# Time-Domain Pitch and Time Scale Modification of Speech Signal

NGUYEN TIEN DUNG and NGUYEN DINH HAI
School of Computing, National University of Singapore

Pitch and time scales modification of speech signal have many application: prosody modification in speech synthesis, voice morphing, singing voice correction, and speech watermarking. In this project, we implemented a complete pitch and time scale modification system using Time-Domain Pitch Synchronize Overlap and Add method. The system can produces synthetic speech with desire pitch scale and time scale modification for arbitrary input speech. The subjective evaluation result suggests that our system has equivalent quality to that of of Praat, a well-know speech analysis and synthesis program which is used widely in speech research community. The evaluation result also shows that low-band spectrum reconstruction method proposed by Mochizuki and Kobayashi [2004] work well with PSOLA and improve the quality of synthetic speech in case of downward pitch scale modification.

General Terms: Speech Modification, TD-PSOLA
Additional Key Words and Phrases: Pitch Estimation, Pitch Marking, Unit Waveform, Low-Band Spectrum Reconstruction, Voice Quality, Mean-Opinion-Score (MOS)

## 1. INTRODUCTION

Pitch and time scale modification of speech play an important role in many audio applications such as speech synthesis, voice morphing, singing voice correcting, and speech watermarking. There are two main categories of pitch and time scale modification of audio signals. They are time-domain techniques (PSOLA) and frequency-domain techniques (Phase Vocoder, Sinusoid plus noise model). The first one is very suitable for single pitch signals such as speech but is not adequate for multipitch signals, and the second one can be applied for both single pitch and multipitch signals such as music.

In this project, a time-domain technique called Time Domain Pitch Synchronize Overlap and Added (TD-PSOLA) is chosen because it is an effective and computational saving technique but still provides high synthetic voice quality. To implement PSOLA, firstly pitch marks have to be placed at each pitch cycles based on pitch estimation information. After that, unit waveforms are extracted from original speech signal using Hanning window with length of two pitch periods produced by three consecutive pitch marks. Based on pitch scale and time scale information, new synthetic pitch marks are calculated. Unit waveforms are placed at new synthetic pitch marks to produce output signal with desired pitch and duration.

The pitch and time scale modification system has three major parts: Pitch Estimation, Pitch Marking and TD-PSOLA illustrated in figure 1.

Because unit waveforms are modified in time domain, the frequency domain parameters are distorted. That leads to the perceptual distortion of modified speech. Chen et al.

[2006] showed that in TD-PSOLA pitch scale downward modification is the main factor of speech quality degradation. Time scale modification and pitch scale upward modification only have some affects. To overcome that restriction, a spectrum reconstruction method proposed by Mochizuki and Kobayashi [2004] was applied to reconstructs the low-band F0 harmonic components when pitch scale is modified downward.

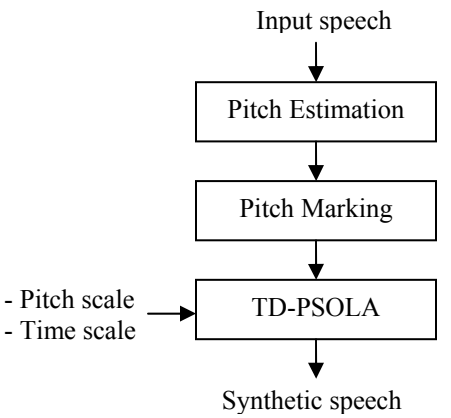


Fig. 1. System Overview

The remaining of this report is organized as follows. In section 2, we present an implementation of pitch estimation based on autocorrelation function. Pitch marking is described in section 3. TD-PSOLA and low-band spectrum reconstruction are described in section 4 and 5. Section 3, 4 and 5 are summary of Mattheyses [2006], Lin et al. [2004], Huang [2001] and Mochizuki [2004] added practical information from our implementations. Section 6 describe perceptual experiments used to measure the quality of whole system with comparison to TD-PSOLA based speech modification of Praat [Broersma 2005], one of the most popular tools used in speech research community. Finally, evaluation results conclusion is presented in section 7 and 8.

## 2. PITCH ESTIMATION

There are many pitch estimation methods for speech signal. Most of them belong to one of two major approaches: time-domain and frequency-domain. Many popular and reliable methods used autocorrelation function (ACF) measurement in time-domain [Broersma 1993]. YIN, the state-of-the-art pitch estimation method, also used ACF as first step of its implementation [Cheveigné 2002].

The ACF function could be defined as:

$$r(\tau) = \sum_{i=1}^{W-\tau} x_i x_{i+\tau}, \qquad (1)$$

with W is the size of rectangular window. Figure 2 shows a sample of a speech signal and it corresponding ACF values. If the input signal is periodic, the ACF is periodic too and show peaks at multiple of pitch period.

ACF is quite interesting. It is defined in time-domain but could be computed efficiently in frequency-domain via fast Fourier transform (FFT) and its inverse transform:

$$r(n) = iFFT\{|\ FFT[x(k)]\ |^2\}, \tag{2}$$

ACF based pitch estimation is closely related to cepstrum-based pitch estimation which replace |.|2 in (2) by log(|.|). Because cepstrum flattens the vocal track spectral shape, it is robust for formants but sensitive to noise. In contract, ACF emphasizes spectrum peaks so it is robust for noise but sensitive to formants. However, this weakness could be overcome by "spectral whitening" methods, center-clipping is one of them [Sondhi 1968].

Center-clipping is defined as

$$C[x(n)] = \begin{cases} x(n) - C_L, & x(n) > C_L \\ 0, & |x(n)| \le C_L \\ x(n) + C_L & x(n) < -C_L \end{cases}, \tag{3}$$

where $C_L$ is clipping threshold and generally is set to 30% of the maximum magnitude of the signal.

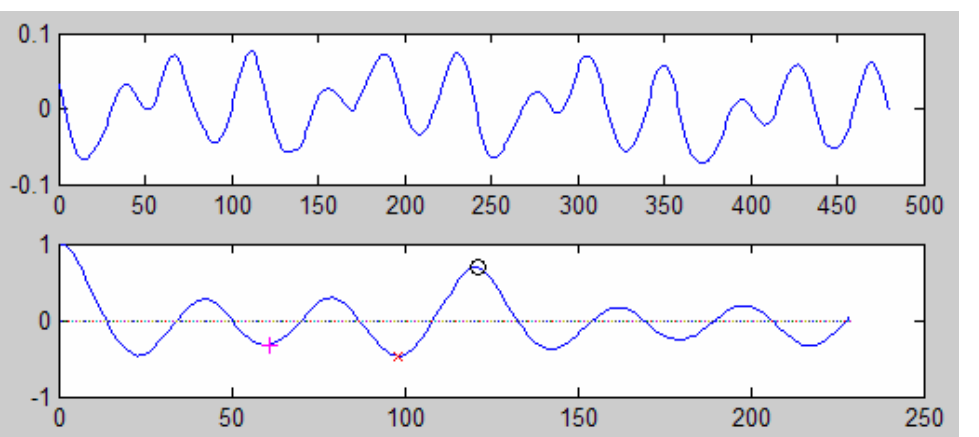


Fig. 2. Speech signal and it corresponding ACF

ACF is a smooth transition between time-domain and frequency-domain methods. It is very important to understanding both single pitch and multi-pitch estimation methods because people often borrow ACF to explain other periodic measurement functions and "auditory models based on autocorrelation are currently one of the more popular ways to explain pitch perception" [Cheveigné 2002].

In this project, ACF is chosen for pitch estimation implementation. The implementation steps are shown in figure 3. Because speech signal has pitch range from 60 Hz to 500Hz, low pass filter is used to cut off the frequency above 900Hz [Rabiner 1976]. This pre-processing step reduces noise and hence improves accuracy. Filtered signal is divided into frames with 30ms frame length and 10ms frame rate. Center clipping is applied to each frame before calculate ACF. Frame pitch is estimated based on ACF of each frame. Voiced / unvoiced decision also bases on ACF. After that, median filter, a post-processing method, is applied to reduce errors of previous steps.

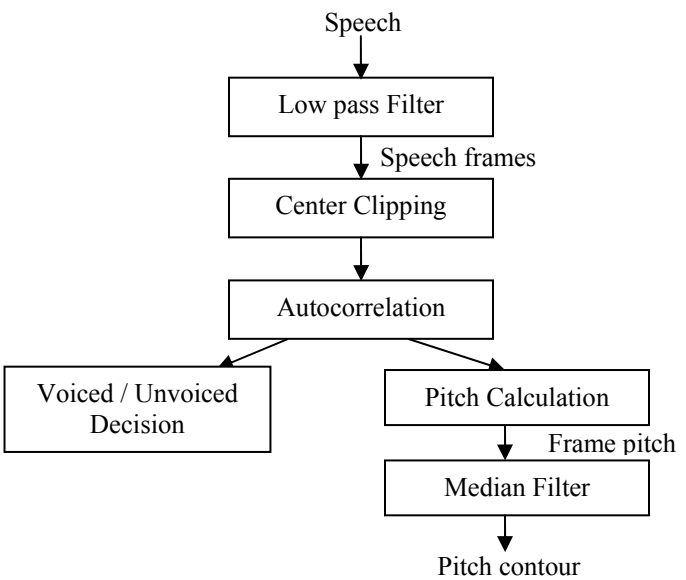


Fig. 3. Pitch Estimation

## 3. PITCH MARKING

The purpose of pitch marking is determine the where is the beginning and ending of a pitch cycle when the signal repeats itself using pitch contour estimated in pervious section. In figure 4, the vertical lines are pitch marks. The horizon line is the pitch contour. The * shapes are pitch mark candidates of search regions.

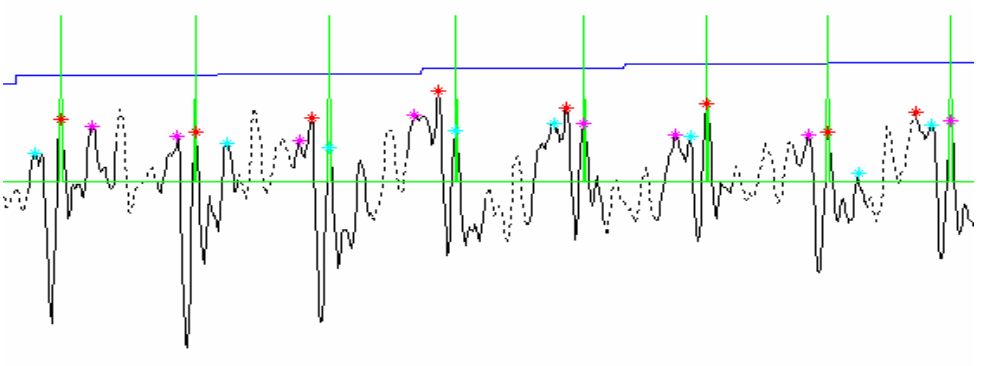

Fig. 4. Pitch marking illustration

Pitch marking is implemented based on Mattheyses [2006] and Lin et al. [2004]. Figure 5 shows steps of pitch marking algorithm. For each voiced speech segment, the global maximum location $t_m$ is the first pitch mark. The first search region is defined using pitch period value at $t_m$:

$$SR = [t_m + f.T_0; t_m + (2 - f).T_0], \quad (4)$$

The next search region is formed by applying formula (4) to $t_m$ chosen at maximum sample of current search region. The f value is usually set to 0.7.

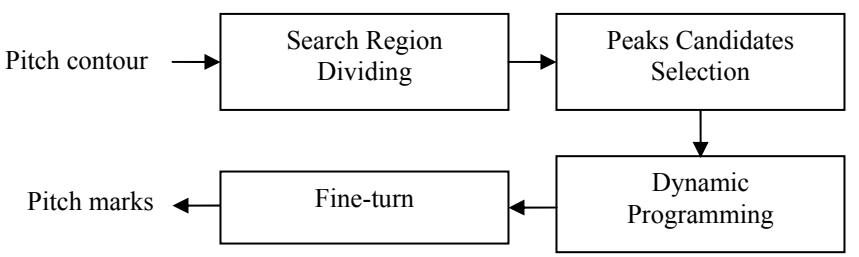


Fig. 5. Pitch Marking

After all search regions in voiced segment are known, three peak candidates for pitch marks are selected from each search regions. For each search region, only one peak candidate is chose as pitch mark. A minimum distance between two consecutive candidates is defined to assure that there is no too close peak candidate pair.

Best pitch mark sequence is found by dynamic programming which maximize probabilistic optimally criterion. There are two probabilities. The first one is relative to peak candidate height, which prefers higher amplitude peaks (5). h(j) is amplitude of j, $h_{min}$ and $h_{max}$ are minimum and maximum amplitude of search region containing j. The second one relates to distance between two pitch mark candidates in two consecutive search regions (6), $T_0$ is pitch period of frame containing i and j; $d_{ij}$ is distance between i and j; α, β and γ are three adjustment parameters which are adapted for each speaker.

$$s(j) = (\frac{h(j) - h_{\min}}{h_{\max -} h_{\min}})^{\alpha}, \tag{5}$$

$$t(i,j) = (\frac{1}{1 + \beta \mid T_0 - d_{ij} \mid})^{\gamma}, \tag{6}$$

Dynamic programming is applied to get the optimal pitch mark sequence with k is current search region that maximize likelihood

$$P(k,j) = \max[P(k-1,i) + \log t_k(i,j)] + \log s_k(j), \tag{7}$$

## 4. TD-PSOLA

In this project, we implement PSOLA method to do pitch and time modification. The final synthesis wave should have the same spectral characteristics as input wave but with a different pitch and/or duration. The general implementation steps are shown in figure 6.

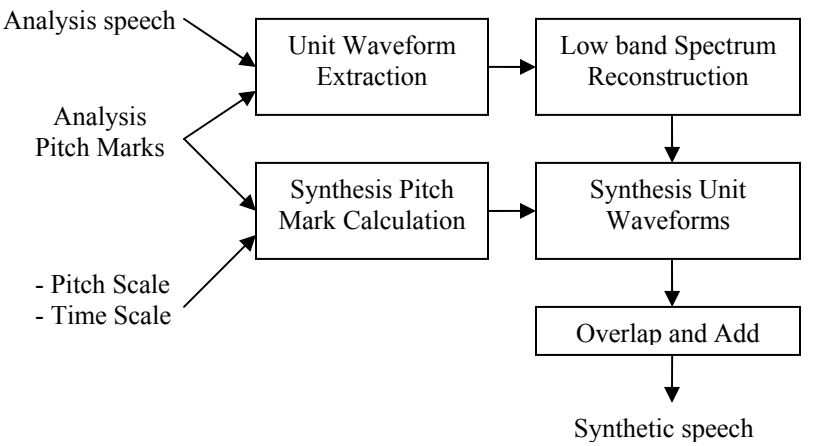


Fig. 6. TD-PSOLA implementation

In order to have better synthesis speech, we classify speech into 2 categories: voiced speech and unvoiced speech. Voiced speech is the speech that has high local energy or low local energy and low amount of zero-crossing. Other speeches which are not voiced speeches are considered as unvoiced speeches.

There are 3 steps in our PSOLA implementation. First, calculate synthesis epoch from analysis epoch. Second, use Hanning window to mapping analysis signal to synthesis signal. And lastly, apply waveform mapping using linear interpolation to have better smooth speech.

### 4.1 Find epoch of synthesis wave

The accurateness of epoch sequence is very important to achieve a high quality prosody modification. Because input signal is voiced signal, so it can be presented as a function of pitch cycles $x_i[n]$:

$$x[n] = \sum_{i=-\infty}^{\infty} x_i[n - t_a[i]] \tag{8}$$

Where $t_a[i]$ are the epochs of the analysis signal and the pitch cycle is windowed version of the input:

$$x_i[n] = w_i[n]x[n] \tag{9}$$

The synthesis epochs are computed so as to meet a specified duration and pitch constraints. In our implementation, we use formula 16.25 [Huang et al. 2001]:

$$t_s[j+1] - t_s[j] = \frac{\alpha}{t_s[j+1] - t_s[j]} \int_{t_s[j]/\alpha}^{t_s[j+1]/\alpha} \beta(t) P_a(t) dt \tag{10}$$

Where $t_s$ is synthesis epoch, $\alpha$ is time scale factor and $\beta$ is pitch scale factor. We assumed that $\alpha$ and $\beta$ are not changed during the speech.

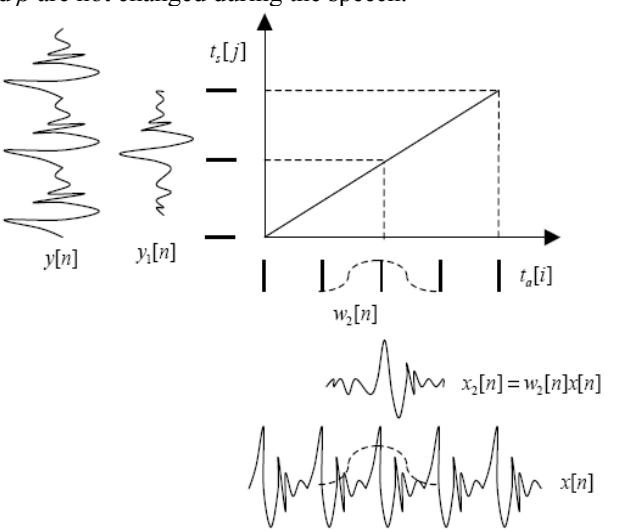


Fig. 7: Mapping between five analysis epochs $t_a[i]$ and three synthesis epochs $t_s[j]$ Duration has been shortened by 40% and pitch period increased by 60%. Pitch cycle $x_2[n]$ is the product of the analysis window $w_2[n]$, in dotted line, with the analysis signal $x[n]$, which is aligned with analysis epochs $t_a[i]$. In this case, synthesis pitch cycle $y_1[n]$ equals $x_2[n]$ and also $y_0[n] = x_0[n]$ and $y_2[n] = x_5[n]$. Pitch is constant over time in this case [Huang et al. 2001]

Figure 7 shows mapping from input signal to output signal using Hanning window, which spans two pitch periods [Huang et al. 2001].

### 4.2 For voiced speech

Replace the analysis epoch sequence $t_a[i]$ with the synthesis epochs $t_s[j]$ and the analysis pitch cycles $x_i[n]$ with the synthesis pitch cycles $y_j[n]$:

$$y[n] = \sum_{j=-\infty}^{\infty} y_j[n - t_s[j]] \tag{11}$$

The synthesis epochs are already calculated from *step 1*.

### 4.3 For unvoiced speech

We equally divide unvoiced speech period into segments of length 10ms. After that, we applied Hanning windows to 2 periods.

### 4.4 Waveform mapping using linear interpolation

In order to increase smoothness of synthesis waveforms, we applied linear interpolation method to voiced frames.

Suppose that $t_a[i] \le t_a^{'}[j] < t_a[i+1]$

Then $y_j[n]$ is given by

$$y_j[n] = (1-\gamma_j)x_i[n] + \gamma_j x_{i+1}[n] \quad (12)$$

Where $$\gamma_j = \frac{t_a^{'}[j] - t_a[i]}{t_a[i+1] - t_a[i]} \quad (13)$$

Although the linear waveform mapping process increase smoothness of voiced frames, it's not suitable to apply to unvoiced frames. For unvoiced framed, this interpolation results in a decrease of the amount of aspiration.

## 5. LOW-BAND SPECTRUM RECONSTRUCTION

Because unit waveforms are modified in time domain, some of its frequency domain parameters are distorted. That leads to the perceptual distortion of modified speech. To overcome that restriction, a spectrum reconstruction method proposed by Mochizuki and Kobayashi [2004] was applied to reconstructs the low-band $F_0$ harmonic components when pitch scale is modified downward.

The steps of spectrum reconstruction are showed in figure 8.

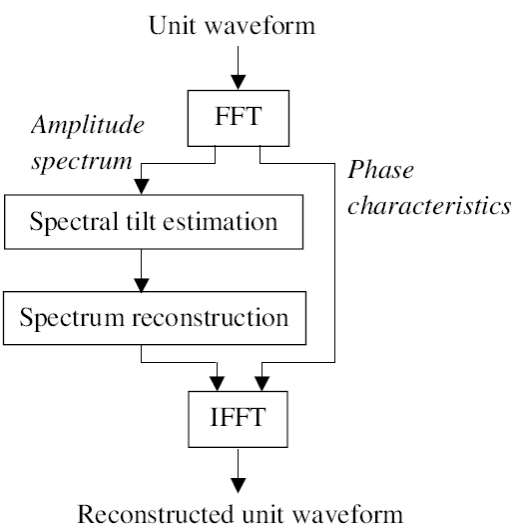


Fig. 8. Spectrum reconstruction produces [Mochizuki et all. 2004]

First of all, the spectrums of unit waveforms are calculated after zero-padding using FFT. The length of each unit waveform is double of its T0 so the zero-padding is necessary to both increase the spectrum resolution and retain characteristic of frequencies lower than

$F_0$ when doing IFFT. Then spectral tilt is estimated. Based on spectral tilt, and new target $F_0$, the spectrum is reconstructed. Finally, unit waveforms are reconstructed using IFFT. The following parts present more detail of these steps. They are summarized and captured from Mochizuki et all. [2004].

### 5.1 Spectral tilt estimation

The spectral tilt α (figure 9) is calculated using Least Mean Square as following:

$$\alpha = \frac{\sum_{\omega_i \in \Omega} \log_2 \frac{\omega_i}{\overline{\omega}} \ln \frac{|S(\omega_i)|}{\overline{|S(\omega)|}}}{\sum_{\omega_i \in \Omega} \left( \log_2 \frac{\omega_i}{\overline{\omega}} \right)^2} \tag{14}$$

Where $\Omega = \{\omega_i \mid \omega_0 < \omega_i < \Pi\}$, $\omega_0$ is angular frequency corresponding to the original $F_0$, $\overline{S(\omega)}$ is the mean spectral amplitude. Only spectrum envelope between $F_0$ and $F_s$ / 2 are used.

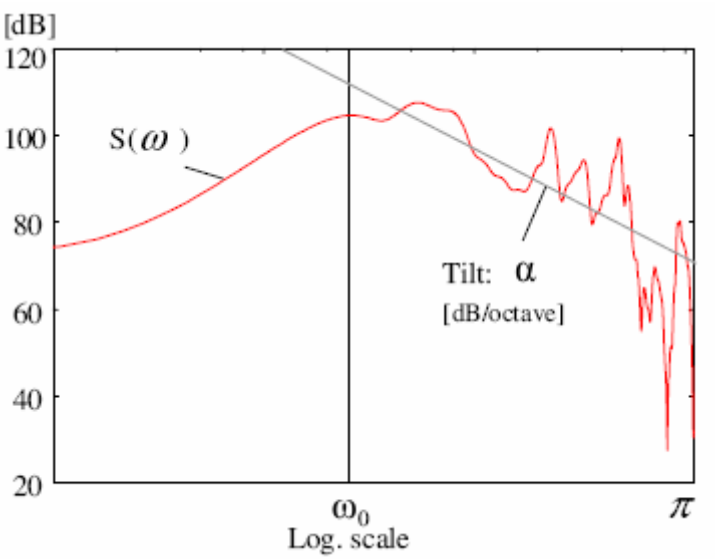


Fig. 9. Estimation of spectral tilt [Mochizuki 2004]

### 5.2 Spectrum reconstruction

The spectrum envelope $S'(\omega)$ is reconstructed by convoluting line spectra with the frequency characteristics of Hanning window and is defined as follows:

$$S'(\omega) = \begin{cases} \sum_{i=1}^{N} A_i \frac{|W(\omega_i)|}{W\max_i} & (\omega < \omega'_0 N) \\ S(\omega) & (\omega'_0 N \le \omega) \end{cases} \quad (15)$$

$$W\max_i = \max|W_i(\omega)|, \qquad N = \left\lceil \frac{\omega_0}{\omega'_0} \right\rceil + 1$$

Where $\omega'_0$ is the angular frequency corresponding to the target $F_0$, i is the harmonic number of $\omega'_0$. $A_i$ is the amplitude of the i-th target line spectrum. $W_i(\omega)$ is the frequency characteristics of window function (Hanning)

$$A_i = \begin{cases} \exp\{\alpha \log_2(i.\omega'_0/\omega_0)\}.S(\omega_0) & (i < \omega_0/\omega'_0) \\ S(i.\omega'_0) & (\omega_0/\omega'_0 \le i) \end{cases} \quad (16)$$

$$W_i(\omega) = F[w_i(t)] \quad (17)$$

$$w_i(t) = w_{han}(t,\tau).\cos(2\Pi i t/T_0) \quad (18)$$

$$w_{han}(t,\tau) = 0.5\{1.0 + \cos(\Pi t/\tau)\} \qquad (|t| < \tau) \quad (19)$$

Where $T_0$ is pitch period, F[.] denotes FFT and $w_{han}(t,\tau)$ is the Hanning window.

## 6. PERCEPTUAL EXPERIMENTS

To evaluate the quality of our implementation and verify the effectiveness of low-band spectrum reconstruction method proposed by Mochizuki and Kobayashi [2004], two perceptual experiments tests were performed.

The first one is Mean-Opinion-Score (MOS) test that measure the MOS score of synthetic utterances produced by our program and Praat [Broersma 2005]. The same input utterances and input parameters are used for both our program and Praat that produces two synthetic utterances at of the same pitch and time scales. The input utterances were selected from TIMIT database. There are four five utterances, spoken by two females and two males, were selected. There are eight synthetic utterances used in MOS experiment, four produced by our program, four produced by Praat. Six listeners were listened eight stimuli in a random order. After the stimulus was finish, the listener was asked to measure the naturalness quality of the stimulus from 0 to 5 (0 is bad and 5 is good). The results were used to estimate the quality of our program and Praat.

The second experiment was carried out to evaluate the quality improvement of low-band spectrum reconstruction method proposed by Mochizuki and Kobayashi [2004].

One female utterance was selected to produce eight synthetic utterances, four unreconstructed and four reconstructed. The same six listeners of the first experiment were listened four stimuli. For each stimulus, one unreconstructed synthetic utterances and the corresponding reconstructed utterance were play in random order. The listener was asked to choose which utterance he preferred.

For second experiment, only female utterance and downward pitch scale modification were interested because downward pitch scale modification is the main factor of speech quality degradation, upward pitch scale modification only has some effect [Chen et al. 2006]. Figure 10 illustrated above conclusion by showing that the low-band spectrum after reconstructing is much more different from unreconstructed one in downward modification than upward modification. Mochizuki and Kobayashi [2004] also implicit stated that the perceptual different between reconstructed utterance and the corresponding unreconstructed one is only significant for female voice. It could due to the average F0 of female voice is much higher than average F0 of male voice.

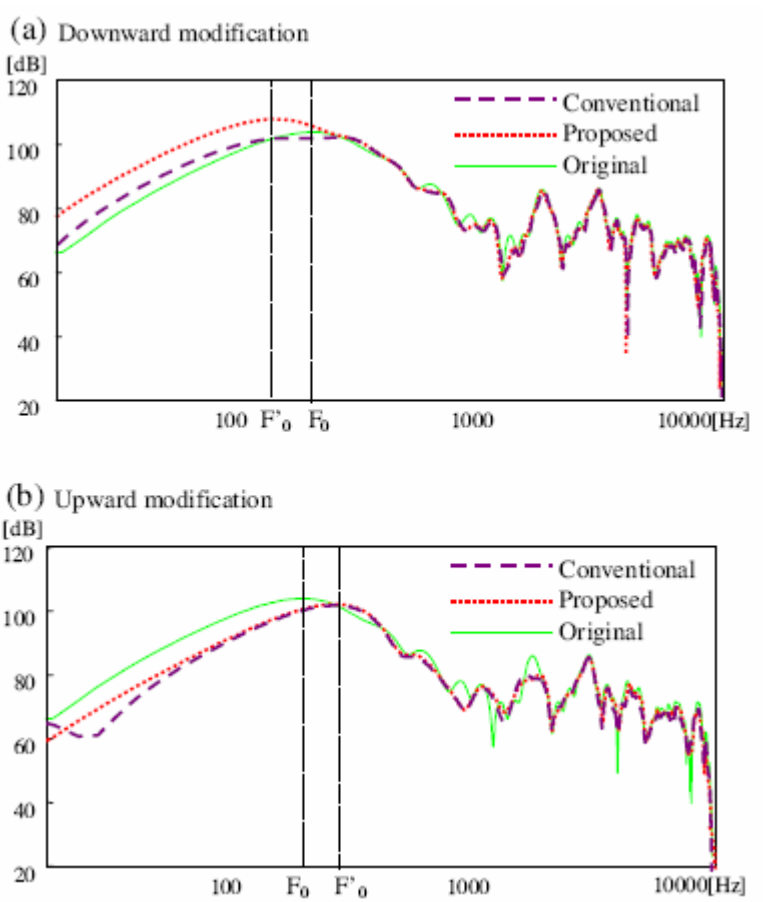


Fig. 10. Spectrum envelope extracted from $F_0$ modified speech [Mochizuki 2004]

## 7. RESULTS

The result of quality evaluation is shown in figure 11. The MOS of Praat is 3.54 and the MOS of our implementation is 3.29 (the absolute score is 5). The results show that the quality of our implementation is equivalent to Praat.

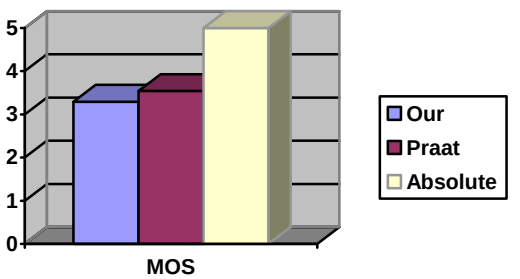


Fig. 11. Mean opinion score

Figure 12 shows the result of quality comparison between unreconstructed (normal) PSOLA and low-band spectrum reconstructed PSOLA. In most cases, the reconstructed is more preferred. It means that the spectrum reconstruction improve quality of synthetic utterance in case of downward pitch scale modification.

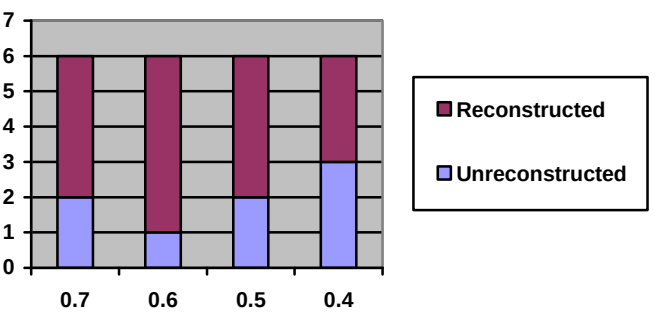


Fig. 12. Preferring test results

## 8. CONCLUSION

In this project, we implemented a complete pitch scale and time scale modification of speech signal. The system includes pitch estimation, pitch marking, PSOLA and low-band spectrum reconstruction. The subjective evaluation suggest that the quality of our system is equivalent to synthetic quality of Praat and low-band spectrum reconstruction method improves the quality in case of downward pitch scale modification.

## REFERENCES

BOERSMA, P. 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of Inst. of Phonetic Sciences 1(17):97,* 110.

BOERSMA, P AND WEENINK, D. 2005. Praat: doing phonetics by computer (Version 4.3.14) [Computer program]. Retrieved May 26, 2005, from http://www.praat.org/

CHEN, S.H, CHEN, S.J, KUO, C.C. 2006. Perceptual Distortion Analysis and Quality Estimation of Prosody-Modified Speech for TD-PSOLA, In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, May 2006

CHEVEIGNÉ, A.D, KAWAHARA. 2002. Yin: A fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America, 111(4)*, 2002

HUANG, X, ACERO, A, HON, H.W. 2001. *Spoken Language Processing - A Guide to Theory, Algorithm and System Development*, Prentice Hall PTR, 2001

LIN, C.Y, JANG, J.S. 2004. A two-phase pitch marking method for td-psola synthesis. In *Proceedings of 8th International Conference on Spoken Language Processing*, 2004

MATTHEYSES, W et al. Robust Pitch Marking For Prosodic Modification Of Speech Using TD-PSOLA, In *Proceedings of the 2006 The second annual IEEE BENELUX/DSP Valley Signal Processing Symposium*, 2006

MOCHIZUKI, KOBAYASHI. 2004. A Low-Band Spectrum Envelope Modeling for High Quality Pitch Modification. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing,* 2004

RABINER L.R. et al. 1976. A comparative performance study of several pitch detection algorithms. *IEEE Trans ASSP* 24, 399-418, 1976

SONDHI, M.M. 1968. New Methods of Pitch Extraction, *IEEE Trans. Audio and Electroacoustics*, 1968

SOOD, S, KRISHNAMURTHY, A. 2004. A robust on-the-fly pitch (OTFP) estimation algorithm. In *Proceedings of 12th annual ACM international conference on Multimedia*, 2004