Computational Linguistics and Chinese Language Processing
Vol. 6, No. 2, Feburary 2001, pp. 31-42
© Computational Linguistics Society of R.O.C.

Pitch Marking Based on an Adaptable Filter and a Peak-Valley Estimation Method

Jau-Hung Chen and Yung-An Kao[*]

Abstract

In a text-to-speech (TTS) conversion system based on the time-domain pitch-synchronous overlap-add (TD-PSOLA) method, accurate estimation of pitch periods and pitch marks is necessary for pitch modification to assure optimal quality of synthetic speech. In general, there are two major tasks in pitch marking: pitch detection and location determination. In this paper, an adaptable filter, which serves as a bandpass filter, is proposed for use in pitch detection to transform voiced speech into a sine-like wave. The pass band of the adaptable filter can be adapted based on the fundamental frequency. Based on the sine-like wave, a peak-valley decision method is proposed to determine the appropriate parts (positive part and negative part) of voiced speech for use in pitch mark estimation. In each pitch period, two possible peaks/valleys are searched, and dynamic programming is performed to obtain pitch marks. Experimental results indicate that our proposed method performs very well if correct pitch information is estimated.

1. Introduction

In past years, the concatenative synthesis approach has been adopted for use in many text-to-speech (TTS) systems [Hamon *et al.* 1989][Iwahashi *et al.* 1995][Shih *et al.* 1996][Chen *et al.* 1998][Chou *et al.* 1998][Charpentier *et al.* 1986]. Concatenative synthesis uses real recorded speech segments as synthesis units and concatenates them together during synthesis. In addition, the time-domain pitch-synchronous overlap-add (TD-PSOLA) [Charpentier *et al.* 1986] method has been employed to perform prosody modification. This method modifies the prosodic features of a synthesis unit according to the target prosodic information. Generally, the prosodic information of a speech unit includes its pitch (the fundamental frequency, f_0), intensity, duration, etc. For a synthesis scheme based on the

[*] Advanced Techonlogy Center, Computer and Communication Research Laboratories, Industrial Technology Research Institute, Chutung 310, Taiwan
Email: chenjh@itri.org.tw, kya@itri.org.tw

TD-PSOLA method, it is necessary to obtain a pitch mark for each pitch period in order to assure optimal quality of synthetic speech. The pitch mark is a reference point for the overlap between speech signals.

A speech synthesizer with various voices is useful for speech synthesis. Sometimes, it is also important for a service-providing company to have a synthesizer with the voice of its own employee or its favorite speaker. For conventional TTS systems, however, it is a demanding and tedious job to create a new voice. Recently, corpus-based TTS systems have been developed which use a large number of speech segments. Some approaches select speech segments as candidates for synthesis units. Establishing synthesis units involves speech segmentation, pitch estimation, pitch marking, and so on. Moreover, pitch marking is very labor-intensive task if no automatic mechanism is available.

In general, there are two major tasks in pitch marking: pitch detection and location determination. Compared to the literature on pitch detection [Rabiner *et al.* 1976][Rabiner 1977][Noll 1967][Markel 1972][Barnard *et al.* 1991][Kadambe *et al.* 1991][Barner 2000][Huang *et al.* 2000], few papers have focused on pitch marking [Moulines *et al.* 1990][Kobayashi *et al.* 1998], which is also a difficult problem because of the great variability of speech signals. Moulines *et al.* [Moulines *et al.* 1990] proposed a pitch-marking algorithm based on the detection of abrupt changes at glottal closure instants. In each period, they assumed that the speech waveform could be represented by the concatenation of the response of two all-pole systems. On the other hand, Kobayashi *et al.* [Kobayashi *et al.* 1998] used dyadic wavelets for pitch marking. The glottal closure instant was detected by searching for a local peak in the wavelet transform of the speech waveform.

In this paper, we propose a pitch-marking method based on an adaptable filter and a peak-valley estimation method. The block diagram of our method is shown in Fig. 1. The input signals are limited to voiced speech because only the periodic parts are of interest. We introduce an adaptable filter, which serves as a bandpass filter, to transform voiced speech into a sine-like wave. FFT (Fast Fourier Transform) is used to transform voice to the frequency domain, and the filter's pass band is determined by finding the spectral peak of the fundamental frequency. Consequently, the pass band can be adapted based on the fundamental frequency. The autocorrelation method is then used to estimate the pitch periods on the sine-like wave. In addition, a peak-valley decision method is employed to determine which part of the voiced speech is suitable for pitch mark estimation. The positive part (the speech with positive amplitude) and the negative part (the speech with negative amplitude) are investigated in this method. This is demonstrated by Fig. 3(a), which shows an example of a waveform having a negative part that reveals explicit periodicity. In general, it is possible to achieve better speech quality if the pitch marks are labeled at the positions of the extreme points (peaks and valleys) of speech. In each pitch period, two possible peaks/valleys are searched. Finally, the pitch marks are obtained through dynamic programming by calculating the degree of pitch distortion.

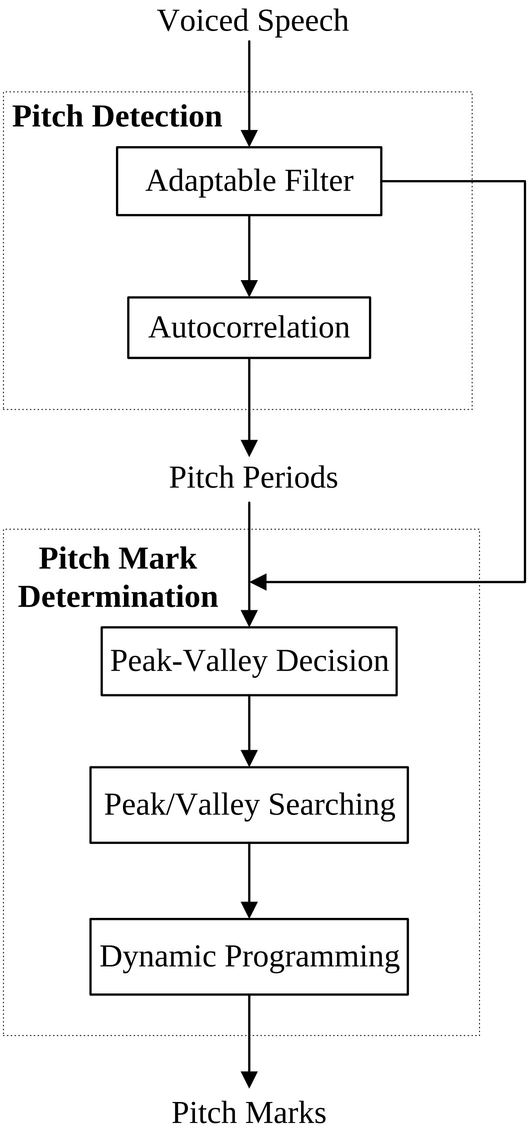

Figure 1 *Block diagram of the proposed pitch-marking method.*

2. Pitch Detection Using an Adaptable Filter Followed by Application of the Autocorrelation Method

The proposed adaptable filter serves as a bandpass filter in which the pass band extends from 50 Hz to the detected fundamental frequency, up to 500 Hz, of the voiced speech. First, we will define the following symbols, which are used in this algorithm:

N: frame size in sample. Consecutive frames do not overlap.

$s_m[n]$: the voiced speech of the m-th frame, where $0 \le n < N$.

$SF_m[k]$: the frequency response of $s_m[n]$, where $0 \le k < N$.

$YF_m[k]$: the pass band frequency response of $SF_m[k]$, where $0 \le k < N$.

$o_m[l]$: the adaptable filter's output signal of the m-th frame, where $0 \le l < N$.

The algorithm of the adaptable filter is described as follows:

Step 1. Use FFT to transform the signal $s_m[n]$ to obtain the frequency response $SF_m[k]$.

Step 2. Find the position k_p of the spectral peak of the fundamental frequency for $SF_m[k]$ by searching the first forty points of $|SF_m[k]|$.

Step 3. Decide on the filter's pass band. Let $YF_m[k]=SF_m[k]$ if $3 \le k \le k_p+2$ or $3 \le N\text{-}k \le k_p+2$; otherwise, let $YF_m[k]=0$.

Step 4. Normalize $YF_m[k]$ by multiplying a scale of $Max_k(|YF_m[k]|)/|YF_m[k_p]|$.

Step 5. Use IFFT (Inverse FFT) to transform the normalized $YF_m[k]$ to the time domain. Let $o_m[n]$ be the real part of the time domain signal.

Finally, the refined pitch periods are obtained by analyzing the filtered speech $o[n]$ using the conventional autocorrelation method. The waveform of $o_m[n]$ after IFFT may be discontinuous at the frame boundaries. A typical example is shown in Fig. 2. However, such waveform discontinuity is not very significant and does not significantly affect the results of pitch period estimation.

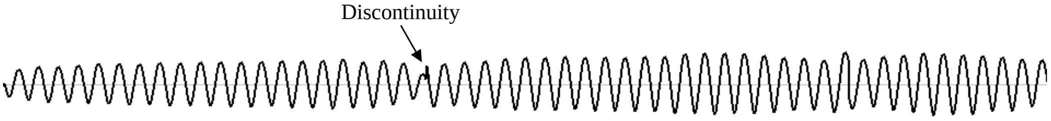

Figure 2 *A typical example of waveform discontinuity after IFFT.*

An example of an adaptable filter is displayed in Fig. 3. Panels (a) and (b) show the waveforms of the original speech and the filtered speech, respectively. It can be seen that the filtered speech is generally a sine-like wave with clear periodicity than the original speech waveform. For a frame in the middle of the voiced speech, the spectral contour is depicted in panel (d). Note that the frequency axis is not linearly plotted to allow inspection of the first spectral peak. The first peak was found at 168 Hz, which was the fundamental frequency. Finally, the pitch periods were obtained by analyzing the filtered speech using the conventional autocorrelation method.

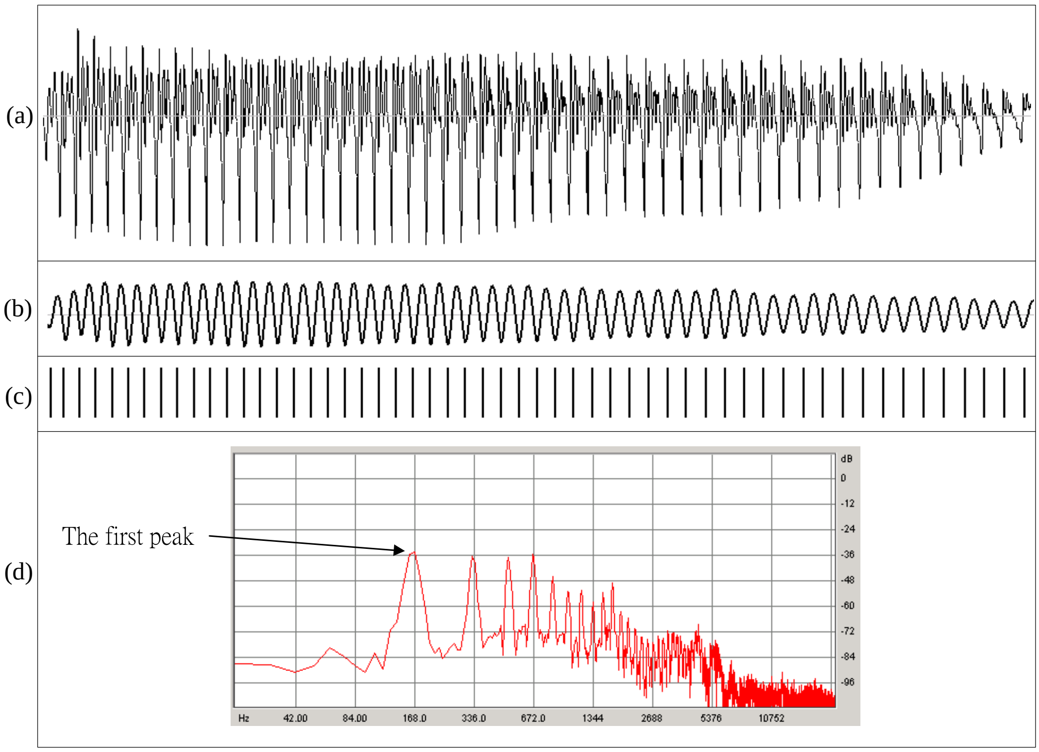

Figure 3 *Results obtained using the adaptable filter and pitch mark determination. (a) Waveform of the voiced speech with explicit periodicity in the negative part. (b) Waveform of the filtered speech. (c) Detected pitch marks. (d) Spectral contour (note that the frequency axis is not linearly plotted).*

3. Pitch Mark Determination Using a Peak-Valley Decision Method and Dynamic Programming

3.1 Peak-Valley Decision

From observations, we have found that voiced speech, $s[\cdot]$, is synchronous with filtered speech, *o*[·], either at peaks or at valleys. The cases illustrated in Figs. 3 (a) and 2 (b) are synchronous at valleys having explicit periodicity instead of at peaks. As a result, the pitch marks can be more easily determined in the negative part than in the positive part. In the following, the peak-valley decision method is used to calculate two costs by summing the amplitudes of *s*[*q*], where *q* represents the position of the local extreme point of *o*[·] over each pitch period:

$$C_{peak} = \frac{1}{N_{peak}} \cdot \sum_{n=1}^{N_{peak}} s[Pos_{peak}[n]], \tag{1}$$

$$C_{valley} = \frac{-1}{N_{valley}} \cdot \sum_{n=1}^{N_{valley}} s[Pos_{valley}[n]], \tag{2}$$

where the symbols are defined as follows:

C_{peak} : cost estimated at the peaks of *o*[·].

C_{valley} : cost estimated at the valleys of *o*[·].

N_{peak} : total number of the peaks of *o*[·].

N_{valley} : total number of the valleys of *o*[·].

$Pos_{peak}[n]$: position of the *n-th* peak of *o*[·].

$Pos_{valley}[n]$: position of the *n-th* valley of *o*[·].

The peak-valley decision is made as follows: If $C_{peak} > C_{valley}$, then the positive part (peak) of *s*[·] is adopted for evaluation of the pitch marks. Otherwise, the negative part (valley) of *s*[·] is adopted.

3.2 Pitch Mark Determination Based on Dynamic Programming

Once the peak or valley, say the peak, has been adopted, the positions of the pitch marks are determined by picking the peaks of *s*[·]. For a speech segment with a length of one pitch period, the PSOLA method can be used to synthesize good quality speech if the pitch mark is denoted at the signal with the largest amplitude. However, the largest peak may not correspond to the largest one in the next period (as shown in Fig. 4). This inconsistency will result in unpleasant speech after the PSOLA method is used. Therefore, the two highest peaks in each period are searched in pitch mark determination. We do not use three peaks or more because this would improve the performance very little. In this paper, we consider that a peak is located at the signal with the largest amplitude among consecutive positive signals. Among peaks, the highest peak is the one with the largest amplitude. The second highest peak is the highest of the two peaks, the left-side and the right-side peaks, neighboring the highest peak.

For the *i-th* pitch period, P_i, suppose the highest and the second highest peaks are located at L_{i1} and L_{i2}, respectively. It might occur that the second one is absent. In this case, we let $L_{i2} = L_{i1}$. For all the detected peaks, pitch mark determination is then performed based on dynamic programming. The distortion of the pitch period, $d_i(j,k)$, and its accumulation, $A_i(j)$, are defined as follows:

$$d_i(j,k) = \left\| L_{ij} - L_{(i-1)k} \right| - P_i \Big| + g(j,k), \text{ for } i=2,\ldots,PN, \tag{3}$$

$$A_i(j) = \min\begin{Bmatrix} d_i(j,1) + A_{i-1}(1), \\ d_i(j,2) + A_{i-1}(2) \end{Bmatrix}, \text{ for } i=2,3,\ldots,PN, \tag{4}$$

where PN is the total number of pitch period and j, k=1,2. In Equation (3), $g(j,k)$ is a penalty function represented by

$$g(j,k) = \begin{cases} 0, \text{if } j = 1 \text{ or } k = 1 \\ \dfrac{1}{\mathrm{PN}}, \text{otherwise} \end{cases} \quad . \tag{5}$$

The penalty function is introduced here due to the preference for the highest peak.

The search path of the dynamic programming is illustrated in Fig. 5. The peak locations (pitch marks) can be obtained by back tracing the peak sequence corresponding to the smallest values of $A_i(1)$ and $A_i(2)$. An example of the results of pitch marking is shown in Fig. 3(c). A procedure similar to that described above can be applied for the case of a “valley.”

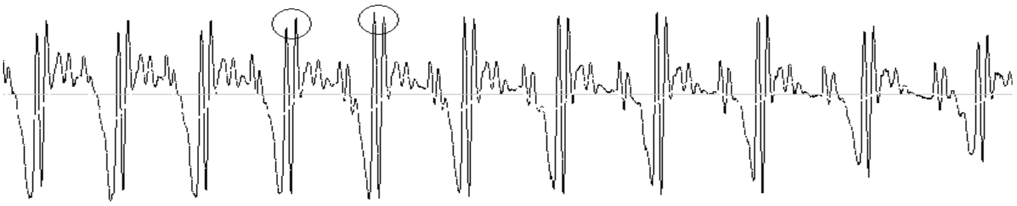

***Figure 4** An example of a waveform (syllable /a/ of tone 3), in which the largest peak does not correspond to the largest one in the next period (indicated by the circles).*

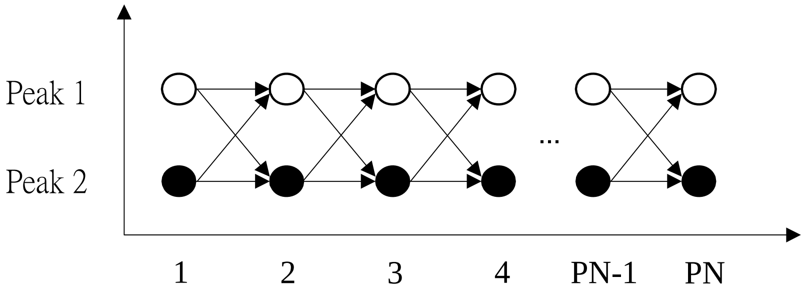

***Figure 5** Illustration of the peak-picking search path of the dynamic programming.*

4 Experiments and Results

4.1 Experimental Environment

A continuous speech database was established which provides the basic synthesis units of our Mandarin Chinese TTS system. This database is composed of 70 phrases, and their lengths are from 4 to 6 Chinese characters. It includes a total of 436 tonal syllables comprising the required 413 basic synthesis units. A native female speaker read them in normal speaking style. The speech signals were then digitized by a 16-bit A/D converter at a 44.1k Hz sampling rate. Syllable segmentation was done manually in order to obtain the precise boundaries of the voiced speech and unvoiced speech. The total duration of the 436 voiced speech segments was about 2.1 minutes. For each syllable, the voiced speech was used to test the proposed methods. The frame size used in the adaptable filter was set to 4096 speech samples (92.8 ms). We used large frame size so that we could deal with signals with very low f_0 values.

For the voiced speech, the waveforms along with the pitch marks obtained using our pitch-marking program were visually displayed. The pitch marks were then checked and corrected by an experienced person through a friendly interface. For evaluation of the experiments, we obtained 436 sets of human-labeled pitch marks, denoted as *H*, which comprises 23,868 pitch marks.

4.2 Performance of the Pitch Marking Method

The peak-valley decision results were verified by human judgment based on visual displays. A success rate of 99.1% was obtained (4 of the 436 results disagreed). For the female speaker, we found that 97.2% of the voiced segments revealed clear periodicity in the negative parts.

The proposed method generated 23,860 pitch marks, denoted as *I*, without any duplication. The success rate of the pitch marking method is calculated as follows:

$$\text{Correct rate} = \frac{\left|\{x \mid x \in I \text{ and } x \in H\}\right|}{|H|} \times 100\% \,. \tag{6}$$

As shown in Table 1, a success rate of 97.2% was obtained (baseline), in contrast with 95% and 97% success rates obtained using the methods proposed in [Moulines *et al*. 1990] and [Kobayashi *et al*. 1998], respectively. Moreover, we found that most of the errors resulted from incorrect pitch detection results. Most of the pitch errors were due to large changes of pitch located at the boundaries of the voiced speech. With correct pitch information provided, our method achieved a success rate of 99.5%.

The tone type of voice significantly affects the results of the detection of f0. The main reason for error detection of f_0 is dependent on the tone types of voice. There are five tones in Mandarin speech, including a high-level tone (Tone 1), a mid-rising tone (Tone2), a mid-falling-rising tone (Tone 3), a high-falling tone (Tone 4), a neutral tone (Tone 5). In our system, it is easy to detect f_0 for tone 1 and tone 2 since the spectral peak of f_0 is prominent (Fig. 3 (d)). For tone 3, tone 4 and tone 5, f_0 may be erroneously detected at the end of the voice segment if the consecutive pitch periods change abruptly (Fig. 6 (a)). For this case, the spectral peak of f_0 is unclear (Fig. 6 (b)), which may result in error detection.

Table 1. *Success rate of the pitch-marking method.*

Condition	Baseline	Using correct pitch
Success rate	97.2%	99.5%

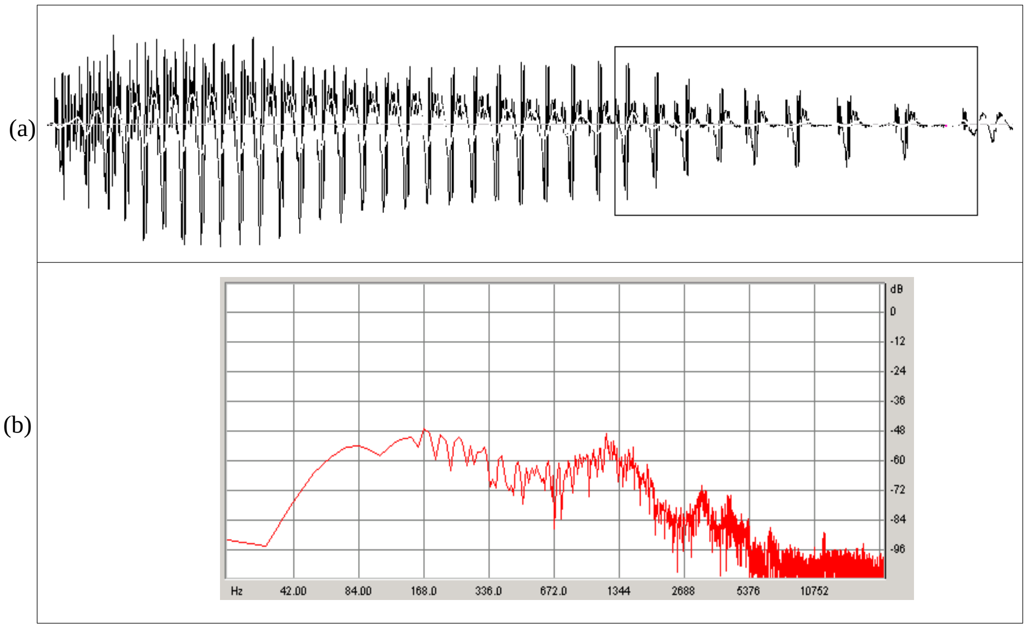

Figure 6 *An example of unclear spectral peaks. (a) Waveform of the syllable /a/ of tone 3. (b) Spectral contour corresponding to the end part of the waveform (note that the frequency axis is not linearly plotted).*

5 Conclusions

In this paper, a preliminary work on pitch marking has been proposed. We have presented an adaptable filter which can be combined with the autocorrelation method to perform pitch detection. On the other hand, a peak-valley decision method has been proposed to select either the positive or the negative part for pitch mark evaluation. Also, a dynamic-programming-based pitch mark determination method has been demonstrated, where two peaks/valleys are searched in each period. In the experiments, our pitch-marking method achieved a 97.2% success rate. Furthermore, a high success rate of 99.5% was obtained when correct pitch information was provided.

Acknowledgement

This paper is a partial result of Project 3XS1B11 conducted by ITRI under sponsorship of the Ministry of Economic Affairs, Taiwan, R.O.C.

References

Hamon, C., E. Moulines, and F. Charpentier, "A diphone synthesis based on time-domain prosodic modifications of speech," *International Conference on Acoustics, Speech, and Signal Processing*, 1989, pp.238-241.

Iwahashi, N. and Y. Sagisaka, "Speech segment network approach for optimization of synthesis unit set," *Computer Speech and Language*, 1995, pp.335-352.

Shih, C. L. and R. Sproat, "Issues in text-to-speech conversion for Mandarin," *Computational Linguistics and Chinese Language Processing*, vol.1, 1996, pp.37-86.

Chen, S. H., S. H. Hwang and Y. R. Wang, "An RNN-based prosodic information Synthesizer for Mandarin text-to-speech," *IEEE Transactions on Speech and Audio Processing*, 6(3), 1998, pp. 226-239.

Chou, F. C. and C. Y. Tseng, "Corpus-based Mandarin speech synthesis with contextual syllabic units based on phonetic properties," *International Conference on Acoustics, Speech, and Signal Processing*, 1998, pp.893-896.

Charpentier, F. J. and M. G. Stella, "Diphone synthesis using an overlap-add technique for speech waveforms concatenation," *International Conference on Acoustics, Speech, and Signal Processing*, 1986, pp. 2015-2020.

Rabiner, L. R., M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A Comparative performance study of several pitch detection algorithms," *IEEE Transactions on Acoustics., Speech and Signal Processing*, 24, 1976, pp. 399-417.

Rabiner, L. R., "On the use of autocorrelation analysis for pitch detection," *IEEE Transactions on Acoustics., Speech and Signal Processing*, 25, 1977, pp. 24-33.

Noll, A. M., "Cepstrum pitch determination," *Journal of the Acoustical Society of America*, 47, 1967, pp. 293-309.

Markel, J. D., "The SIFT algorithm for fundamental frequency estimation," *IEEE Transactions on Audio Electroacoustics*, Au-20, 1972, pp. 367-377.

Barnard, E., R. A. Cole, M. P. Vea, and F. A. Alleva, "Pitch detection with a neural-net classifier," *IEEE Transactions on Signal Processing*, 39(2), 1991, pp. 298-307.

Kadambe, S., G. F. Boudreaux-Bartels, "A comparison of a wavelet functions for pitch detection of speech signals," *International Conference on Acoustics, Speech, and Signal Processing*, 1991, pp. 449-452.

Barner, K. E., "Colored L-l filters and their application in speech pitch detection," *IEEE Transactions on Signal Processing*, 48(9), 2000, pp. 2601-2606.

Huang, H. and F. Seide, "Pitch tracking and tone features for Mandarin speech recognition," *International Conference on Acoustics, Speech, and Signal Processing,* 2000, pp.1523-1526.

Moulines, E., F. Emerard, D. Larreur, J. L. Le Saint Milon, L. Le Faucheur, F. Marty, F. Charpentier, and C. Sorin, “A real-time French text-to-speech system generating high-quality synthetic speech,” *International Conference on Acoustics, Speech, and Signal Processing*, 1990, pp.309-312.

Kobayashi, M., M. Sakamoto, T. Saito, Y. Hashimoto, M. Nishimura, and K. Suzuki, “Wavelet analysis used in text-to-speech synthesis,” *IEEE Transactions on Circuits and Systems-II, Analog and Digital Signal Processing*, 45(8), 1998, pp. 1125-1129.