A Fast Pitch Searching Algorithm Using Correlation Characteristics in CELP Vocoder

상관관계 특성을 이용한 CELP 보코더의 고속 피치검색 알고리듬

JooHun Lee*, MyungJin Bae**, SouGuil Ann*
이 주 헌*, 배 명 진**, 안 수 길*

ABSTRACT

The major drawback to the Code Excited Linear Prediction(CELP) type vocoders is their large computational requirements. In this paper, a simple method is proposed to reduce the pitch searching time in the pitch filter almost without degradation of quality. Bease upon the observational regularity of the correlation function of speech, the searching range can be restricted to the positive side in pitch search. This is done by skipping the negative side with the width which is estimated from the previous positive envelope. In addition to that, the maximum number of available lags can be limited by the threshold, L_T, which is set on 58 empirically. So, only the limited numbers of lags are considered in pitch search, which is less than a half of that of the full search method. By using the proposed method in pitch search, its required computations are greatly reduced. Experimental result shows 51% time reduction almost without lowering the speech quality in segmental SNR measure.

요 약

CELP 타입의 보코더에서 가장 큰 단점은 계산량이 상당히 커서 실시간 구현에 어려움이 많다는데 있다. 이러한 계산량의 부담을 줄이기 위해서 본 논문에서는 음질의 저하없이 피치검색시간을 단축하는 간단한 방법을 제안한다. 음성신호의 상관함수에서 발견되는 몇 가지의 특성으로 부터 피치검색은 상관함수의 양의 구간만으로 한정될 수 있다. 이러한 피치검색구간의 한정은 상관함수에서 음의 진폭구간을 앞선 양의 진폭 구간의 폭만큼으로 추정하여 건너뜀으로서 구현할 수 있다. 또한 검색되는 피치래그의 갯수를 일정한 수로 제한할 수도 있는데 실험적으로 약 58로 제한된다. 따라서 제안된 수의 피치래그에서만 피치검색이 수행된다. 제안된 방법으로 피치검색을 수행한 결과 기존의 방법에 비하여 음질의 저하없이 약 51%의 시간단축이 되었다.

I. INTRODUCTION

After the intorduction of the Code Excited Linear Prediction(CELP) speech coder in 1984 [1], there have been many researches to achieve high quality speech below 4.8 kbps within reduced computational requirements. The major drawback in CELP type analysis-by-synthesis speech coders is their large computational requirements in codebook and pitch searches [2]. CELP analysis con-

*Department of Electronics Engineering, Seoul National University.
**Department of Telecommunication Engineering, Soongsil University.
접수일자 : 1994년 6월 28일

sists of three basic functions : 1) short delay spectrum prediction, 2) long delay pitch search, and 3) residual codebook search. The spectrum analysis is performed once per frame by open-loop, usually 10th order autocorrelation LPC analysis using no preemphasis and 15 Hz bandwidth expansion with a Hamming window [3]. The codebook search is performed by closed-loop analysis using conventional minimum mean squared prediction error criterion of the perceptually weighted error signal. The pitch search is done usually using one of the followings : filtering [4], self-excited [5], or adaptive codebook [6] methods. Since the pitch search is performed four times per frame based upon analysis by-synthesis technique and all of the available pitch lags are exhaustively searched, it requires great computational complexity. These computational requirements of the pitch search are almost same as those of the codebook search and time reduction of the pitch search can reduce the overall computational requirements in CELP considerably.

In this paper, a simple method is proposed to reduce the pitch searching time in the correlation based pitch predictor with audible distortion of speech quality. On the basis of the observational regularity of the correlation function in pitch search, the searching range can be restricted to the positive side by estimating the width of negative envelope with the width of previous positive envelope. Experimental result shows that 34% reduction can be achieved by doing so. In addition to that, by limiting the maximum number of available lags in pitch searching by 58, more reduction can be achieved up to 51% almost without lowering the speech quality in segmental SNR measure.

II. PITCH SEARCHING IN PITCH FILTER

Fig. 1 shows a typical flow for pitch search using one-tap pitch filter. Pitch search is performed based on analysis-by-synthesis technique to select parameters such as the pitch lag L and pitch gain b for pitch prediction filter which minimize the weighted error between the input speech and the synthesized speech. In Fig. 1, ZIR is zero input response and α is perceptual weighting constant and $A(z)/A(z/\alpha)$ is a perceptual weighting filter. Pitch synthesis filter is given as

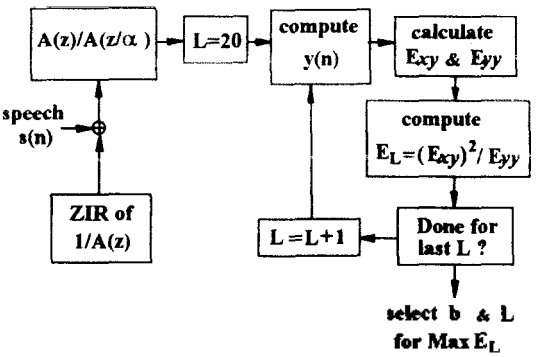

Fig 1. An example of implementation flow for pitch search.

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

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

$$\mathrm{MSE} = \frac{1}{L_p} \sum_{n=0}^{L_p-1} (x(n) - by_L(n))^2 \tag{2}$$

$$= \frac{1}{L_p} \sum_{n=0}^{L_p-1} (x(n) - by(n-L))^2 \tag{2}$$

where L_p is the length of pitch analysis frame. The objective is to choose the L and b which minimze MSE. This is equivalent to maximizing

$$E_L = \frac{(E_{xy})^2}{E_{yy}} \tag{3}$$

where

$$E_{xy} = \sum_{n=0}^{L_p-1} x(n)\, y_L(n)$$

$$E_{yy} = \sum_{n=0}^{L_p-1} y_L(n)\, y_L(n).$$

The optimum b for the given L is found to be

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

This search is repeated for all allowed values of L (usually from 20 to 147). The lag L and the pitch gain b that maximize E_L are chosen for transmission. Since pitch search is done four times per frame by this exhaustive search (every 5 or 7.5 msec), it requires very large computations. To reduce the burden, several methods such as recursive convolution [3][7], approximations of correlation function [4][8] and delta search [3] are used. Delta search method exploits the natural smoothness of pitch lag. For odd subframes, all of available lags searched while for even subframes, only 32 lags relative to the previous subframe are searched. The delta search greatly reduces the computational complexity and data rate while causing no perceivable loss in speech quality. However, since the formulation used for the pitch filter aimes at that it removes long-term correlations whether due to actual pitch excitation or not (strictly it is not a true pitch estimator), the chosen pitch lag L can be an improper lag even for voiced speech and shows doubling and halving (i.e., submultiples of pitch lag) frequently. To overcome this practically, the MSE criterion is usually modified to check the error at submultiples of the lag so to determine if it is within an allowable level of MSE [3]. From the fact that mentioned above, the optimum b can be restricted to be positive [7] so the E_L which produced a negative value of b, E_{xy}/E_{yy}, is ignored in the search.

Ⅲ. PROPOSED PITCH SEARCH METHOD

In connection with the pitch estimation method based on the correlation, the true pitch lag for voiced speech is always located at the peak of a positive envelope in the correlation function [9]. Based upon this fact, pitch lag search in long delay prediction filter can be done in the correlation function and the search range can be restricted to the positive side of correlation function, if possible [7]. The correlation function shows some regularity and has the following properties. The envelope of correlation function varies slowly, for speech signal is highly correlated. The positive and negative envelopes are alternative and the width of each envelope is usually maintained by the effect of the first formant of voiced speech [9].

Based upon the properties of correlation function as mentioned above, the width of a negative envelope can be estimated by the width of the previous positive envelope. By skipping the lags corresponding to that width, pitch search range reduction can be achieved. Since the positive peaks of correlation function are maintained, the performance in segmental SNR does not change. To cope with some bad case, when the width of negative envelope is prolonged so that the skipping by just the same width as that of the previous positive envelope can not exclude the negative correlation lage efficiently, we introduce the adjusting constand $d(\geq 1)$, which is multiplied to the width of the previous positive envelope.

Also, from the fact that, for viced speech, both the numbers of positive correlation lags and negative ones are approximately same and that the maximum value of L is 147, the maximum number of available lags at which the correlation E_{xy} is calulated and checked can be limited. By counting the lags taken part in correlation computations with L_C and restricting this maximum number by the threshod, L_T, pitch searching time can be more considerably reduced with negligible degradation of speech quality in segmental SNR. So, during the pitch search for one frame, if L_C get to L_T, the rest of lags are not considered to be a proper L and the L search of that frame comes to end. In case of that the number of lags which have positive correlation is much more than 84 ($=20+148/2$)), it is considered that the frame lies in the unvoiced or silence segment. Actually, when d is set on 1, 2, the maximum number of

available lags can reduced up to 58 (= 128/(1 + 1.20)) almost withous SNR degradation. As a result mentioned so far, our proposed method can reduce the overall pitch search time requirements remarkably though it introduced extra operations for L_r and d so to incresase the time burden little. Fig. 2 shows the flow of the proposed pitch search algorithm. L is the lag index which varies from 20 to 147 and L_r is the counted number of lags where the correlation of E_{xy} is calculated. Status points out whether the positive envelope appeared or not during pitch search up to present lag.

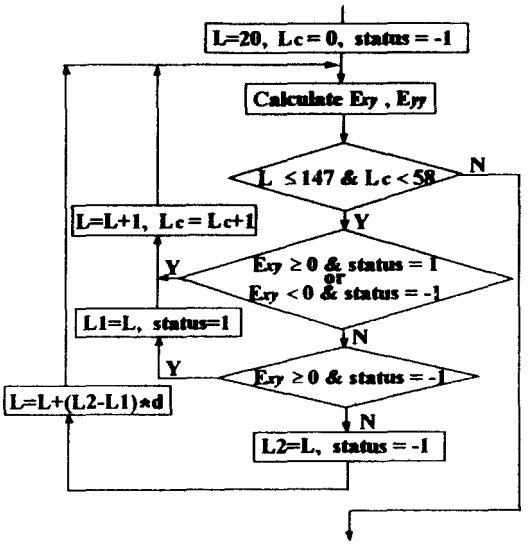

Fig 2. Proposed pitch search algorithm.

IV. EXPERIMENTAL RESULT

For experiment, phonetic balanced five Korean sentences pronounced five times by three male speakers and two female speakers were used for test data base. The speech signal was sampled at 8 kHz and lowpass filtered at 4 kHz and digitized with a 16 bits A/D converter. We used a 20 ms frame size with four 5 ms subframes. For spectrum analysis, 10th order autocorrelation LPC analysis using no preemphasis with a 20 ms Hamming window was performed on every frame by one-loop. In pereceptual weighting, we choose $\alpha = 0.8$ and, in pitch search, lags from 20 to 147 were searched. Under the above condition, the segmental SNR and mean of computation time reduction ratio between the conventional full search method and the proposed method with various d and L_T were obtained. The average required computation time of the conventional method and the proposed one for test speech data were measured and compared on personal computer (IBM 486DX-II).

From the results for various d with no restriciton on L_T as shown in Table 1, d can be determined empirically as 1.2. This is reasonable because too large d may skip even the next positive peak point so to introduce the distortion of speech due to the missing of proper pitch lag and, also, too small d may include many needless lags which have negative correlation value so to deteriorate the time

Table 1. Comparison result between the conventional full search method and the proposed method with various values of d.

Sentence number	Conventional full search (dB)	Proposed method (dB)				
		d = 1.0	d = 1.2	d = 1.5	d = 2.0	d = 2.5
S1	11.67	11.64	11.45	10.18	8.86	9.01
S2	12.41	12.32	12.19	11.49	10.86	9.46
S3	11.95	11.95	11.95	11.79	11.30	10.62
S4	12.03	11.95	11.94	11.44	10.90	9.58
S5	11.52	11.51	11.51	11.34	11.33	11.28
Mean of SNR(dB)	11.92	11.87	11.81	11.25	10.65	9.99
Mean of time reduction ratio(%)	–	31.3	34.4	38.2	42.3	47.2

Table 2. Comparison result between conventional full search method and proposed method with various values of L_f when $d = 1.2$

Sentence number	Convenlional full search (dB)	Proposed method (dB)					
		$L_f = 74$	$L_f = 64$	$L_f = 58$	$L_f = 50$	$L_f = 40$	$L_f = 30$
S1	11.67	11.20	11.10	11.10	10.88	9.83	9.56
S2	12.41	12.17	12.03	11.92	11.90	11.88	9.30
S3	11.95	11.86	11.86	11.86	11.85	11.54	11.46
S4	12.03	11.93	11.89	11.78	11.97	11.61	11.58
S5	11.52	11.51	11.51	11.45	11.42	11.39	11.17
Mean of SNR(dB)	11.92	11.74	11.68	11.62	11.60	11.25	10.61
Mean of time reduction ratio(%)	–	40.8	46.9	51.3	572.	64.8	72.2

L_T = maximum threshold of L_C(samples/frame)

reduction efficiency in computations. With the above method in which L_T is not considered, the time reduction was achieved up to 30% with around 0.1 dB of degradation of segmental SNR.

As shown in Table 2 (d is set on 1.2), by limiting the number of candidate lags, the required computation time can be more reduced up to 51% while the SNR degradation is less than 0.3 dB. As considered before, L_f may be proper to set on 58. However, empirically, smaller L_f up to 50 can be suitable for the proposed algorithm with 0.5 dB of degradation.

V. CONCLUSION

In this paper, we proposed a simple method which preserves the quality of CELP vocoder with reduced complexity. The basic idea is that, by restricting the pitch search range to positive side of envelope in the correlation function and also limiting even the number of available lags to be searched by proper small number, the required pitch searching time can be greatly reduced with negligible degradation of speech quality. Those can be achieved by using several characteristics of speech signal such that the envelope of correlation function of speech signal varies slowly and the positive and the negative envelopes alternatively appear with maintaining the width of the previous envelope in a sufficiently short interval due to the first formant of voiced speech. Employing the proposed method, we can get approximately 51% complexity reduction in the pitch search. Since the proposed method is performed in each subframe, great reduction of computational complexity can be expected when it combined with the delta search method.

REFERENCES

1. B. S. Atal and M. R. Schroeder, "Stochastic Coding of Speech at Very Low Bit Rates," Proceedings of ICC, pp. 1610-1613, 1984.
2. M. R. Schroeder and B. S. Atal, "Code-Excited Linear Prediction(CELP) : High Quality at Low Bit Rates," Proceeding of ICASSP85, pp.937-940, 1985.
3. J. P. Campbell Jr., V. C. Welch and T. E. Tremain, "An Expandable Error-Protected 4800bps CELP Coder (U.S. Federal Standard 4800bps Voice Coder)," Proceedings of ICASSP 89, pp. 735-738, 1989.
4. R. P. Ramachandran and P. Kabal, "Pitch Prediction Filter in Speech Coding," IEEE Trans. on Acoustics Speech and Signal Processing, vol. ASSP-37, no.4., pp.467-478, April 1989.
5. R. Rose and T. Barnwell, "Quality Comparison of Low Complexity 4800 bps Self Excited and Code Excited Vocoders," Proceedings of ICASSP87, pp. 1637-1640, 1987.
6. D. Lin, "Speech Coding Using Efficient Pseudo Stochastic Block Codes," Proceedings of ICASSP

87, pp.1354-1357, 1987.
7. Vocoder software : high level design. Qualcomm inc., 1992.
8. J. Menez, C. Galand, M. Rosso, and F. Bottau, "Adaptive Excited Linear Predictive Coder (ACE-LPC)," Proceedings of ICASSP89, pp.132-135, 1989.
9. L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signal*, Prentice Hall, 1978.

▲JooHun Lee : Department of Electronics Engineering, Seoul National University
▲MyungJin Bae : Vol.13, No.1E 참고
▲SouGuil Ann : Vol.11, No.4 참고