On a Reduction of Pitch Searching Time by Separating the Speech Components in the CELP Vocoder

성분분리에 의한 CELP 보코더의 피치 검색시간 단축에 관한 연구

J.I. Hyun*, K.J. Byun*, K.C. Han*, J.J. Kim*, H.Y. Yoo*, J.S. Kim*, D.S. Kim**, M.J. Bae**

현진일*, 변경진*, 한기천*, 김종재*, 유하영*, 김재석*, 김대식**, 배명진**

* 본 논문은 한국전자통신연구소의 1994년도 수탁과제 연구지원에 의해 수행되었습니다.

ABSTRACT

Code Excited Linear Prediction(CELP) vocoder exhibits good performance at data rates below 4.8 kbps. The major drawback of CELP type coders is their 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 pregrasps preliminary pitches about signal and performs pitch search only about the preliminary pitches. Applying the proposed method to the CELP vocoder, we can reduce complexity about 90 % in the pitch search.

요 약

무호여기된 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 피치를 검색하고자하는 신호에 대해 음소 성분 분리를 통해 예비피치 주기를 사전에 파악하고 이들 예비피치에 대해서만 본격적인 피치 검색을 수행하는 것이다. 제안한 방법을 CELP 보코더에 적용하므로써, 피치검색에서 기존의 방법에 비해 약 90%의 복잡성이 감소되었다.

I. INTRODUCTION

Linear predictive speech coders(LPC) 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

*Dept. of VLSI Architecture - ETRI
**Soongsil University, Dept. of Telecommunication Engineering
접수일자 : 1994년 8월 30일

analysis method, has long been used in areas other than speech coding. This method has also been successfully applied to several speech coding techniques, such as multipulse-excited LPC and code-excited LPC (CELP). With a perceptually meaningful distortion measure, the analysis part of a speech coding scheme can be optimized to minimize the chosen distortion measure between the original speech and the reconstructed speech.

For mathematical tractability, a perceptually-weighted mean-squared-error (WMSE) is used as the distortion measure. Coders that employ WMSE are basically waveform coders. In this sense, the resultant coder should be less speaker dependent and more robust against background acoustic noise as the synthesized speech is reconstructed to mimic the individual or the corrupted speech waveform (in a perceptually weighted sense). However, in terms of speech quality, a better distortion measure for narrowband speech coding applications could be used.

The decoder (or the synthesis part) of an analysis-by-synthesis LPC (A-by-S LPC) is identical to that of LPC. To enhance the perceived speech quality, an adaptive post-filter is used. The spectrum filter is typically a tenth-order all-pole filter. The excitation signal can assume a wide range of different models depending on the available data rate. The transfer function of the adaptive post-filter is given as following ;

$$Q(z)=\frac{(1-\mu z^{-1})A(z/a)}{A(z/b)} \qquad (1\text{-}1)$$

where $A(z)=1-\sum_{i=1}^{10} a_i z^{-i}$ is the transfer function of the spectrum filter ; $0<a<b<1$ are design parameters ; and $=ck_1$, where $0<c<1$ is a constant, and k_1 is the first reflection coefficient.

The perceptual weighting filter, W(z), used in the WMSE distortion measure is defined as

$$W(z)=\frac{A(z)}{A(z/\gamma)} \qquad (1\text{-}2)$$

where $0<\gamma<1$ is a constant controlling the amount of spectral weighting.

In CELP vocoder, 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 decided optimal correlation value by searching the correlation of all pitch lags being pitches with two signals. But this pitch searching procedure must search about all pitch intervals, therefore it is difficult to implement with DSP chip and has many handling time.

In this paper, we propose a new pitch searching method that preserves the quality of the CELP vocoder reducing complexity. This method is to pre-grasp the period of preliminary pitch about signal which will search pitch, and perform pitch search only about these preliminary. Applying the proposed method to the CELP vocoder, we can reduce complexity about 90% in the pitch search.

Ⅱ. THE PRINCIPLE OF CELP VOCODER

Fig. 2-1 is a schematic diagram of the CELP speech coder. The excitation signal is formed by filtering the selected random sequence through the selected pitch synthesizer. For the closed-loop excitation analysis, a suboptimum 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 and the pitch synthesizer coefficients. Pitch synthesizer fixed, a closed-loop method is then used to find the best excitation random sequence, C_i, and compute the corresponding gain, G. To save in computation, a first-

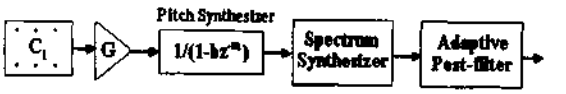

Fig 2-1. A schematic diagram of the CELP speech coder

order pitch synthesizer is used.

For speech coding at 4.8 kbit/s, the sampling rate is 8 kHz and the frame size is 160 samples. The spectrum filter uses the 26-bit coding scheme. 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.

Due to the block processing nature in the computation of the spectrum filter, the spectrum filter parameters can have abrupt change in neighboring frames. To smooth out the abrupt change and to synchronize with the excitation, each excitation subframe uses a different set of LSP(line spectrum pair) parameters. These LSP parameters are obtained through interpolation using the quantized LSP parameters of the current frame and the previous frame.

The procedures for the closed-loop pitch synthesizer analysis and the selection of the best random excitation sequence are identical. For every excitation subframe, each codeword is used as the input signal to the synthesizing filter. Codeword C_i, together with its corresponding gain G, which minimizes the WMSE between the original speech and the synthesized speech, is selected as the best excitation. The minimization step can be formulated as

$$E_{ii}(G, C_i) = \sum_{n=1}^{N} [S_w(n) - GY_{ii}(n)]^2 \qquad (2\text{-}1)$$

where N is the total number of samples in a subframe : $S_w(n)$ denotes the weighted residual signal after the memory of the synthesizing filter has been subtracted from the speech signal : and $Y_w(n)$ denotes the combined response of the synthesizing filter and W(z) to the input signal C_i. The optimum value of the gain term, G, can be derived as

$$G = \sum_{n=1}^{N} S_w(n)\, Y_{ii}(n) / \sum_{n=1}^{N} Y_{ii}^2. \qquad (2\text{-}2)$$

The excitation codeword (C_i) which maximizes the following term is selected as the best excitation codeword :

$$E_{ii}(C_i) = [\sum_{n=1}^{N} S_w(n) Y_{ii}(n)]^2 / \sum_{n=1}^{N} S_{ii} Y_{ii}^2. \qquad (2\text{-}3)$$

For the closed-loop analysis of the random excitation, the synthesizing filter is the combination of the spectrum synthesizer and the pitch synthesizer. For the closed-loop analysis of the pitch synthesizer, the codeword Ci corresponds to the different pitch synthesizer memory due to different pitch periods. The gain term G corresponds to the pitch synthesizer coefficient b.

The closed-loop analysis of the random excitation and the pitch synthesizer require extremely high computational complexity. For real-time implementation using current DSP chips, substantial reduction of the computational complexity is essential. For the pitch synthesizer, it is obvious that the different pitch synthesizer memory as the excitation signal form a set of overlapped excitation sequence. The computation of $Y_w(n)$ in Eq. (2-2) and (2-3) can thus be greatly simplified. For the random excitation, the same technique can be employed by using a codebook with overlapped random sequences. By using these complexity reduction techniques, real-time implementation of the CELP coder becomes practical.

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 :

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

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 Eq. (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 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 THE PITCH SEARCHING TIME BY AVERAGE PERIOD OF THE FIRST FORMANT

The pitch searching is to obtain the pitch gain and pitch lag when synthesized speech signal is similar to original speech[4-6], that is, the correlation with time delay is maximum. To obtain the time lag which has maximum correlation, it needs to search the duration with sequential pitch. Because this sequential pitch searching method consume too much time, we will perform pitch searching only about preliminary pitch which has minimum period component of the voiced signals.

According to the speech 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 3㎑, the average ZCR of unvoiced speech is higher than that of voiced signal. The voiced sounds are 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 of air 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 in time domain voiced signal is dominated to the effect of F_1. In one pitch interval, the inverse of average 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 0 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 search average period of F_1 in one frame have been proposed. In the case of voiced signal average ZCR approximately can be obtained as follow :

$$ZCR=L/2\sum_{n=0}^{L-1}|\,sgn\,[s(n)-sgn\,[s(n-1)]| \quad (4\text{-}1)$$

where N is the size of frame. ZCR represents the main period of frequency which control the

waveform in time domain. In the case of voiced signal, the energy of the first formant is remarkably larger than that of the other formant, so we obtain the ZCR which is proportional to F_1. The average period of the first formant, F_1, is obtained by average ZCR as follow :

$$F_1^{-1} = 2N/ZCR. \tag{4-2}$$

For 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, $DI = F_1^{-1}$, for searching the preliminary period. But if the value was shorter than or equal to minimum pitch, decimation interval (DI) is 20.

First, a frame of DI samples is invested with duration number i. At this time, with computing the maximum peak of i-th composed DI 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 Hanning :

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

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

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

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 $E(L) = E_{xy}/E_{yy}$ and $T_p(i)$, maximum E ($T_p(i)$), determined the pitch value of pitch filter, L. Then the coefficient of pitch filter is

$$b_L = E_{xy}/E_{yy} = \frac{\sum_{n=0}^{L-1}(s(n)s(n-L))}{\sum_{n=0}^{L-1}(s(n-L)s(n-L))}. \tag{4-5}$$

The peak and valley are searched one per DI samples by considering separately the interval of 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 :

$$T_R = \frac{2}{DI} \times 105 \le 11\%. \tag{4-6}$$

Where adding 5% in computation time is considered the time that performing decimation to find the preliminary pitch. Figure 4-1 is to roughly present this algorithm mentioned above.

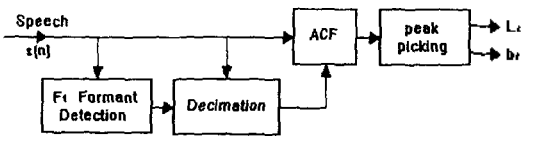

Fig 4-1. The pitch search algorithm proposed in this paper.

V. EXPERIMENTAL 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 16 bit/samples. On each utterance, the frame length is 160 samples and the sub-frame length with 40 samples is processed each frame overlapped. 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 GONG 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. 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 pitch searching range(from 20 to 147).

The second part of processing is implemented by the proposed method. First, the decimation interval DI is determined by separating the components of speech signal and the preliminary pitches are searched. This method performs the separation with phoneme component which approximately obtains the first formant by detecting the ZCR. The period of the first formant obtained like this is applied decimation interval to obtain the preliminary pitch. We can get the preliminary pitch by decimation interval. This method is performed by searching peak-valley.

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 0.68 sec, resultingly pitch search time is reduced about 90%. As the estimated time value is different according to computer types, we have considered only relative time reduction rate in evaluation as shown table 5-1 and 5-2. But, the prediction gain of proposed method is degraded by 0.82 dB in average, in clean speech signal 0.87 dB degradation and in noisy environment 0.76 dB.

Table 5-1. Performance of SNR

SNR(dB) / Method	Clean	20	6	0
Full search	11.64	11.31	8.71	7.63
Proposed search	10.77	10.45	7.92	6.87
Degradation	0.87	0.86	0.79	0.76

Table 5-2. Performance of pitch search

Time / Method	Average pitch search time
Full search	7.52 sec
Proposed search	0.68 sec

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 half of overall coding time. Accordingly, we proposed a new algorithm to reduce the pitch searching time by using a preliminary pitch.

In this paper, first, we pre-grasp the period of preliminary pitch about signal that will search pitch and perform pitch search only about these preliminary. The pitches of speech signals are detected generally above 2.5 ms and are equal or longer than the period of the first formant. Detection of the first formant by separating components is performed with ZCR in time domain. With this proposed algorithm, the result of performing pitch search have been degraded average 0.82 dB than that of the the sequential pitch search, but the pitch searching time have been reduced about 90%.

REFERENCES

1. L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signal*, Prentice-Hall, 1978.
2. J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*, Springer Verlag, New York, 1976.
3. A. N. Ince, *Digital Speech Processing* (speech coding, synthesis, and recognition), Kluwer Academic Publishers, 1992.
4. 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.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. Grant Davidson and Allen Gersho, "Complexity Reduction Methods for Vector Excitation Coding," Proc. Int. Conf. on Acoustics, Speech and Signal Processing, 1986.
11. 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.
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.
13. A. N. Ince, Digital Speech Processing(speech coding, synthesis, and recognition), Kluwer Academic Publishers, 1992.
14. 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.
15. R. C. Rose and T. P. Barnwell, "Design an 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.
16. 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.
17. U. Balss, U. Kipper, H. Reiniger and D. Wolf, "Improving the Speech Quality of CELP coders by Optimizing the Long-term Delay Determination," EUROSPEECH, pp. 59-62, 1992

▲J.I. Hyun : Vol. 13, No. 1E 참조
▲K.J. Byun : Vol. 13, No. 1E 참조
▲K.C. Han : Vol. 13, No. 1E 참조
▲J.J. Kim : Vol. 13, No. 1E 참조
▲H.Y. Yoo : Vol. 13, No. 1E 참조
▲J.S. Kim : Vol. 13, No. 1E 참조
▲D.S. Kim : Vol. 13, No. 1E 참조
▲M.J. Bae : Vol. 13, No. 1E 참조