Pitch Detection of Speech Signals using the Cross-Correlation Technique

Salina Abdul Samad, Aini Hussain and Low Kok Fah
Multimedia Signal Processing Research Group
Dept. of Electrical, Electronic and Systems Engineering
Faculty of Engineering
Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor
Malaysia
e-mail: sas@ieee.org

Abstract : To generate high quality speech using the Linear Predictive Coding (LPC) technique, a method for detecting pitch contour is critical since the human ear is sensitive to small pitch variation in speech. The auto-correlation method, though simple to implement with digital signal processors (DSPs), can result in perceptible unnaturalness. This paper describes the cross-correlation technique that can be used to obtain the pitch information more accurately than the auto-correlation method for certain speech samples. Experimental results will illustrate the pitch contour detected using both techniques. In general, the cross-correlation method generates less error than the auto-correlation method for pitch determination in a LPC scheme while having the advantage of requiring less computation.

Keywords:
Speech processing, pitch detection, Linear Predictive Coding.

I. INTRODUCTION

In the classical model of speech production, a source is passed through a filter with the vocal tract response to produce speech. The simplest implementation of this is known as the Linear Predictive Coding (LPC) synthesizer. At every frame, the speech is analyzed to compute the filter coefficients, the energy of the excitation, a voicing decision, and a pitch value if voiced.

In order to generate high quality speech with voice synthesis method that is based on LPC, a method for detecting the pitch contour is critical since the human ear is sensitive to small pitch variation in speech. There are a variety of methods and algorithms which serve this purpose such as the auto-correlation function, average magnitude difference function (AMDF) and simplified inverse filter tracking (SIFT) [1]-[4].

For implementation with digital signal processors, the classic auto-correlation technique is highly suitable due to its multiply and add operations. This method can detect the high pitch frequency with good results; however, it is well known that it cannot provide good results for detecting the low pitch frequency in a male voice. This paper describes the cross-correlation technique that can be used to obtain the pitch information more accurately than the auto-correlation method for certain speech samples. Fewer calculations are afforded for pitch detection since the cross-correlation technique uses a fraction of a frame length as one of its signal components.

II. PITCH DETECTION

The auto-correlation and cross-correlation techniques detect the peak amplitude in the function to estimate the pitch period. The second largest peak relative to the true origin, or zero delay of the correlation, is used to determine the pitch period. One of the advantages of using the cross-correlation function is that correlation peaks in the function tend to remain large and can be easily detected. On the other hand, the peaks in the auto-correlation function tend to fall off linearly starting from the first peak.

Equation (1) defines the auto-correlation function, ACF.

$$ACF_\tau = \frac{1}{L}\sum_{j=1}^{L} S_j S_{j-\tau} \qquad \tau = 0,1,2,...\tau_{max} \qquad (1)$$

where

S_j = j-th sample of the speech waveform ($S1,S2,S3,...S_L$)
L = size of speech segment
τ = delay
τ_{max} = maximum delay shift ($\tau_{max} = L$)

The cross-correlation function, CCF, is defined by the following equation.

$$CCF_{\tau} = \frac{1}{L'} \sum_{j=1}^{L'} S_j' S_{j-\tau}, \quad \tau = 0,1,2,...\tau_{max} \qquad (2)$$

where

S_j = j-th sample of the speech waveform ($S_1,S_2,S_3...S_L$)
S_j' = j-th sample of the sub- interval speech waveform, ($S_1,S_2,S_3,...S_{L'}$)
L' = size of speech segment, (L' < L)
τ = delay
τ_{max} = maximum delay shift ($\tau_{max} \leq L - L'$)

The combination of speech segment length L and the sub-interval speech segment length L' must be suitably chosen. Usually, the value for the length of the sub-interval speech segment L' is between one to two times the length of the expected maximum pitch period. The cross-correlation function has relatively less computation requirement compared to the auto-correlation function. The cross-correlation computation is up to the length of sub-interval speech segment L' which is less than the whole length of the speech segment L. Fig. 1 shows the comparisons between the auto-correlation and the cross-correlation for pitch determination.

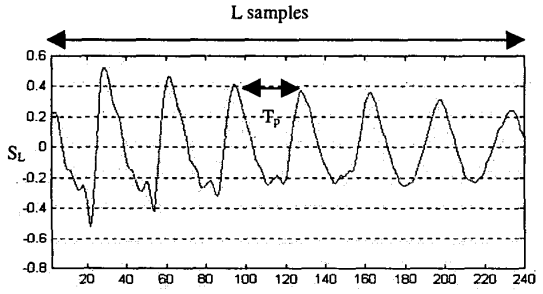

Fig. 1(a): A sample speech segment with L samples

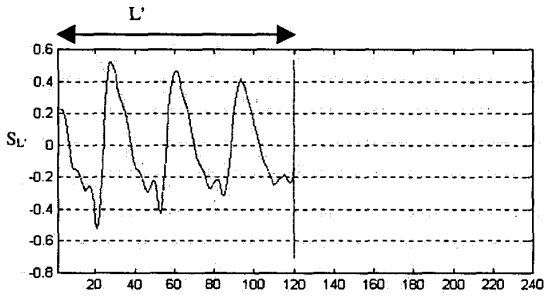

Fig. 1(b): Sub-interval containing L' samples from the speech segment

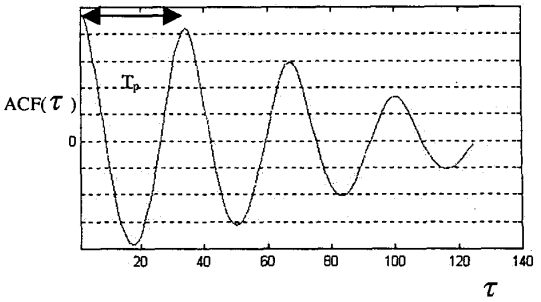

Fig. 1(c): Part of the auto-correlation function of S_L

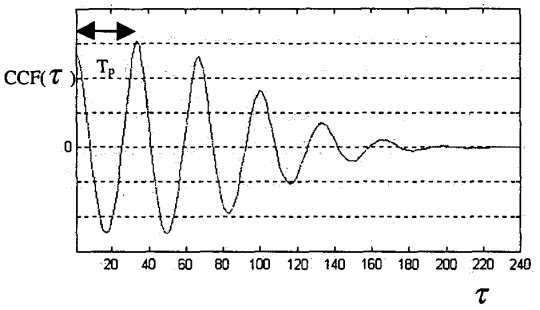

Fig. 1(d): Part of cross-correlation function of S_L and $S_{L'}$

Fig. 1: Comparisons between auto-correlation and cross-correlation.

III. RESULTS AND DISCUSSION

To compare the performance of the cross-correlation with the auto-correlation technique, the algorithms are tested using a set of speech waveforms. The sample words are recorded with an 8 kHz sampling frequency, and with 16 bits representing each sample. The preprocessing stage involves a low-pass filter at a cut-off frequency of 900 Hz to approximately select the first formant of the speech and remove the higher formants that tend to reduce the accuracy of the detected pitch [5].

The speech data are divided into several frames using the Hamming window. A linear prediction order of 10 is used for both pitch detection schemes with identical voiced-unvoiced decision algorithm. The method used to detect the pitch is a simple search for a second peak starting from the first maximum peak. No correction algorithm is applied to correct pitch values that are wrongly detected.

The following are the results obtained with speech samples from male speakers. The results for female speakers are similar using both techniques due to the inherently higher pitch in a female voice, and are not shown.

Fig. 2 and Fig. 3 show the detected pitch for the word *'satu'* of a male speaker with the auto-correlation and cross correlation method, respectively. The frame size is 180 samples while the sub-interval speech segment length is 120. Fig. 4 and Fig. 5 show the detected pitch for the word *'empat'*. The frame size in this case is 240 samples while the sub-interval speech segment length is kept at 120. Fig. 6 and Fig. 7.show the results of the pitch detected using the auto-correlation and cross correlation method, respectively, for the word *'lapan'*. In this case, the sub-interval speech segment length and the length of the frame size are 100 and 180 samples, respectively.

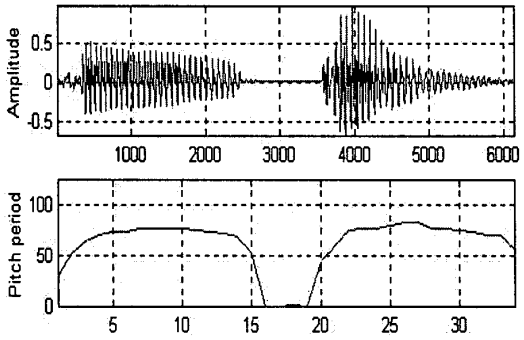

Fig. 2: The pitch detected using auto-correlation for the word *'satu'*

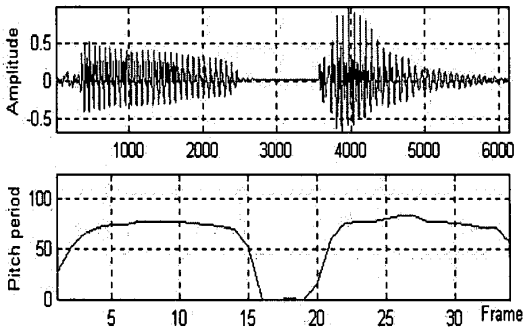

Fig. 3: The pitch detected using cross-correlation for the word *'satu'*

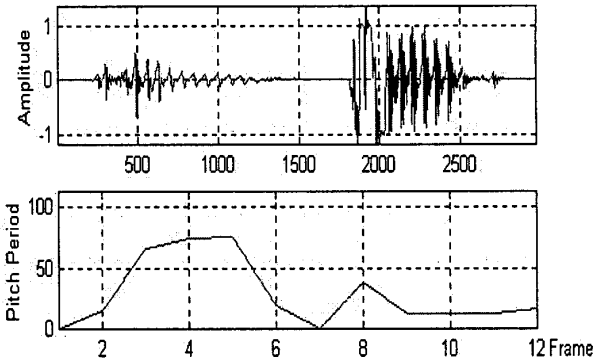

Fig. 4: Pitch detected using auto-correlation for the word *'empat'* .

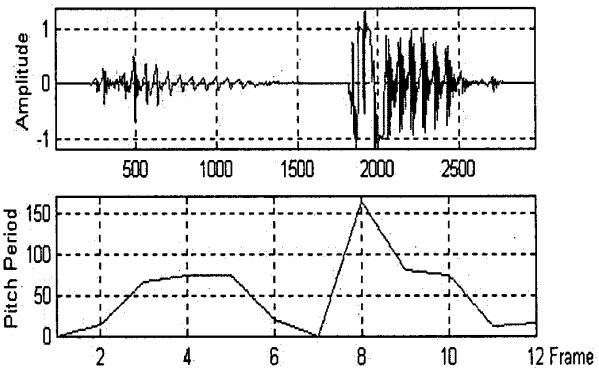

Fig. 5: Pitch detected using cross-correlation for the word *'empat'* .

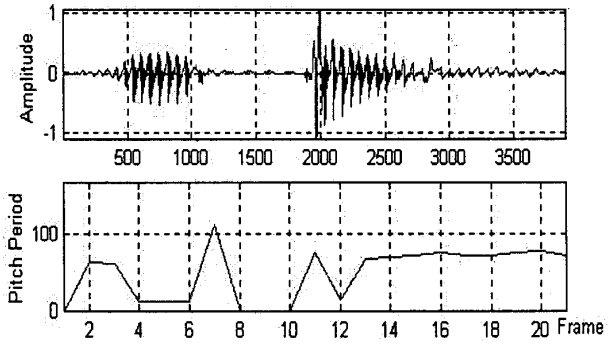

Fig. 6: Pitch detected using auto-correlation for the word *'lapan'*

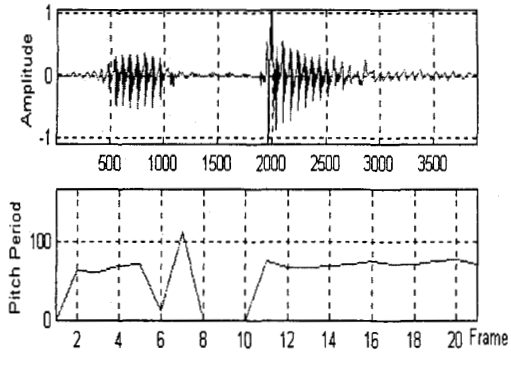

Fig. 7: Pitch detected using cross-correlation for the word '*lapan*'

In general, the cross correlation method performs better than the auto-correlation although both methods have difficulties detecting pitch at certain regions. Fig. 2 and Fig. 3 show that both the auto-correlation and the cross-correlation methods can detect the correct pitch period for the speech word '*satu*' of a male speaker. Both figures show the same contour pattern for the detected pitch period. There is little difference in the value of the pitch period detected at certain frames. At the region where the transition from voiced-unvoiced or unvoiced-voiced in the speech segment occur, the difference of the pitch is often neglected. In this region it is difficult to estimate the pitch since in some cases there is no significant information present for pitch detection. This phenomenon occurs at frame number 20 where a part of the speech signal in the frame is unvoiced and the other part is voiced.

Fig. 4 and Fig. 5 show that the resulting pitch detected using the auto-correlation and cross-correlation for the word '*empat*'. Fig. 4 shows that errors in the detected pitch occur at frame numbers 8 to 11 using the auto-correlation method. The pitch is detected correctly for frame numbers 9 and 10 by the cross-correlation method as shown in Fig. 5. Both methods fail to give good estimate for frame numbers 8 and 11 which are in the unvoiced-voiced transition region. Less error is encountered in estimating the pitch with the cross-correlation method compared to the auto-correlation method.

Fig. 6 is the result for the pitch detected using auto-correlation for the word '*lapan*'. Frame number 6 is where the voiced-unvoiced transition region occurs and both pitch detection algorithms fail to provide good estimate. The auto-correlation method fails to detect the pitch at frames 4,5,6, and 12. The cross-correlation correctly detect the pitch for frames 4, 5 and 12 as shown in Fig. 7.

IV. CONCLUSIONS

This paper has demonstrated the cross-correlation method with respect to the classic auto-correlation method for pitch detection of speech signals. The cross-correlation method has the advantage of having less computation since only a fraction of the window length is used as one of the signal components. For female speakers, who inherently have higher pitch than male speakers, both methods give similar outcomes. For male speakers, generally better results are obtained during pitch detection with the cross-correlation method. This suggests that the cross-correlation technique is better at detecting lower pitch in speech signals compared to the auto-correlation technique.

However, both methods described in this paper are not robust enough to detect the pitch in some region of speech, especially for speech with very low frequency and with rapid changing pitch. The pitch detection algorithm can be improved by adding correction algorithm to correct the error of the detected pitch period. Accurate and reliable pitch detection cannot be easily achieved since the speech waveform is not always periodic and sometimes the pitch varies within the duration of the frame.

V. REFERENCES

[1]. M.J.Ross, H.L.Shaffer, A. Cohen, R. Freudberg, and H.J. Manley, "Average Magnitude Difference Function Pitch Extractor", IEEE Trans. Acoustic, Speech, and Signal Processing, pp. 353-362 Oct. 1974.

[2] D. Takin, "A Robust Algorithm for Pitch Tracking (RAPT)", Speech Coding and Synthesis, Netherlands: Elsevier Science. 1995.

[3] I.A. Atkinson, M. Kondoz and B.G. Evans, "Time Envelop Vocoder, A New LP Based Coding Strategy for Use of Bit-Rate 2.4kb/s and Below", IEEE Journal on Selected Areas on Communications, Vol. 13, No. 2, Feb. 1995.

[4] A. M. Kondoz, Digital Speech: Coding for Low Bit Rate Communications Systems, Wiley, England. 1995.

[5] B.Gold and L. Rabiner, "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain", J. Acoust. Soc. Amer., vol. 46, pp. 442-448, Aug. 1969.