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

Jau-Hung Chen and Yung-An Kao

Advanced Technology Center, Computer and Communication Research Laboratories,

Industrial Technology Research Institute, Chutung 310, Taiwan

Email: chenjh@itri.org.tw, kya@itri.org.tw

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 an optimal quality of the synthetic speech. In general, there are two major issues on pitch marking: pitch detection and location determination. In this paper, an adaptable filter, which serves as a bandpass filter, is proposed for pitch detection to transform the voiced speech into a sine-like wave. Based on the sine-like wave, a peak-valley decision method is investigated to determine the appropriate part (positive part and negative part) of the voiced speech for pitch mark estimation. At each pitch period, two possible peaks/valleys are searched and the dynamic programming is performed to obtain the pitch marks. Experimental results indicate that our proposed method performed very well if correct pitch information is estimated.

1. Introduction

In past years, the approach of concatenative synthesis has been adopted by many text-to-speech (TTS) systems [1]–[6]. The concatenative synthesis uses real recorded speech segments as the synthesis units and concatenates them together during synthesis. Also, the time-domain pitch-synchronous overlap-add (TD-PSOLA) [6] method has been employed to perform prosody modification. This method modifies the prosodic features of the synthesis unit according to the target prosodic information. Generally, the prosodic information of the speech includes pitch (the fundamental frequency), intensity, and duration, etc. For a synthesis scheme based on TD-PSOLA method, it is necessary to obtain a pitch mark for each pitch period in order to assure an optimal quality of the synthetic speech. The pitch mark is a reference point for the overlap of the speech signals.

It is useful to have a speech synthesizer with various voices 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 the speaker of its favorite. For conventional TTS systems, however, it is a professional but tedious job to create a new voice. Recently, corpus-based TTS systems have been appreciated which use a large amount of speech segments. Some approaches selected the speech segments as the candidates of synthesis units. Establishing the synthesis units includes speech segmentation, pitch estimation, pitch marking, and so on. However, pitch marking is very labor-intensive among them if there involved no automatic mechanism.

In general, there are two major issues on pitch marking: pitch detection and location determination. Compared to pitch detection [7]-[14], few papers have been presented for pitch marking [15][16], which is also a difficult problem because of the great variability of the speech signals. Moulines *et al.* [15] proposed a pitch-marking algorithm based on the detection of abrupt changes at glottal closure instants. At 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.* [16] used dyadic wavelet 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 is shown in Fig. 1. The input signals are constrained to the voiced speech because only the periodic parts are interested. We introduce an adaptable filter, which serves as a bandpass filter, to transform the voiced speech into a sine-like wave. The autocorrelation method is then used to estimate the pitch periods on the sine-like wave. Also, a peak-valley decision method is presented 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 motivated from Fig. 2(a), which displays an example of waveform having the negative part reveals explicit periodicity. In general, it could synthesize better speech quality if the pitch marks are labeled at the positions of extreme points (peaks and valleys) of the speech. At each pitch period, two possible peaks/valleys are searched. Finally, the pitch marks are obtained by the dynamic programming by calculating the pitch distortion.

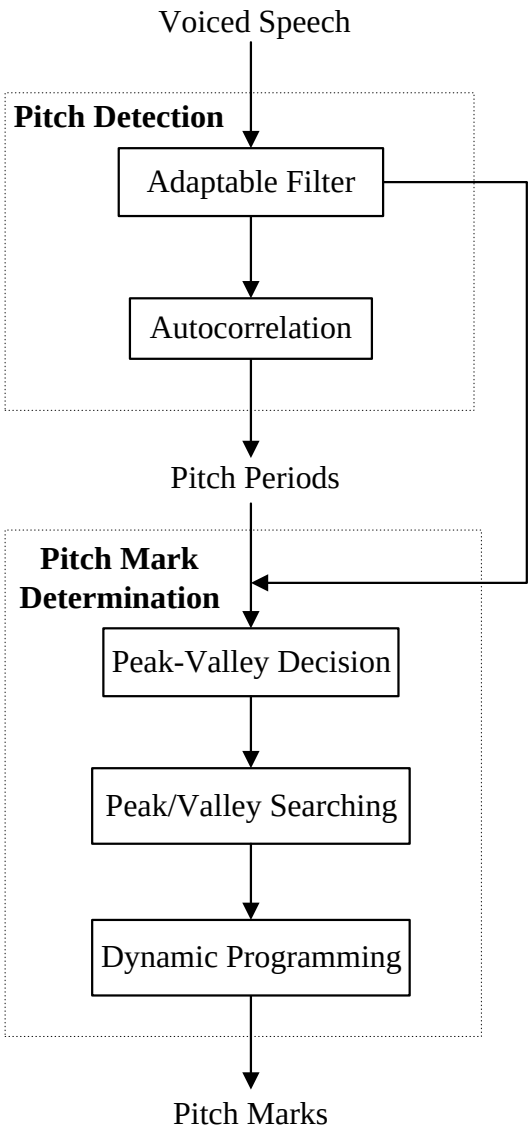

Figure 1: Block diagram of the proposed pitch-marking method.

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

The proposed adaptable filter serves as a bandpass filter in which its pass band is from 50 Hz to the detected fundamental frequency, up to 500 Hz, of the voiced speech. The adaptable filter is achieved by the following three steps.

Step 1. It computes the FFT (Fast Fourier Transform) to transform the voiced speech into the frequency domain.

Step 2. The fundamental frequency, f_0, is detected by searching the first peak of the spectral contour.

Step 3. The IFFT (Inverse FFT) is invoked over the passband between 50 Hz and f_0 to obtain the filtered speech.

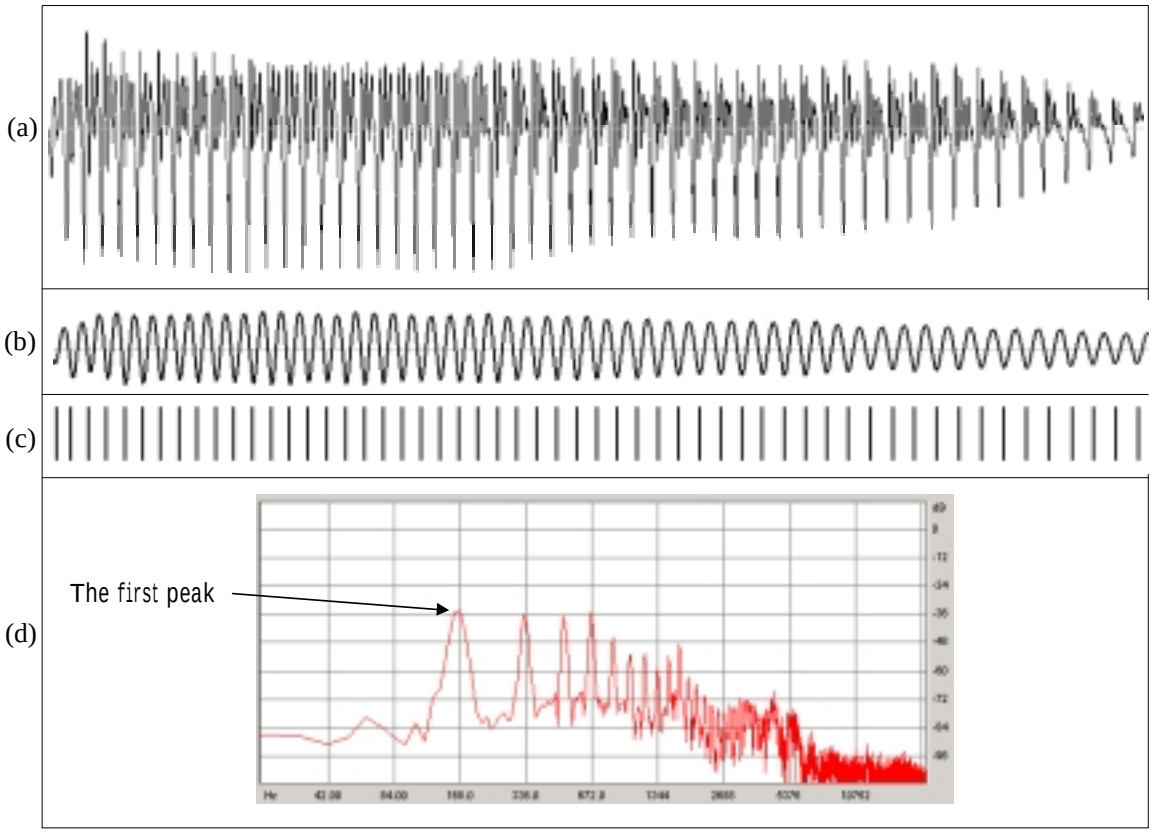

Figure 2: Results of the adaptable filter and pitch mark determination. (a) Waveform of the voiced speech with explicit periodicity on the negative part. (b) Waveform of the filtered speech. (c) Detected pitch marks. (d) Spectral contour (note that the frequency axis is not linearly plotted).

An example of the adaptable filter is displayed in Fig. 2. Panel (a) and (b) shows 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 that reveals clear periodicity than that on 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 for the reason of inspecting the first spectral peak. The first peak was found at 168 Hz, which is the fundamental frequency. Finally, the pitch periods are obtained by analyzing the filtered speech using the conventional autocorrelation method.

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

3-1 Peak-Valley Decision

From observations, we found that the voiced speech, $s[\cdot]$, is synchronous with the filtered speech, $o[\cdot]$, either at peaks or at valleys. For the case illustrated in Fig. 2 (a) and 2 (b), they are synchronous at valleys having explicit periodicity instead of those at peaks. As a result, the pitch marks could be easily determined at the negative part than those at the positive part. In the following, peak-valley decision method calculates two costs by summing the amplitudes of $s[m]$, where m represents the position of the local extreme point of $o[\cdot]$ 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 $o[\cdot]$.

C_{valley} : Cost estimated at the valleys of $o[\cdot]$.

N_{peak} : Total number of the peaks of $o[\cdot]$.

N_{valley} : Total number of the valleys of $o[\cdot]$.

$Pos_{peak}[n]$: Position of the *n-th* peak of $o[\cdot]$.

$Pos_{valley}[n]$: Position of the *n-th* valley of $o[\cdot]$.

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

3-2 Pitch mark determination Based on Dynamic Programming

Once the adoption of the peak or valley has been decided, say peak, the positions of pitch marks are determined by picking the peaks of $s[\cdot]$. For the *i-th* pitch period, P_i, two highest peaks in the corresponding voiced speech are searched. 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. For this case, we let $L_{i2} = L_{i1}$. For all the detected peaks, the determination of pitch mark is then performed based on dynamic programming. The distortion of pitch period, $d_i(j,k)$, and its accumulation, $A_i(j)$, are defined as follows:

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

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

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 as

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

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

The search path of the dynamic programming is illustrated in Fig. 3. The peak locations (pitch marks) can be obtained by back tracing the peak sequence corresponding to the smallest value of $A_i(1)$ and $A_i(2)$. An example of the results of pitch marking is shown in Fig. 2(c). Similar procedures described above can be applied to the case of "valley".

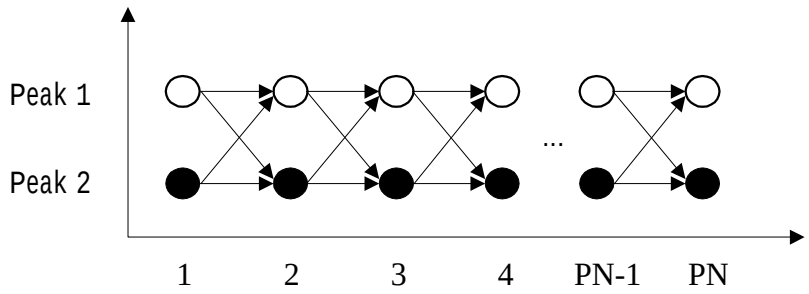

Figure 3: Illustration of the peak-picking search path of the dynamic programming.

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 between 4 to 6 Chinese characters. It includes an amount 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. The syllable segmentation was manually done in order to obtain the precise boundaries of voiced speech and unvoiced speech. The total duration of the 436 voiced speech is 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).

For the voiced speech, the waveforms along with the pitch marks obtained from our

pitch-marking program were visually displayed. The pitch marks were then checked and corrected by an experienced person through a friendly interface. For the evaluation of the experiments, we obtained 436 sets of human-labeled pitch marks, denoted as *H*, which comprises 23868 pitch marks.

4-2 Performance of the pitch marking method

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

The proposed method generated 23860 pitch marks, denoted as *I*, without any duplication. The success rate of the pitch marking method is defined 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% is obtained (baseline), in contrast with the 95% and 97% success rates of the methods of [15] and [16], respectively. However, we found that most of the errors are resulted from the incorrect results of pitch detection. Most of the pitch errors are due to large changes of pitch locating at the boundaries of the voiced speech. Providing correct pitch information, our method leads to a success rate of 99.5%.

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

Condition	Baseline	Using correct pitch
Success rate	97.2%	99.5%

5 Conclusions

In this paper, a preliminary work on pitch marking has been proposed. We present the adaptable filter combined with the autocorrelation method for pitch detection. On the other hand, a peak-valley decision method is introduced to select either the positive or the negative parts for evaluation of pitch mark. Also, a dynamic-programming-based pitch mark determination method is demonstrated where two peaks/valleys are searched at each period. In the experiments, our pitch-marking method achieves 97.2% success rate.

Furthermore, a high success rate of 99.5% is obtained providing correct pitch information.

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

1. Hamon, C., E. Moulines, and F. Charpentier, “A diphone synthesis based on time-domain prosodic modifications of speech,” in Proc ICASSP, 1989, pp.238-241.
2. Iwahashi, N. and Y. Sagisaka, “Speech segment network approach for optimization of synthesis unit set,” Computer Speech and Language, 1995, pp.335-352.
3. Shih, C. L. and R. Sproat, “Issues in text-to-speech conversion for Mandarin,” in Computational Linguistics and Chinese Language Processing, vol.1, 1996, pp.37-86.
4. Chen, S. H., S. H. Hwang and Y. R. Wang, “An RNN-based prosodic information Synthesizer for Mandarin text-to-speech,” IEEE Trans. on Speech and Audio Processing, Vol. 6, No. 3, 1998, pp. 226-239.
5. Chou, F. C. and C. Y. Tseng, “Corpus-based Mandarin speech synthesis with contextual syllabic units based on phonetic properties,” in Proc. ICASSP, 1998, pp.893-896.
6. Charpentier, F. J. and M. G. Stella, “Diphone synthesis using an overlap-add technique for speech waveforms concatenation,” in Proc. ICASSP, 1986, pp. 2015-2020.
7. Rabiner, L. R., M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, “A Comparative performance study of several pitch detection algorithms,” IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-24, 1976, pp. 399-417.
8. Rabiner, L. R., “On the use of autocorrelation analysis for pitch detection,” IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-25, 1977, pp. 24-33.
9. Noll, A. M., “Cepstrum pitch determination,” J. Acoust. Soc. Amer., Vol. 47, 1967, pp. 293-309.
10. Markel, J. D., “The SIFT algorithm for fundamental frequency estimation,” IEEE Trans. Audio Electroacoust., Vol. Au-20, 1972, pp. 367-377.

11. Barnard, E., R. A. Cole, M. P. Vea, and F. A. Alleva, “Pitch detection with a neural-net classifier,” IEEE Trans. On Signal Processing, vol. 39, No. 2, 1991, pp. 298-307.
12. Kadambe, S., G. F. Boudreaux-Bartels, “A comparison of a wavelet functions for pitch detection of speech signals,” in Proc. ICASSP, 1991, pp.449-452.
13. Barner, K. E., “Colored L-l filters and their application in speech pitch detection,” IEEE Trans. On Signal Processing, Vol. 48, No. 9, 2000, pp. 2601-2606.
14. Huang, H. and F. Seide, “Pitch tracking and tone features for Mandarin speech recognition,” in Proc. ICASSP, 2000, pp.1523-1526.
15. 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,” in Proc. ICASSP, 1990, pp.309-312.
16. Kobayashi, M., M. Sakamoto, T. Saito, Y. Hashimoto, M. Nishimura, and K. Suzuki, “Wavelet analysis used in text-to-speech synthesis,” IEEE Trans. on Circuits and Systems-II, Analog and Digital Signal Processing, Vol. 45, No. 8, 1998, pp. 1125-1129.