# Pitch-Scale Modification Based on Formant Extraction from Resampled Speech

Pei-Chih Su (蘇培智) and Pao-Chi Chang (張寶基)

Department of Communication Engineering, National Central University
Chung-Li, Taiwan 320
pcchang@ce.ncu.edu.tw

## Abstract

The speech pitch-scale modification that can change the tone and the prosody is useful in the applications of privacy protection and entertainment. The analysis-synthesis method is one of the approaches for pitch-scale modification. It provides the freedom for synthesizing arbitrary voices once the speech parameters such as LPC coefficients and residual signals are obtained. In this paper we propose a pitch-scale modification method based on formant extraction from resampled speech. A dual-resampling mechanism is used to obtain the modified formants and modified pitch harmonics, respectively. In addition, the cross-correlation coefficients are calculated to locate the synchronization point, i.e., the pitch mark. Experimental results show that the speech can be successfully modified to different timbre and tone with high quality.

## 1. Introduction

Speech transformation changes the speech signals in two directions. The first is to change the speed of the speech signal in the time domain, called time scaling. It is the process of compressing or stretching the time basis of the speech signal without changing its spectral contents. The second is to change the tone of the speech signal, called pitch scaling. It modifies the spectrum of the signal without changing its playback time. These two speech transformation functions are important in many applications such as speech transmission and storage, audio-visual systems, speech recognition, and text to speech conversion. The time scaling modification is relatively easy that can be performed by TD-PSOLA technique with very good performance [1][2]. Therefore we focus on the pitch scaling speech transformation in this work.

Pitch-scale modification that can change the tone and the prosody of speech is useful in privacy protection and entertainment. One of the approaches for pitch-scale modification is the analysis-synthesis method [3]. It has the freedom for synthesizing arbitrary voice once the speech parameters such as LPC coefficients and residual signal are obtained.

A pitch-scale modification method based on formant extraction from resampled speech is proposed. The formants, which are the spectrum envelopes of speech signals, can be extracted by LPC analysis. This formant extraction procedure, so-called de-formant, eliminates the short-term correlation incurred by vocal tract filter. The frequency response of LPC synthesis filter determines the timbre of synthesized speech. The residual signals mainly consist of long-term components, the pitch harmonics, which determine the tone of speech and can be easily modified by using the resampling technique [4][5][6][7]. A dual-resampling mechanism is used to obtain the modified formants and modified pitch harmonics, respectively. The pitch-scale modification mentioned above is only performed in voiced frames because they have high energy and are relatively stable. In addition, the cross-correlation coefficients are calculated to locate the synchronization point, i.e., the pitch mark. Pitch synchronization will then be performed to produce the modified speech output.

The rest of this paper is divided into three sections. Section 2 describes the proposed method of pitch-scale modification based on resampling. Section 3 shows some experiment results. Finally, conclusions are made in section 4.

## 2. The proposed system

The proposed pitch-scale modification system architecture is shown in Figure 1. The first step is the voiced/unvoiced decision. The proposed pitch-scale modification is only performed in voiced frames because they have high energy and are relatively stable. For a voiced frame, the pitch period is estimated and if it is in the transition region, from either voiced to unvoiced or unvoiced to voiced, it will be divided into two sub-frames for a better time resolution. Both sub-frames will be decided as voiced or unvoiced with the same procedure, respectively. The core of the proposed system is the pitch modification and formant modification. The basic concept is shown in Figure 2, in which the

pitch and formant are separately modified based on the resampling method and finally combined.

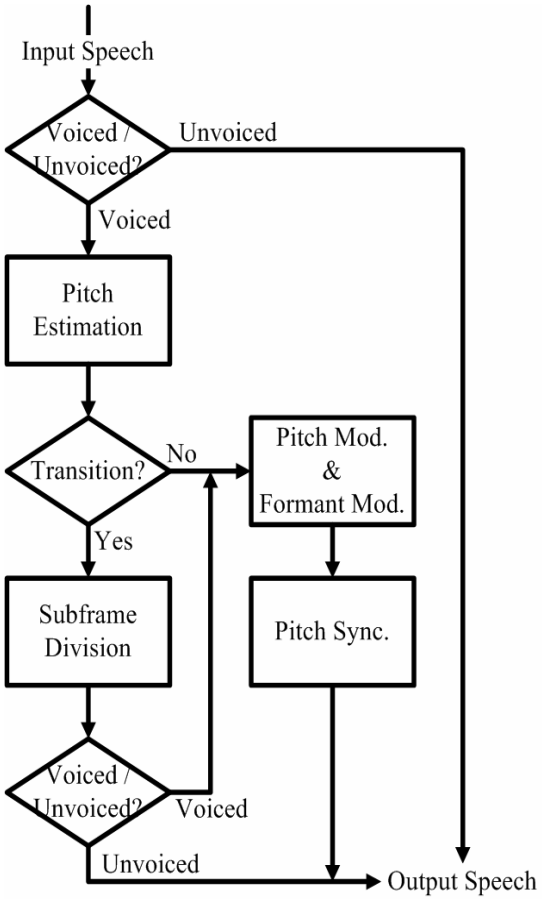


Figure 1: System architecture

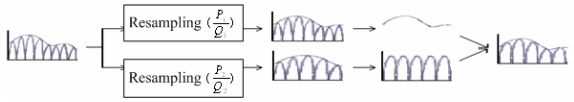


Figure 2: The basic concept of pitch modification and formant modification

A very important step is the pitch synchronization that solves the problem of out-of-synchronization between the current frame and the previous frame. Finally, the synthesized speech is obtained after the low-pass filter. The details of each major block are discussed as follows.

## 2.1 Voiced/Unvoiced decision

The speech signal is first classified as either voiced or unvoiced. Each frame is tested based on three parameters [8]: energy of signal, normalized energy of prediction error, and normalized autocorrelation with unit delay. Any parameter that is below the threshold will result in an unvoiced decision as shown in Figure 3.

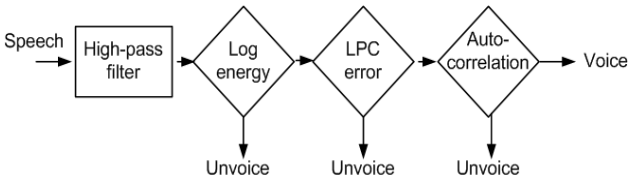


Figure 3: Voiced/Unvoiced decision

The speech samples of frame *s(n)* are filtered by high-pass filter and the following three parameters are calculated.

(a) Energy of signal *Es*:

$$Es = 10 \times \log_{10}\left( \varepsilon + \frac{1}{N}\sum_{n=1}^{N} s^2(n) \right) \quad (1)$$

where $N$ is the length of frame, $\varepsilon$ is a small positive constant.

(b) Normalized energy of prediction error $Ep'$:

$$Ep' = 10 \times \log_{10}\left( 10^{-6} + \left| \sum_{k=1}^{p} \alpha_k \phi(0,k) + \phi(0,0) \right| \right) \quad (2)$$

$$Ep = Es - Ep' \quad (3)$$

where $Ep$ is the energy of prediction signal, $\alpha_k$ is the predictor coefficient, and

$$\phi(i,k) = \frac{1}{N}\sum_{n=1}^{N} s(n-i)s(n-k) \quad (4)$$

is the *(i ,k)* term of covariance matrix of the speech samples.

(c) Normalized autocorrelation coefficient with unit delay $R_1$ :

$$R_1 = \frac{\sum_{n=1}^{N} s(n)s(n-1)}{\sqrt{\left(\sum_{n=1}^{N} s^2(n)\right)\left(\sum_{n=0}^{N-1} s^2(n)\right)}} \quad (5)$$

## 2.2 Pitch estimation

When estimating the pitch period, a method called circular average magnitude difference function (CAMDF) [9] is often used and it is defined as:

$$D(k) = \sum_{n=0}^{N-1} \left| s_w\left(\mathrm{mod}(n+k, N)\right) - s_w(n) \right| \quad (6)$$

Figure 4 shows an example of CAMDF result of a voiced speech segment. By detecting the local

minima of CAMDF, the candidates of pitch period are obtained.

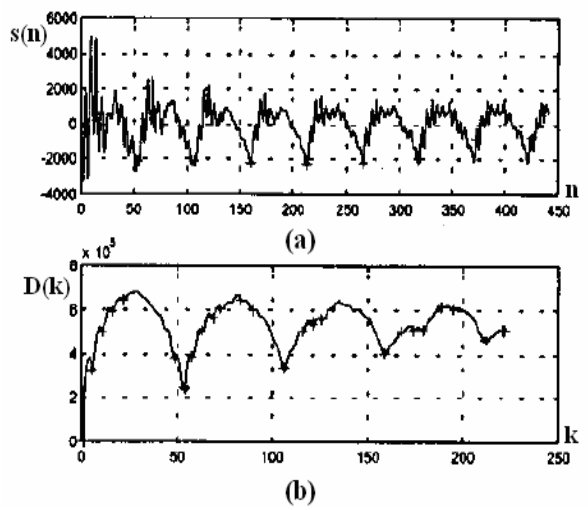


Figure 4: Example of CAMDF [8]. (a) voiced speech segment. (b) CAMDF result

## 2.3 Pitch modification and formant modification

Once the voiced speech is resampled, its formant structure and pitch harmonics in spectral domain will be changed at the same time. However, the variation of formants between natural voices is usually different from that of pitch harmonics. Thus a dual-resampling mechanism is used to perform pitch modification and formant modification, as shown in Figure 5.

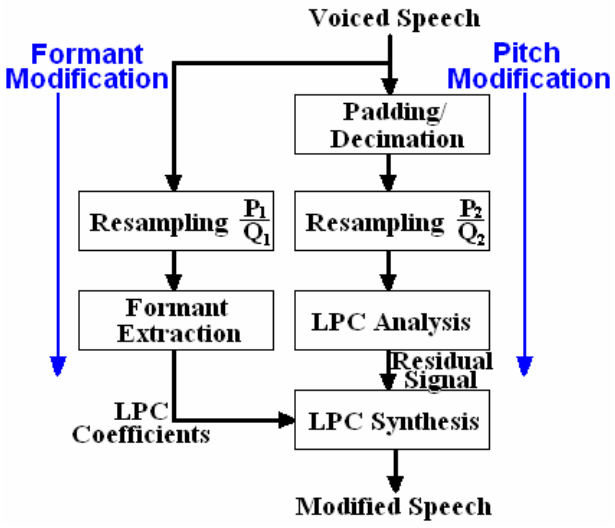


Figure 5: The block diagram of pitch modification and formant modification

By using different resampling ratios, the tone and timbre of speech can be changed independently.

## 2.4 Pitch synchronization

The pitch across two frames is often out of synchronization when the pitch period or the frame length is changed. Pitch synchronization in this work contains two steps: one is the window extraction; the other is the search of synchronization position. The cross-correlation function is first calculated as

$$R(\tau) = \sum_{n=1}^{N} X_p(n) \times X_c(n+\tau) \qquad (7)$$

where $X_p$ is the signal of the previous frame and $X_c$ is the signal of the current frame.

(a) Window extraction

The signal in the current window will be changed when performing the pitch-scale modification. It is important to find the high-correlation window achieve pitch synchronization. We compute the backward cross-correlation between the residual frame of the previous window and the current window to find the high-correlation window. The maximum length of shift is the pitch period. This procedure is shown in Figure 6.

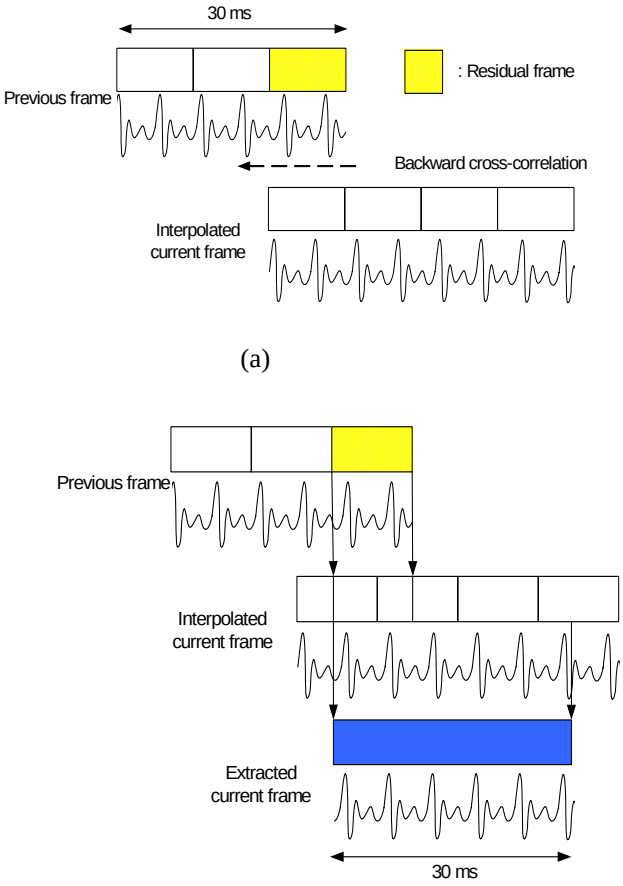


Figure 6: Procedure of pitch synchronization. (a) Backward cross-correlation. (b) Window extraction.

(b) Search of pitch synchronization

By calculating the forward and backward cross-correlation, we can find the position of synchronization between the previous window and the current window. The maximum cross-correlation values of these two searches are obtained as

$$Max\_R_B = \max(R_B(\tau)),\ \tau_B = \max^{-1}(R_B(\tau)) \qquad (9)$$

where $\tau_F$ is the shift of maximum forward cross-correlation and $\tau_B$ is the shift of maximum backward cross-correlation.

Figure 7(a) shows the method of forward search, which is calculated between the current window and the previous window. The maximum length of right shift in the current window is half of pitch period. Then, we obtain the start $S_{\tau_F}$ and end $E_{\tau_F}$ position of current window from the shift of the maximum forward cross-correlation.

Figure 7(b) shows method of backward search, and the way of search is reverse to the forward search. The maximum length of left shift in the current window is also half of pitch period. Then, we obtain the start $S_{\tau_B}$ and end $E_{\tau_B}$ position of current window from the shift of the maximum forward cross-correlation.

Only one of these two searches is chosen as the new pitch position based on the maximum value of cross-correlation.

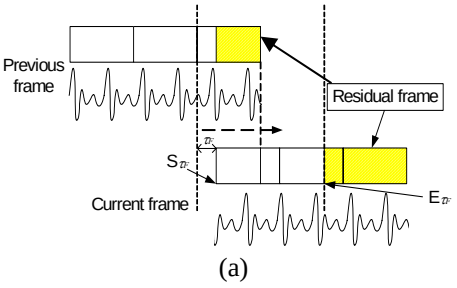


Figure 7: Search of pitch synchronization. (a) Forward search. (b) Backward search.

## 2.5 Frame boundary compensation

Since the signal at frame boundary may not be continuous, a triangle window is used to smooth the synthesized signal between the previous frame and the current frame, as illustrated in Figure 8. Finally, the synthesized signal is filtered by a low-pass filter to remove high frequency noise.

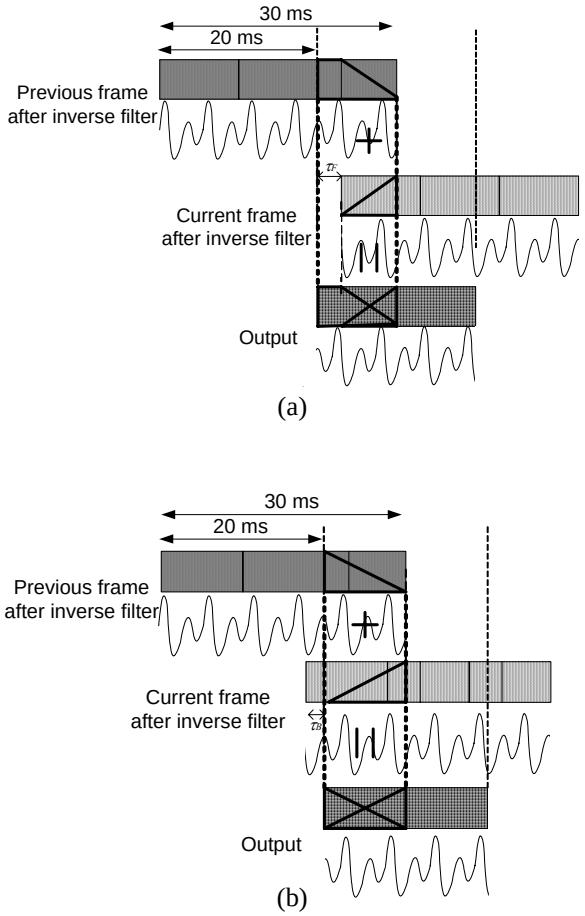


Figure 8: Frame boundary compensation. (a) with forward search. (b) with backward search.

Because of the discontinuity of signals between the previous frame and current frame, the synthesized signal is interpolated on the frame boundary and the compensated window boundary. This smoothing method is shown in Figure 9.

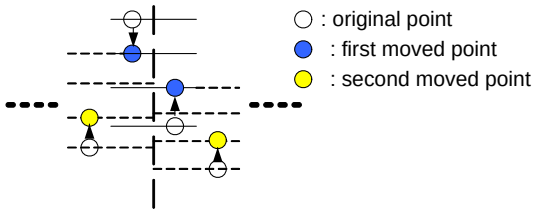


Figure 9: Smoothing method

## 3. Experimental results

To evaluate the performance of the proposed method, an informal subjective assessment of mean opinion score (MOS) evaluation is used. The speech quality assessment includes intelligibility (I), distortion (D) and naturalness (N), all are ranged from 1 to 5 with 5 meaning the best. These MOS evaluation results of male-to-female and female-to-male modifications are shown in Table 1 and Table 2, respectively. The evaluation results, shown as in these tables, indicate that when the pitch modification ratio is higher than 1/2, the experiment results are satisfactory and the speech signals sound natural.

Table 1: Male to Female

| Intelligibility (I) Distortion (D) Naturalness (N) | | | Formant modification ratio | | | |
|---|---|---|---|---|---|---|
| | | | 5/6 | 4/5 | 3/4 | 3/5 |
| pitch modification ratio | 2/3 | I | 4.65 | 4.79 | 4.79 | 4.59 |
| | | D | 3.93 | 4 | 3.93 | 3.86 |
| | | N | 4.36 | 4.43 | 4.22 | 3.29 |
| | 1/2 | I | 4.57 | 4.58 | 4.65 | 4.29 |
| | | D | 3.72 | 3.86 | 3.93 | 3.86 |
| | | N | 4.43 | 4.50 | 4.50 | 3.29 |
| | 2/5 | I | 4.51 | 4.44 | 4.52 | 4.29 |
| | | D | 3.43 | 3.43 | 3.50 | 3.51 |
| | | N | 3.72 | 3.72 | 4.07 | 3.72 |

Table 2: Female to Male

| Intelligibility (I) Distortion (D) Naturalness (N) | | | Formant modification ratio | | | |
|---|---|---|---|---|---|---|
| | | | 5/6 | 4/5 | 3/4 | 3/5 |
| pitch modification ratio | 3/2 | I | 4.65 | 4.71 | 4.86 | 4.64 |
| | | D | 3.64 | 3.79 | 4.14 | 3.29 |
| | | N | 4.00 | 4.15 | 4.29 | 3.43 |
| | 2/1 | I | 4.65 | 4.50 | 4.50 | 4.29 |
| | | D | 3.72 | 3.71 | 3.57 | 3.29 |
| | | N | 4.29 | 4.22 | 4.36 | 3.43 |
| | 5/2 | I | 4.07 | 4.15 | 4.22 | 4.14 |
| | | D | 3.29 | 3.50 | 3.36 | 3.22 |
| | | N | 3.79 | 3.22 | 3.50 | 2.93 |

## 4. Conclusion

A pitch-scale modification method based on formant extraction from resampled speech is proposed. The proposed method can successfully synthesize high quality speech. By changing the pitch and formant resampling ratios, this method can direct control the variation of formant structure and pitch harmonics separately. Experimental results show that the speech can be successfully modified to different timbre and tone with high quality.

## References


[1] F. Charpentier and M. Stella, "Diphone synthesis using an overlap-add technique for speech waveform concatenation," *Proc. ICASSP* Tokyo, pp. 2015-2018, 1986.

[2] H. Valbret, E. Moulines, and J.P. Tubach, "Voice transformation using PSOLA technique," *Proc. ICASSP*, vol. 1, pp. 145 – 148, 1992.

[3] J. Makhoul, "Linear Prediction: a tutorial review," *Proc. of the IEEE*, vol. 63, pp. 561-580, 1975.

[4] 王鴻彬, 國語聲訊處理, 碩士論文, 國立交通大學, 1995

[5] G. J. Lin, S. G. Chen, and T. Wu, "High Quality and Low Complexity Pitch Modification of Acoustic Signals," *Proc. ICASSP*, vol. 5, pp. 2987-2990, 1995.

[6] F. M. Gimenez de los Galanes, M. Savoji, and J. M. Pardo, "Speech synthesis system based on a variable decimation/interpolation factor," *Proc. ICASSP*, pp. 636-639, 1995.

[7] R. Vergin, D. O'Shaughnessy, and A. Farhat, "Time Domain Technique for Pitch Modification and Robust Voice Transformation," *Proc. ICASSP*, pp. 947-950, April 1997.

[8] B. Atal, L. Rabiner, "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 201 –212, Jun 1976.

[9] W. Zhang, G. Xu, and Y. Wang, "Pitch Estimation Based on Circular AMDF," *Proc.* ICASSP, pp. I-341-344, 2002.