order LPC coefficients of all pole structure. LPC coefficients are encoded after converting to LSP coefficients because the distortion in quantizer is large, and they are converted to LPC coefficients again when decoding. The LPC coefficients are encoded every frame of 20 ms, and provide differently each subframe of 20 ms after interpolation. Also the excited source parameter

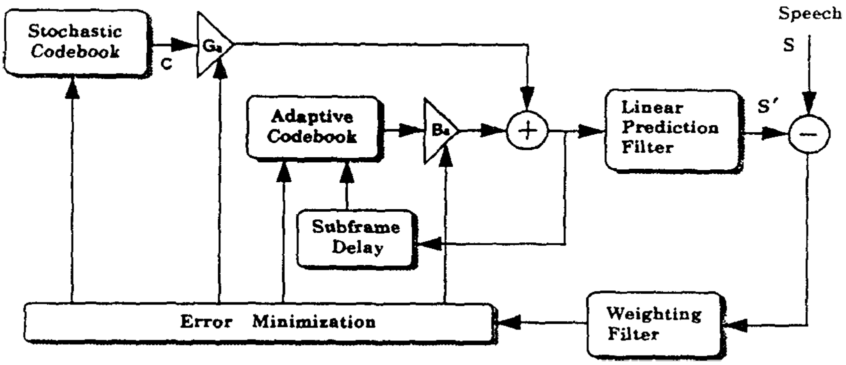

Fig. 2-1. A CELP speech coder

is converted newly every subframe of 5 ms.

CELP encoder and decoder make use of two excited sources. The first excited source is long-term(pitch) predictive state or adaptive codebook. The second one takes in the excited codebook. The size of codebook is 128 at a low transmission rate. These two excited sources are multiplied by the gain term conforming to those, and then the results are summed. These are the combined excited sequence. The excited output of each subframe is applied to change the long-term filter state of the adaptive codebook to be utilized in the next subframe.

In coding method of CELP vocoder, because of coding with applying vector quantization technique to the residual signal that strains formant information, the data applied to transmit the residual signal component is the index of codebook. Consequently, the transmission rate can be lowered below 4.8 kbps and if these parameters are transmitted with additional error correcting code, it is robust coding method under transmission noise. Also, because it is analyzed repeatedly to maintain the talk quality of optimum with applying analysis-by-synthesis method, the quality is excellent at a low transmission rate.

The CELP vocoder has the complicated structure as it must always synthesize speech signal and compare original signal with synthesized one. Specially, it is required a lot of computation when coding, and wasted the most of computation time in the process to find the input excitation signal and the coefficient of pitch filter, $p(z)$. The pitch searching is to obtain the pitch period information corresponding to the long term correlation of speech signal.

Ⅲ. PITCH SEARCHING METHOD

The pitch searching procedure is to decide the value satisfying pitch delay condition optimally between original speech and synthesized one. That is, this detects autocorrelation value with altering gradually time delay about original speech signal, and is the procedure detecting pitch period with time delay that indicates the maximum correlation of the values.

From to new the proposed methods to improve pitch search are self-excited structure[6], expanded adaptive codebook structure[8], and delta pitch search structure[9]. These are methods reducing pitch search time by using correlation between adjacent pitch periods when searching pitch period usually.

Pitch analysis is performed each 5 ms about speech signal sampled in 8 kHz. Spectrum analysis is fulfilled with open circuit structure, but pitch analysis must be done with closed one. That is, the value satisfying pitch lag condition optimally is determined through repetitive comparison. Pitch synthesis filter is given as

$$\frac{1}{P(z)}=\frac{1}{1-bz^{-L}}\ . \tag{3-1}$$

When $x(n)$ is the perceptually weighted input speech and $y(n)$ is the perceptually weighted synthesized speech, the mean squared error(MSE) equation through pitch filter is

$$\begin{aligned} MSE &= \frac{1}{L_p}\sum_{n=0}^{L_p-1}[x(n)-y_L(n)]^2 \\ &= \frac{1}{L_p}\sum_{n=0}^{L_p-1}[x(n)-by(n-L)]^2 \end{aligned} \tag{3-2}$$

where L_p is the length of pitch analysis frame. The objective is to choose the L and b which minimize the MSE, y_L is the synthesized speech waveform of pitch lag L. This is equivalent to maximizing

$$\begin{aligned} E_L &= \frac{(E_{xy})^2}{E_{yy}} \\ &= \frac{\sum_{n=0}^{L_p-1}[x(n)\,y(n-L)]}{\sum_{n=0}^{L_p-1}[y(n-L)\,y(n-L)]} \end{aligned} \tag{3-3}$$

where

$$E_{xy}=\sum_{n=0}^{L_p-1} x(n)\ y_L(n) \text{ and } E_{yy}=\sum_{n=0}^{L_p-1} y_L(n)\ y_L(n).$$

The optimum b for the given L is found to be

$$b_L=\frac{E_{xy}}{E_{yy}}\ . \tag{3-4}$$

This search is repeated for allowed values of L (usually from 17 to 143 or 20 to 147). The lag L and the ptich gain b that maximize E_L are chosen for transmission. The pitch gain is quantized to have relative gain by using the resulting pitch lag before.

In searching pitch, the correlation value $E_L(.)$ of Eq.(3-3) computed along with time delay appears as Fig. 3-1(b). At this time, the correlation is obtained nearly 100% each pitch period, and the similarity differs according to the amplitude variation and periodicity of waveform. Whenever time delay is conformed to the constant times of the periodicity of speech waveform, the autocorrelation has maximum value.

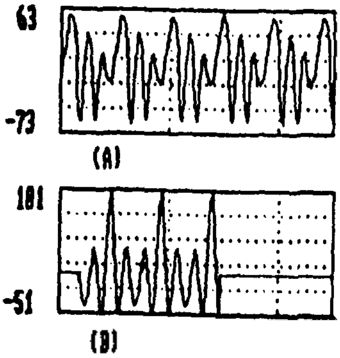

Fig. 3-1. An autocorrelation for time delay
(a) speech waveform
(b) autocorrelation waveform

To obtain most desirable time delay in pitch search method, correlation equation as Eq.(3-3) must be repeatedly performed about all pitch delays as much as possible. This is required much computation to perform multiplication and addition each L_p times every time delay(L is from 20 to 147 samples). In this reason, the pitch searching time of CELP vocoder needs over 5 MIPS when implementing with the latest DSP chip, and this computation time is occupied half of vocoder implementation time. Therefore, the pitch searching technique needs to reduce only searching time as far as it has no effect on pitch search error.

Ⅳ. REDUCTION OF PITCH SEARCHING TIME

The pitch searching in CELP vocoder is to obtain pitch gain and pitch lag that speech signal synthesized with residual signal appears most likely to original speech[1-3], and at this time finds the case that the correlation according to time delay have the highest value. Then the correlation with time delay is searched for maximum. To obtain the time lag which has maximum correlation it needs to search the duration of being pitch sequentially. As the full pitch searching method requires too much of time in processing, first we need to know the duration of high correlation found by preprocessing. By restricting the range of pitch search, computation time can be reduced.

The pitch in speech signals is defined as interval between the repetitive peaks or valleys. In case of pitch detection by using the peaks, the autocorrelation appears high only about time delay that salient peaks exist. Likewise, by using the valleys, we can obtain high autocorrelation only about time delay that prominent valleys exist. If peaks and valleys in the waveform are previously detected, the correlation can be computed as following equation ;

$$R(L)=\sum_{n=-1}^{1} s(n-L)[s(n)+s(n-2L)]$$

$$+\sum_{k=-1}^{1} s(k-L)[s(k)+s(k-2L)] \tag{4-1}$$

$$L=20,\ 21,\ \ldots,\ 147$$

where $s(n)$ indicates the peak, $s(k)$ does the valley in residual signals, n=0 does the vertex of peak, and k=0 does the vertex of valley. In order that correlation value is not affected by impulse noise we take into consideration the correlation value from n+1 to n−1 in standard n=0 point. Method to find peak that comes under pitch period with standard to distinctive peak is to make use of the property that correlation value of Eq.(4-1) forms maximum correlation peak every vertex of peak.

If the correlation of Eq.(4-1) is computed for residual signal, the correlation value is formed positive peak whenever peaks exist. Therefore the duration of positive correlation peaks consider as preliminary pitches, and makes the combination $\{L_1, L_2, \ldots, L_{N-1}\}$ of these. The detected preliminary pitch combination is applied to correlation Eq.(3-1), the pitch value of pitch filter L is determined by maximum E(L), and the coefficient of pitch filter is

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

$$\frac{\sum_{n=0}^{M-1}[s(n)s(n-L)]}{\sum_{n=0}^{M-1}[s(n-L)s(n-L)]}\,. \qquad (4\text{-}2)$$

Above procedure preliminary pitch detection required six multiplication, seven addition, and one comparision per a time delay, but because of existing only a few preliminary pitches, pitch search time can be fairly reduced. The number of preliminary pitches is usually related to the first formant frequency in a pitch period. Because the frequency of the first formant is between 250 Hz and 750 Hz, the maximum number of peaks in pitch search interval is 750/(8000/147)=13.78. In the full pitch searching method Eq.(3-1) is processed 18 times, but we can reduce the computation less than 14 times by using the proposed method added simple preprocessing operation. If the number of preliminary pitches is founded more than 14, then present frame can be considered with unvoiced, mixed, or background noise. As pitch search has the meaning only about voiced speech, the number of preliminary pitches can be limited to 14.

V. EXPERIMENT AND RESULTS

To simulate the proposed algorithm, we use the IBM-PC/486DX2(50) interfaced with A/D converter for input and output of speech signals. The sampling frequency is 8 kHz and quantization level is 12 bits/sample. For the estimation of performance of experimental results, the speech data is 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 are spoken each 5 times.

Utterance 1) /INSUNE KOMANUN CHUNJAE
SONYUNWL JOAHANDA/
Utterance 2) /JESUNIMKESEO CHUNJICHANGJOWI
KIOHUNWL MALSUMHASEOSSDA/
Utterance 3) /SOONGSILDAE JUNGBOTONGSIN
KONGHAKKWA UMSEOUNG
SINHOCHURI YUNGU/
Utterance 4) /KONGILISAMSAORUKCHILPALGU/

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

The pitch searching is performed with the C-language. In this algorithm we proposed, the block of preprocessing previously grasping correlation and the block detecting preliminary pitches with the obtained results beforehand are indicated in dotted line. Where 1/A(z) is the transfer function of formant filter, A(z)/A(z/a) is perceptual weighting filter response, and ZIR is zero input filter response of previous condition. $y_L(n)$ is the synthesized speech waveform of pitch lag L, E_{xy} is cross-correlation between the input speech and the synthesized speech, and E_{yy} is the autocorrelation of speech waveforms.

To compare the performance of pitch searching method the procedure of computer simulation is devided into two parts. First, full pitch search method is to search pitch with increasing one by one the pitch lag L about the interval of pitch searching(from 20 to 147 samples). The performance result shows in Fig.5-2(c). In this case the distinct corelation is obtained each pitch period. Second, as the method we proposed, maximum peak and minimum valley are detected in duration of 50 samples of a frame. The correlation peaks are extracted by increasing of time delay in pitch searching interval by being based on these peaks and valleys. The waveform of detected cor-

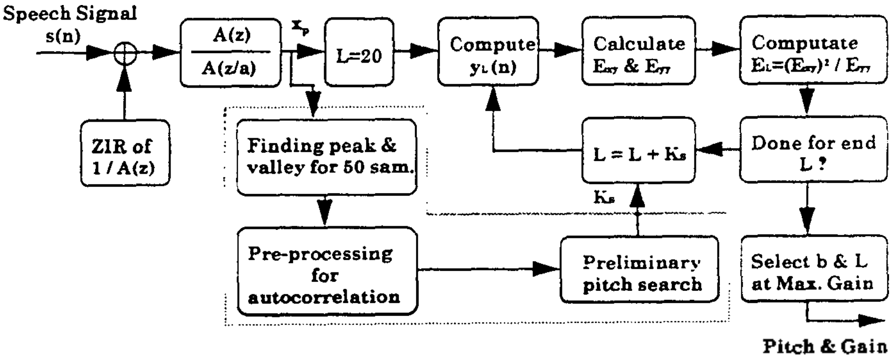

Fig. 5-1. The pitch search algorithm proposed in this paper.

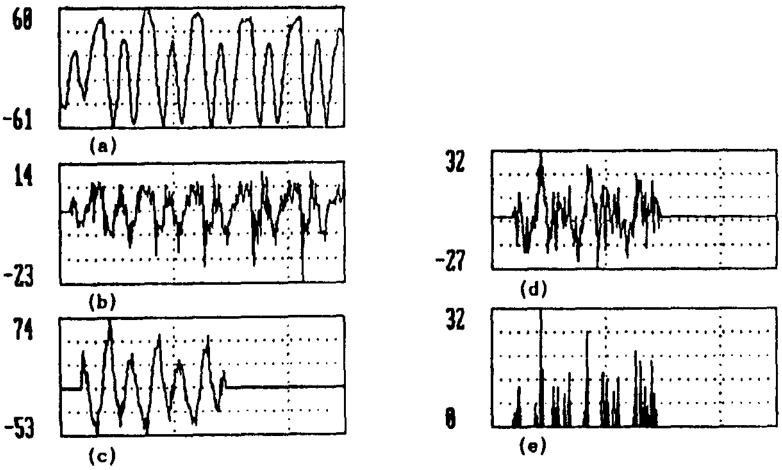

Fig. 5-2. A result for utterance 1.
(a) speech signal
(b) formant residual signal
(c) the result of full pitch search
(d) the result of pre-processing for correlation
(e) the pitch search interval detected by preliminary pitch

relation shows in Fig.5-2(d). If the detected correlation value makes peak, at this time the points are determined as preliminary pitch. And then pitch search is not performed about the interval (K_s) that has not preliminary pitch, but finds pitches that the prediction gain of pitch filter among preliminary pitches has maximum value. The correlation value seus zero in chese ship duration. This result sec Fig.5-2(e). In this case the maximum of position among peake of correlation concords correctly all together.

To obtain the difference of pitch search time in two procedure, the average searching lime of 1 sec unit abtained for above utterances. Full pitch search method needs 25.2 sec and the proposed method needs average 5.8 sec, resultingly pitch search time reduced about 77%. As the measured time value is different according to computer types we considered only relative time reduction rate in evaluation.

Ⅵ. CONCLUSION

The CELP vocoder obtains high talk quality by using analysis-by-synthesis that compares input speech signal with synthesized one. But it is difficult to implement in real time with the existing DSP chip because the required computation is very much. In the CELP vocoder pitch searching time holds about half of coding time. Accordingly, we proposed a new pitch search method that can improve pitch searching time of CELP vocoder.

The full pitch searching method requires too much time in processing, because of searching the duration of being pitch sequentially to obtain the time delay that has maximum correlation. We detect the duration of high correlation by a simple preprocessing, and then apply the time delay obtained by the preprocessing to seaching the pitch finally. Preprocessing method in this paper, is to make use of the property that the correlation value of speech signals forms maximum correlation peak(valley) every vertex of peak(valley) according to time delay. It was required six multiplication, seven addition, and one comparision per a time delay. Because there is existing only a few preliminary pitches to computate correlation equation, pitch search time can be fairly reduced ; i.e., the full pitch searching method is processed 128 times about autocorrelation computation, but reduced less than 14 times by using the proposed method with the simple preprocessing.

The result of performing pitch search with this proposed preprocessing method is similar to that of full search, but pitch search time reduced about 77%.

REFERENCES

1. J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*, Springer Verlag, New York, 1976.
2. L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signal*, Prentice-Hall, 1978.
3. A. N. Ince, *Digital Speech Processing(speech coding, synthesis, and recogintion)*, Kluwer Academic Publishers, 1992.
4. 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.
5. Grant Davidson and Allen Gersho, "Complexity Reduction Methods for Vector Excitation Coding", Proc. Int. Conf. on Acoustics, Speech and Signal Processing, 1986.
6. R. C. Rose and T. P. Barnwell III, "Quality Compression of Low Complexity 4800 bps Self Excited and Code Excited Vocoders," Proc. Int. Conf. on Acoustics, Speech and Signal Processing, 1987.
7. 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.
8. 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.
9. Joseph P. Campbell, Jr., Vanoy C. Welch, and Thomas E. Tremain, "An Expandable Error-Prote-

cted 4800 bps CELP Coder(U. S. Federal Standard 4800 bps Voice Coder)," Proc. Int. Conf. on Acoustics, Speech and Signal Processing, 1989.

10. 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.
11. R. C. Rose and T. P. Barnwell, "Design a Performance of an Analysis-by-Synthesis Class of Predictive Speech Coders," IEEE Trans. Acoustics, Speech and Signal Processing, Vol.38, No.9, pp. 1489-1503, Sep. 1990.
12. 이해군, 금홍, 배명진, "전처리에의한 CELP 보코더의 피치검색 시간 단축에 관한 연구," 제10회 음성통신 및 신호처리 워크샵 논문집, pp.195-199, 1993.

▲Daesik Kim : Vol.13, No.1E 참고

▲Myungjin Bae : Vol.13, No.1E 참고

▲Jongjae Kim : Researcher, IC Technology Dept., ETRI

▲Kyungjin Byun : Researcher, IC Technology Dept., ETRI

▲Kichun Han : Researcher, IC Technology Dept., ETRI

▲Hahyoung Yoo : Researcher, IC Technology Dept., ETRI