A Pitch Synchronous Method for Speech Modification

Chih-Ting Kuo, Hsiao-Chuan Wang
Department of Electrical Engineering, National Tsing Hua University, Hsinchu

Abstract-- **The speech modification is a mechanism of changing speech characteristics and prosody for some specific applications. It is used in voice conversion, pronunciation correction, tone perception, and language learning. The most important part is the change of pitch in an utterance. Pitch extraction is an essential process for speech modification. This paper presents an efficient pitch extraction algorithm based on the normalized second standard deviation function (NSSDF) of magnitude difference. A pitch synchronous method for modifying speaking rate and pitch trajectory is proposed. The speaking rate is modified by inserting or deleting pitch periods in voiced segments. The pitch trajectory change is accomplished by modifying the pitch period of residual signal obtained from pitch synchronous linear prediction (LP) analysis and reconstructing speech signal by LP filter. A speech modification system is developed for Mandarin perception which is used to help hearing impaired students in pronunciation learning.**

Keywords— speech modification, pitch extraction, pitch trajectory change, pitch synchronous linear prediction, Mandarin perception

I. INTRODUCTION

The speech modification is a process to change speech prosody or convert the speech characteristics for some specific applications [1][2]. For example, a Mandarin language learning system may need to change speaking rate so that the pronunciation is much clear, or to change pitch trajectory so that the Mandarin tones can be perceived. Pitch extraction is a necessary process for speech modification. The pitch detection is a basic operation to extract prosodic information, and has been developed in many ways [3]. An efficient method to estimate fundamental frequency of voiced segments is still attractive to many researchers.

In this paper, we review the autocorrelation function (ACF) method [4], the average magnitude difference function (AMDF) method [5], and YIN method proposed by Cheveigne and Kawahara [6]. Then a new method based on the normalized second standard deviation function (NSSDF) of magnitude difference is proposed. It shows the advantage of being able to reliably extract pitch period in frame bases even when the speech signal is drifted temporally. Finally, a speech modification system is developed for Mandarin perception which is used to help hearing impaired students in pronunciation learning.

II. PITCH EXTRACTION

Many methods have been proposed for extracting pitch information of the speech signal. Typical methods are based on the autocorrelation function (ACF) and the average magnitude difference function (AMDF) [4][5]. They are calculated for each frame with fixed length. For example, the speech signal is sampled at 16 kHz. The frame length is 256-point and with 128-point shift. Since the fundamental frequency of speakers is in the range of 80 ~ 400 Hz, a frame of speech contains at least one pitch period. ACF and AMDF are defined by the following equations,

$$ACF(\delta) = \sum_{n=0}^{N-1} x(n)x(n+\delta) \tag{1}$$

and

$$AMDF(\delta) = \sum_{n=0}^{N-1} (|\, x(n) - x(n+\delta)\,|) \tag{2}$$

where N is the frame length, n is the time index in a frame, δ is the shift, and $x(n)$ is the sampled signal.

It is clear that the pitch period can be estimated by finding the maximum on ACF or the minimum on AMDF. Usually, the double or half pitch periods will be extracted. It needs extra processes to find the pitch period. Cheveigne and Kawahara proposed a normalized square difference function [6] to estimate the pitch period. They called the function YIN. The computation equations are as follows.

$$DF(\delta) = \sum_{n=0}^{N-1} (x(n) - x(n+\delta))^2 \tag{3}$$

and

$$YIN(\delta) = \begin{cases} 1, & \delta = 0 \\ \dfrac{DF(\delta)}{\dfrac{1}{1+\delta}\sum_{\tau=1}^{\delta} DF(\tau)}, & \delta \neq 0 \end{cases} \tag{4}$$

The pitch period can be estimated by finding the minimum on YIN. Because of the normalization, a threshold for detecting peaks is much easier to be determined. Figure 1 shows the comparison of ACF, AMDF, and YIN.

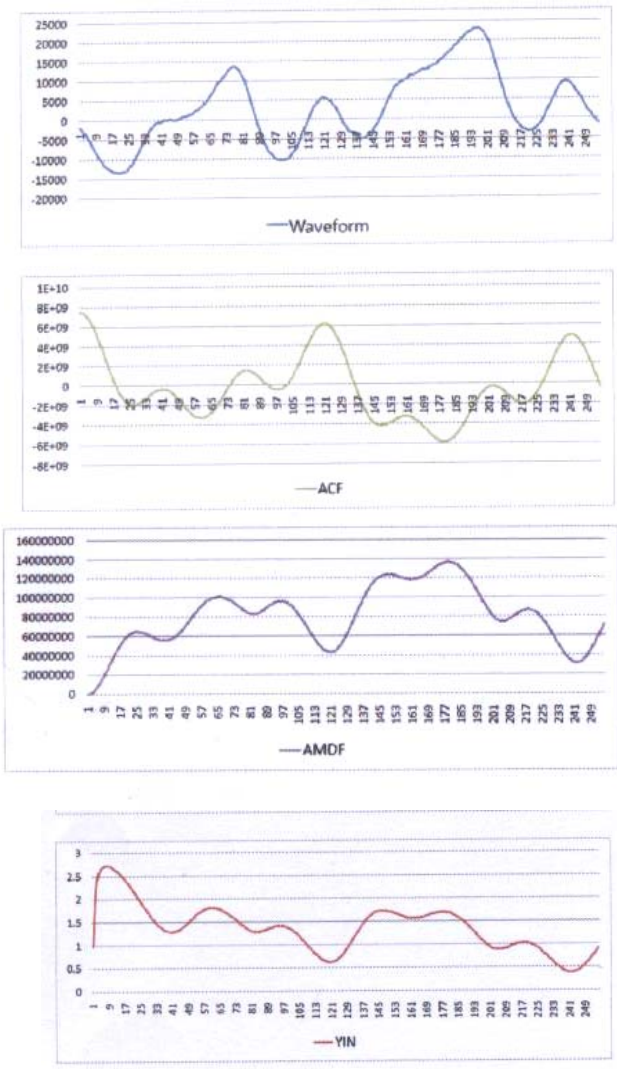

Figure 1. Comparison of ACF, AMDF, and YIN of a voiced speech signal.

Usually, the level drift of signal may cause difficult to find reliable peak locations for estimating the pitch period. In this paper a new method based on the normalized second standard deviation function (NSSDF) of magnitude difference is proposed. The standard deviation is given by

$$SD(\delta) = \left(\frac{1}{N-1}\sum_{n=0}^{N-1} \left(Dif(n,\delta) - \overline{Dif(\delta)}\right)^2\right)^{1/2} \quad (5)$$

where

$$Dif(n,\delta) = x(n) - x(n+\delta) \quad (6)$$

and

$$\overline{Dif(\delta)} = \frac{1}{N}\sum_{n=0}^{N-1} Dif(n,\delta) \quad (7)$$

The second standard deviation is then calculated by

$$SSD(\delta) = \left(\frac{1}{N-1}\sum_{n=0}^{N-1} \left(SDif(n,\delta) - \overline{SDif(\delta)}\right)^2\right)^{1/2} \quad (8)$$

where

$$SDif(n,\delta) = SD(n) - SD(n+\delta) \quad (9)$$

and

$$\overline{SDif(\delta)} = \frac{1}{N}\sum_{n=0}^{N-1} SDif(n,\delta) \quad (10)$$

The same normalization as YIN is used to obtain

$$NSSDF(\delta) = \begin{cases} 1, & \delta = 0 \\ \dfrac{SSD(\delta)}{\dfrac{1}{1+\delta}\sum_{\tau=1}^{\delta} SSD(\tau)}, & \delta \neq 0 \end{cases} \quad (11)$$

Since the mean is removed during the computation, the effect due to signal drift is minimized. Figure 2 shows the NSSDF of a voiced speech signal.

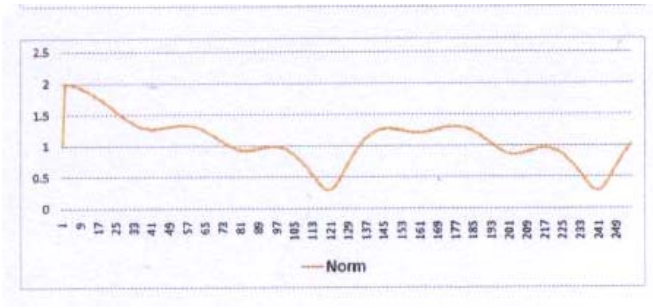

Figure 2. NSSDF of a voiced speech signal.

To show the effectiveness of our proposed method, a simple experiment was conducted. The testing data were sampled at 16 kHz and windowed in 256-point length (16 ms). Each pitch period in a testing utterance was manually labeled. Four pitch detection methods, ACF, AMDF, YIN, and NSSDF, were evaluated. For the speech in a window, ACF, AMDF, YIN, or NSSDF method was applied to get the corresponding function curve. A search process along the curve was performed starting from the point of 2 ms (32 points). When a peak (for ACF) or valley (for AMDF, YIN, NSSDF) was found at point n_p, it continued to search the following $0.8\times n_p$ region for another bigger peak or valley. If no other bigger peak or valley could be found, n_p was indicated the pitch period. No further smoothing algorithm on the estimated pitch periods was applied. The estimated pitch period was compared with the reference. The reference pitch period was the average of three consecutive periods, i.e., the previous, the current, and the next pitch periods of manually labeled data. If the estimated period was different from the reference within 5% of the reference period length, the estimate was considered to be correct. Table 1 gives the experiment results.

Table 1. Comparison of four pitch detection methods

Accuracy of estimated pitch periods (%)				
	ACF	AMDF	YIN	NSSDF
Utterance#1	95.8	94.4	97.2	98.6
Utterance#2	79.0	82.5	86.0	93.0
Utterance#3	73.1	74.6	79.1	85.1
Utterance#4	80.3	77.3	83.3	89.4
Utterance#5	81.5	85.2	85.2	92.6

** Utterances#1,#4 – female, Utterances#2,#3,#5 -- male

The result shows that NSSDF is more reliable than YIN in pitch period detection.

Only those voiced frames can get pitch periods. Then the consecutive voiced frames form a voiced segment. With the help of pitch information, we search for the zero-crossing point just before the maximum amplitude in a pitch period, and set this point as the beginning of a pitch period.

III. SPEECH MODIFICATION

In Mandarin, the speech modification includes the volume adjustment, the speaking rate change, the pitch modification, and the tone modification.

A. Volume Adjustment

The volume adjustment is accomplished by scaling the signal amplitude. A clipping method is applied to avoid the overflow of binary coding of sampled signal amplitudes. This function can be used to strengthen a syllable so that a specific stress pattern is obtained.

B. Speaking Rate Change

In the speaking rate change, a scaling factor for changing the utterance duration is given. The duration of each silence portion is proportionally scaled according to the scaling factor. This operation is to properly insert or delete silence frames so that the duration of silence portion is changed. For those voiced segments, the insertion or deletion of pitch periods is equally distributed over the voiced segment to slow down or speed up the speaking rate. The unvoiced segments are kept unchanged.

C. Pitch Synchronous Process

The voiced segment is formed by consecutive voiced frames where the beginning point of each pitch period is labeled during the pitch extraction. Those operations related to pitch period changing are performed in a pitch synchronous manner. A consecution of two pitch periods is used for the linear prediction (LP) analysis. A set of 12 LP coefficients is calculated for two consecutive pitch periods. By using these LP coefficients, we construct the LP filter and its inverse filter. Taking the speech signal of corresponding pitch period as the input to the inverse filter, we can obtain the residual signal. The pitch period change is performed on the residual signal instead of the speech signal. After the pitch period of residual signal is modified, the modified residual signal is the input to its corresponding LP filter so that the speech signal is reconstructed. A smoothing process is necessary to remove those discontinuities between adjacent pitch periods. Figure 3 presents the block diagram of this pitch synchronous process.

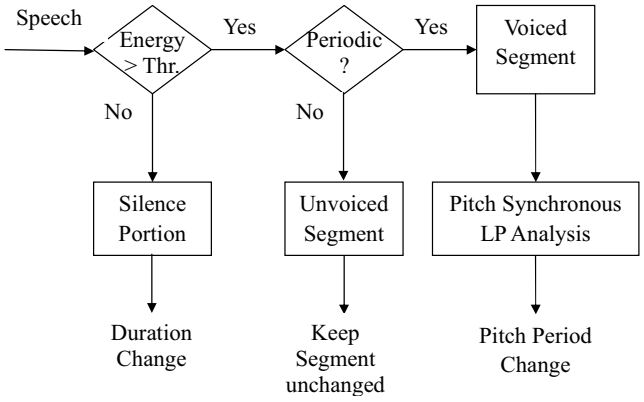

Figure 3. Pitch synchronous process

The procedure for LP analysis and pitch period change is shown in Figure 4.

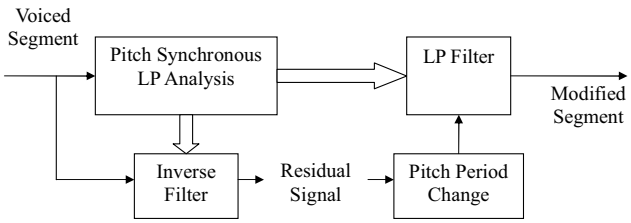

Figure 4. LP analysis and pitch period change

D. Pitch Modification

The pitch modification is performed on those voiced segments only. The operation is done on the residual signal. A pitch period is shortened by discarding the ending potion of a pitch period or lengthened by inserting small random values to the end of a pitch period according to a given scaling factor. Then the modified residual signal is input to its corresponding LP filter to generate speech signal. It results in the raising or lowering the pitch of speech signal. Since we want to keep the speaking rate unchanged, the process to maintain the duration of voiced segment is required. This can be obtained by using the process as being described in section *B*. The reason of applying LP analysis and using LP filter to reconstruct speech signal is that we need to keep the spectral envelope of original speech signal unchanged.

E. Tone Modification

In Mandarin, the tone is an essential element of a syllable. The perception of Mandarin tone is mainly depending on the

fundamental frequency (F0) trajectory in a syllable. When we change the pattern of F0 trajectory, we may change the tone of a syllable. In some applications, such as the Mandarin pronunciation learning, we may need a tone change function to allow students perceiving different tones of a Mandarin syllable. A set of tone patterns is defined by a set of F0 trajectories. When we modify pitch periods to fit a given pattern of F0 trajectory, a specific tone is generated. The most important fact is to maintain the original spectral envelope in the voiced segments. The LP analysis approach ensures that the desired property can be accomplished. If we change the trajectory of a tone pattern, we can change the tone, or strengthen a tone for perception test.

IV. A Speech Modification System for Mandarin Perception

A speech modification system is developed for Mandarin perception. The system is designed for hearing impaired students to help them in the pronunciation learning. It provides a tool for the tutor to prepare teaching materials. During the teaching-learning process, the speech extracted from a recorded file can be modified to adjust its volume, change the speaking rate, modify its pitch, and change the tone of a specific syllable. The desired pitch pattern can be selected from default setting, or manually determined by manipulating the pitch contour on the F0 trajectory shown in the screen. Figure 5 shows the example of tone modification. The second syllable of a Mandarin word has been changed from first tone to second tone.

V. Conclusion

This paper introduces a new pitch extraction method. It has demonstrated the advantage of reliable pitch period estimation. The pitch synchronous linear prediction (LP) analysis is proposed for the pitch modification of speech signal. It changes pitch trajectory on the residual signal and reconstruct the speech signal using LP filter. A speech modification system is developed for Mandarin perception. It is helpful to hearing impaired students in pronunciation learning. The functions include the volume adjustment, the speaking rate change, the pitch modification, and the tone modification.

Acknowledgement

This research was sponsored by the National Science Council, Taiwan, under contract number NSC96-2221-E-007-144.

References

[1] S.S. Nagarajan, X. Wang, M.M. Merzenich, C.E. Schreiner, P. Johnston, W.M. Jenkins, S. Miller, and P. Tallal, "Speech modification algorithm used for training language learning – impaired children," *IEEE Trans. Rehabilitation Engineering*, v.6, no.3, pp. 257-268, 1998.

[2] Y. Nejime, T. Aritsuka, T. Imamura, T. Ifukube, and J. Matsushima, "A portable digital speech-rate converter for hearing impairment," *IEEE Trans. Rehabilitation Engineering*, v.4, no.2, pp. 73-83, 1996.

[3] D. Gerbard, *Pitch Extraction and Fundamental Frrequency: History and Current Techniques*, Technical Report, TR-CS 2002-06, 2003, University of Regina, Canada.

[4] D. O'Shaughnessy, *Speech Communications: Human and Machine*, 2nd edition, IEEE Press, 2000.

[5] M. Ross, H. Shafer, A. Cohen, R. Freuberg, and H. Manley, "Average magnitude difference function pitch extractor," *IEEE Trans. ASSP*, v.22, pp. 353-362, 1974.

[6] A. D. Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *.J. Acoustical Society of America*, v.111, no.4, pp. 1917-1930, 2002.

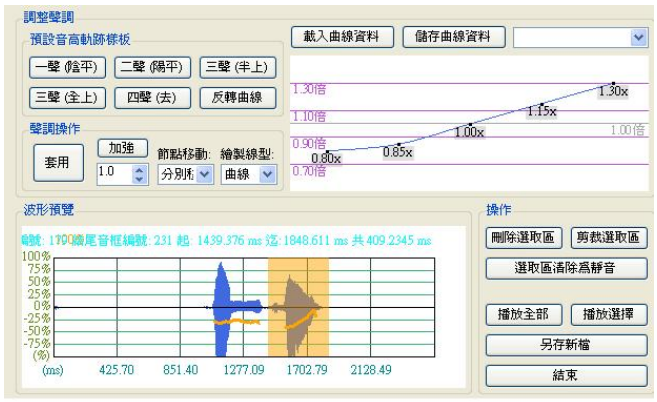

Figure 5. Example: the second syllable has been changed from first tone to second tone