Integrating Speaker-nonspecific Timbre Transformation to an HNM Based Speech Synthesis Scheme

Hung-Yan Gu* and Chen-Lin Tsai

Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology, Taipei 106, Taiwan

Abstract

In this paper, a speech signal synthesis scheme based on harmonic-plus-noise model (HNM) is extended to provide the function of speaker-nonspecific timbre transformation. To transform synthetic speech's timbre, we have developed a formant based frequency mapping method called piece-wise linear frequency mapping (PLFM). Additionally, a commonly adopted method is frequency axis scaling (FAS). Both methods have been integrated into the HNM based signal synthesis scheme, and a real-time speech synthesis system has been implemented according to this scheme. By using the speech files synthesized, perception tests are conducted. The results show that the proposed scheme can indeed transform the source timbre of a female adult into the timbre of a male adult, boy, or girl. In addition, the method PLFM is shown to be better than FAS in obtaining more masculine timbre.

Keywords: speech synthesis; timbre transformation; harmonic plus noise model; frequency axis scaling

Subject index: CO18
*Corresponding author. Email: guhy@mail.ntust.edu.tw

1. Introduction

To obtain natural synthesized speech, a large number of utterances from a speaker must be recorded, labeled, and segmented in order to train the relevant models, e.g., prosodic model (Chen *et al.* 1998) or hidden Markov model (HMM) (Hsia *et al.* 2010, Tokuda *et al.* 2002). Nevertheless, just one specific synthetic timbre can be produced from the training sentences uttered by any one speaker. A speech synthesis system will be more versatile if it can provide several timbres, e.g., the timbres of a man, a woman, a girl, or a boy, for the user to select. A direct approach to achieve this goal is to record utterances from each speaker who provides a timbre and then train each timbre's relevant models. Apparently, duplicated efforts and money will be spent to record and process training sentences from each new speaker. Therefore, we are motivated to study a more economical approach. The approach is to synthesize multiple speaker-nonspecific timbres using just one source speaker's utterances. Here, a speaker-nonspecific timbre means that the owner of the synthetic timbre cannot be identified but its gender (male or female) and rough age (child or adult) can be identified.

A basic technique to transform timbre is to scale the frequency axis of a speech signal's spectrum. On average, the vocal tract of a male adult is longer than that of a female adult. Hence, the formant frequencies of a male are lower than those of a female (O'Shaughnessy 2000). To transform a female's timbre into a male's timbre, we can scale down the frequency axis of a synthetic-speech spectrum to lower formant frequencies. However, there are other factors that must be considered and satisfied simultaneously besides frequency-axis scaling (FAS). For example, speaking rate and pitch-contour need to be independently controlled when a speech signal is to be synthesized.

In the past, a well known method that supports independent control of FAS and speaking rate is the use of a phase vocoder (Moore 1990, O'Shaughnessy 2000). However, it does not provide a precise and flexible mechanism for modifying pitch-contour. As for the technique of formant synthesis (O'Shaughnessy 2000), it can indeed support independent control of the three factors, FAS, speaking rate, and pitch-contour. Nevertheless, the parameters for formant synthesis may be analyzed

with incorrect values, and manual adjustment of those values is inevitably required. On the other hand, the recently proposed method by A. Mousa (Mousa 2010) is an example of a time-domain synthesis method that is based on PSOLA (pitch synchronous overlap and add) (Moulines and Charpentier 1990). Unfortunately, this method cannot support independent control of the three factors. What it can support is independent control of speaking rate and pitch-contour, and FAS is dependent on pitch contour. This implies that a boy's timbre cannot be synthesized when the source speech is recorded from a female adult. Among the time-domain synthesis methods proposed previously, we have however found a method named TIPW (time-proportioned interpolation of pitch waveform) (Gu and Shiu 1998) that can support independent control of the three factors and is also a variant of PSOLA. Nevertheless, the synthetic speech produced by TIPW still has some drawbacks in signal quality, common with most time-domain synthesis methods. Reverberation is perceivable when pitch-contour is considerably changed, and SNR (signal to noise ratio) is significantly degraded as compared to the recorded source speech.

To synthesize speech of high signal quality, some recently proposed methods such as STRAIGHT (Kawahara *et al.* 1999) can be considered. Nevertheless, it should be checked whether the method to be adopted can support convenient and independent control of the three factors mentioned. Furthermore, the factor of computation burden is also very important because we intend to build a speech synthesis system capable of timbre transformation and speech synthesis in real-time. Therefore, we have decided to adopt an HNM (Stylianou 2005, Stylianou 1996) based speech synthesis scheme (Gu and Zhou 2008). We will improve and extend this scheme to support the function of timbre transformation.

The method of FAS is effective for timbre transformation and is often used in phase vocoder based signal processing methods (Moore 1990, Tang *et al.* 2001). Nevertheless, according to our experiments, FAS is not satisfactory. For example, the male timbre, transformed from scaling down a female voice spectrum's frequency axis, may be perceived as sissy. Therefore, we have developed a different method, named piece-wise linear frequency mapping (PLFM), to accomplish timbre transformation. Furthermore, a more masculine timbre can be obtained if PLFM and FAS are combined and used to transform a female's speech signal. In Section 2, the methods of FAS and PLFM are explained in details. In Section 3, the HNM based speech synthesis scheme is improved and extended to support timbre transformation. Then, system construction and experiment results are given in Section 4. Finally, concluding remarks are given in Section 5.

2. Timbre transformation methods

To transform timbre in the frequency domain, one kind of signal model must be adopted to model the magnitude spectrum of a signal frame. In the past, many kinds of signal models have been proposed, including source-filter model (O'Shaughnessy 2000), excitation plus resonances model (Bonada and Serra 2007), sinusoidal model (Quatieri 2002), HNM, and others. Here, we adopt HNM to model the magnitude spectrum of a voice frame. HNM was proposed by Y. Stylianou (Stylianou 2005, Stylianou 1996). In HNM, a maximum voiced frequency (MVF) detection method is provided to divide a voice frame's spectrum into lower and higher frequency parts. The lower-frequency part is modeled as a sum of harmonic partials as in a sinusoidal model. In contrast, the higher-frequency part is roughly modeled with a smoothed spectral envelope (i.e. spectral magnitude envelope) that is represented with a few cepstrum coefficients. A figure that shows the division of magnitude spectrum into two parts is drawn in Figure 1. In this figure, the pulses of unequal magnitudes on the left side represent the harmonic partials while the smooth curve on the right side represents the spectral envelope of the high frequency noise.

For the harmonic part of a source spectrum (i.e., before timbre transformation) as shown in Figure 1, we denote the frequency (in Hz), amplitude, and phase (in radian) of the i-th partial with f_i, a_i, θ_i. As for the noise part, the spectral envelope across the entire frequency range is represented with 20 cepstrum coefficients although in Figure 1, only the envelope curve behind the MVF is drawn.

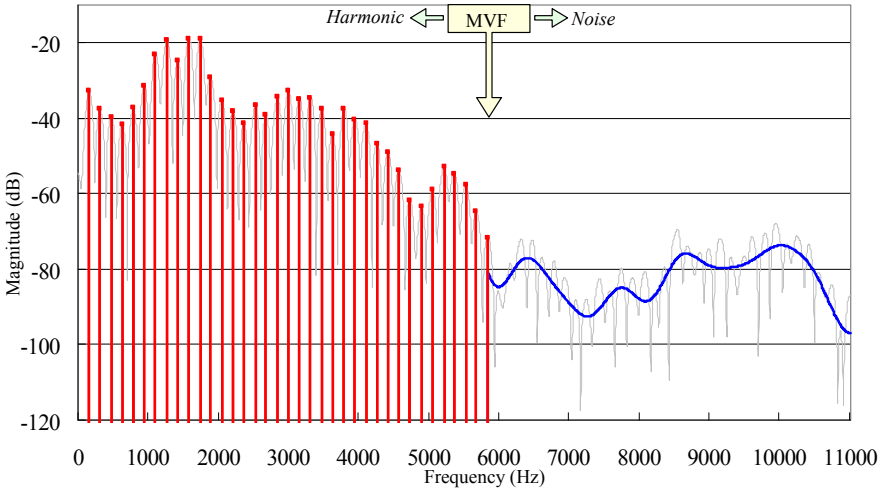

Figure 1. A spectrum divided into harmonic and noise parts.

When the 20 coefficients are appended with zeros and discrete Fourier transformed, the frequency bins and their amplitudes are denoted here with g_j and b_j, respectively. Based on the signal model of HNM, we have studied two timbre transformation methods, FAS and PLFM. They will be explained in the following subsections.

2.1. Frequency axis scaling

As indicated by the name FAS, this method just multiplies the frequencies, f_i and g_j, with a scaling factor, α, and keeps the amplitude and phase values intact. That is, set the frequency value of the i-th harmonic partial to $f_i' = \alpha f_i$. Similarly, the frequency value of the j-th bin in the noise part is set to $g_j' = \alpha g_j$. If α is smaller than 1, the transformed spectrum would have the formant frequencies lowered, which is equivalent to lengthening the vocal tract. Then, a masculine timbre can be synthesized by using the transformed spectrum. An example spectrum obtained by transforming the spectrum in Figure 1 with FAS and $\alpha = 0.8$ is shown in Figure 2. From this figure, it can be seen that the spectral envelopes for the harmonic and noise parts are both shrunk in frequency. As a result, the formant frequencies and MVF are all lowered. Also, another effect is that the spectrum would become empty within the frequency range from near 9,000Hz to the Nyquist frequency 11,025 Hz. Therefore, we need not synthesize noise signals within this frequency range.

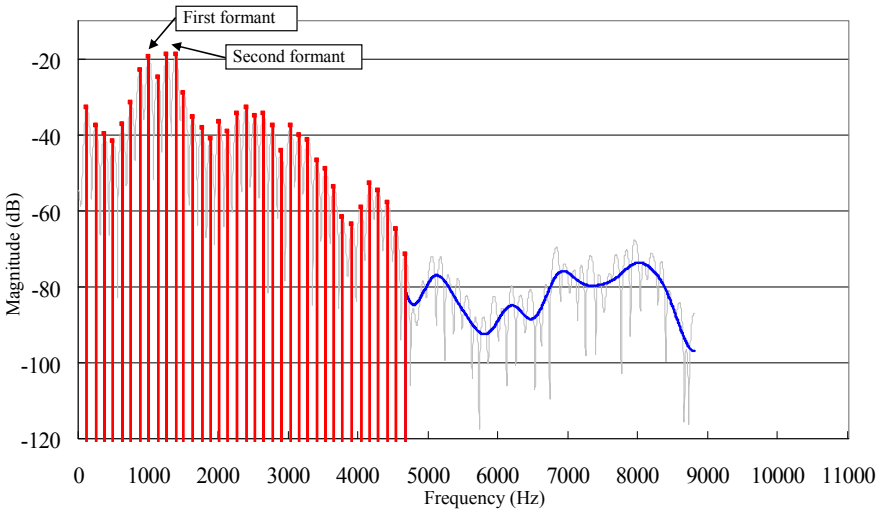

Figure 2. Transformed spectrum with FAS under $\alpha = 0.8$.

In contrast, if α is greater than 1, the transformed spectrum will have the formant frequencies raised, which is equivalent to shortening the vocal tract. This raising of formant frequencies can be

seen from the transformed spectrum drawn in Figure 3. Also, another effect that can be seen from Figure 3 is that the envelope curve for the noise part is cut out partially for those bins with frequencies, g'_j, greater than the Nyquist frequency.

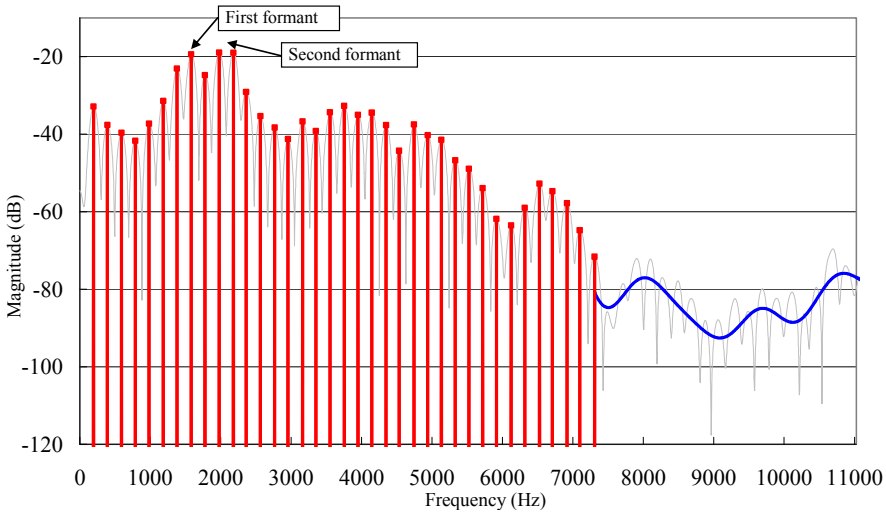

Figure 3. Transformed spectrum with FAS under $\alpha = 1.25$.

2.2. Piece-wise linear frequency mapping

This method is developed by applying acoustic knowledge of vowel production (O'Shaughnessy 2000). In details we know that /*u*/, /*a*/, and /*i*/ are the basic vowels employed by almost all languages, and that /*u*/, /*a*/, and /*i*/ are located at the three corners of the vowel triangle (O'Shaughnessy 2000). Also, we know that *F*1 and *F*2 of /*u*/ are very close, *F*1 and *F*2 of /*a*/ are very close, and *F*2 and *F*3 of /*i*/ are very close due to acoustic coupling (O'Shaughnessy 2000). By taking the average of the nearby formant frequencies as a reference point for frequency mapping, we may construct a mapping function with fewer reference points, and have the formant frequencies of the three vowels of the source speaker being mapped to the corresponding formant frequencies of the reference speaker in many phonetic contexts. On the other hand, since the timbre transformation methods studied here are intended for providing distinctive transformed timbres, it is acceptable that the transformed timbres are not similar to the reference speaker's timbre.

The developed method need not know the voice content in advance, and can conveniently be integrated into a real-time speech synthesis system. In this method, the formant frequencies (in Hz), *F*1, *F*2, and *F*3, of the vowels /*u*/, /*a*/, and /*i*/ uttered by the female who records the speech units (i.e. syllable here) for analyzing HNM parameters are analyzed first. Then, the values of *F*1 and *F*2 for the vowel /*u*/ are averaged to define the first reference frequency, R_1. The values of *F*1 and *F*2 for the vowel /*a*/ are averaged to define the second reference frequency, R_2. As to the third reference frequency, R_3, it is defined by averaging the values of *F*2 and *F*3 for the vowel /*i*/. On the other hand, we invited a male to record the vowels to analyze his formant frequencies. Then, we similarly obtained another set of reference frequencies, U_1, U_2, and U_3. Next, the two sets of reference frequencies were associated one to one, and 5 frequency pairs hence obtained, i.e., (R_1, U_1), (R_2, U_2), and (R_3, U_3) plus (0, 0) and (11,025, 11,025).

According to the 5 frequency pairs, a piece-wise linear frequency mapping function, $M(\bullet)$, can thus be constructed. An example mapping function constructed by using the pairs, (660, 447), (1,405, 1,065), and (3,455, 2,700), is illustrated in Figure 4. By using this mapping function, a source spectrum's formant frequencies that are near 660, 1,405, or 3,455 Hz will be mapped to the frequencies near 447, 1,065, or 2,700 Hz. In general, the frequency of the *i*-th partial, f_i, in the harmonic part of a source spectrum is mapped to $M(f_i)$. Similarly, the frequency, g_j, of the *j*-th bin in the noise part is mapped to $M(g_j)$. Take the spectrum in Figure 1 as an example source spectrum. When this spectrum is transformed by using the method PLFM, the resulting spectrum will be the one drawn in Figure 5. From this figure, it can be seen that in the harmonic part, the spacing between two

adjacent partials is incrementally increased from 0 Hz to the MVF.

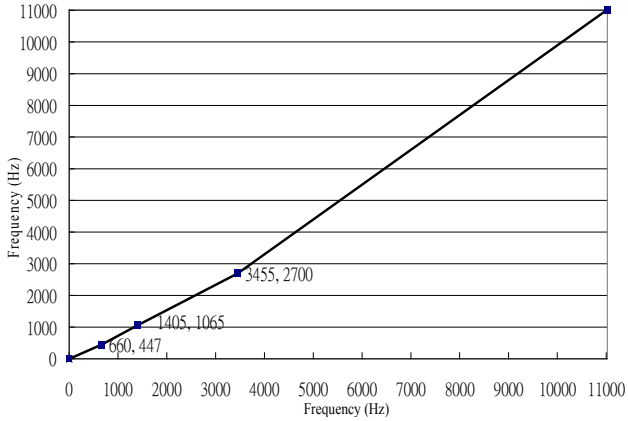

Figure 4. A piece-wise linear frequency mapping function.

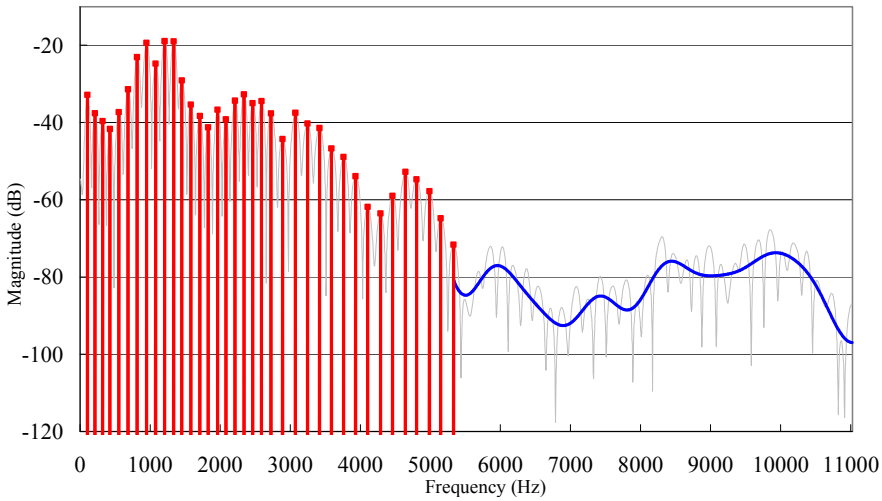

Figure 5. Transformed spectrum with PLFM.

In another spectrum transformation method, we can combine the two methods, PLFM and FAS. That is, the source spectrum is first transformed with PLFM to obtain an intermediate spectrum. Then, the intermediate spectrum is transformed again with FAS to obtain a final transformed spectrum. This combined transformation method is denoted as PLFM+FAS.

3. Timbre-transformation integrated HNM speech synthesis scheme

As mentioned earlier, we adopted the HNM based syllable-signal synthesis scheme (Gu and Zhou 2008) as the host speech synthesis scheme to consider the problem of timbre transformation. Therefore, we must inspect the processing flow of the original synthesis scheme and consider how to integrate the function of timbre transformation. We find that the function of timbre transformation can be implemented as a module and inserted between two original modules. After this insertion, the processing flow then becomes the one shown in Figure 6 where Block (c) is the newly inserted module responsible for timbre transformation.

Since the timbre transform methods, FAS and PLFM, executed in Block (c) of Figure 6 have already been explained in Section 2, they will not be discussed in the following. As for the functions executed in Block (a) and (b), they may be replaced if an HMM based synthesis method (Tokuda *et al.* 2002) is adopted. When an HMM is used to generate the spectral parameters for a control point (the term "control point" is used instead of "frame" in the synthesis stage), the spectral parameters may be mel-frequency cepstrum coefficients (MFCC) (O'Shaughnessy 2000), line spectrum frequencies (LSF) (O'Shaughnessy 2000), or discrete cepstrum coefficients (DCC) (Cappe and Moulines 1996).

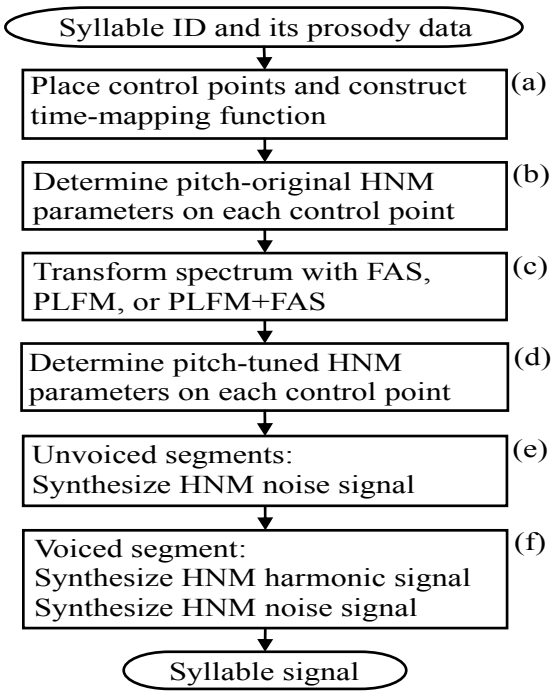

Figure 6. Timbre-transformation integrated HNM speech synthesis scheme.

No matter which kind of spectral parameter is used, a continuous spectral envelope can be computed from the spectral parameters. In terms of the continuous spectral envelope, we can sample it in some selected frequencies beforehand, and feed the sampled spectrum to the step of Block (c) for timbre transformation. Afterward, the steps of Block (d), (e), and (f) can still be followed. That is, the signal model, HNM, is just responsible for signal waveform synthesis no matter the spectral envelope on each control point (or frame) is generated by an HMM or not.

Currently, we do not use an HMM to generate the spectral parameters for a control point. One reason is that the number of recorded Mandarin sentences is only 375 (totally 2,925 syllables), which seems insufficient for training syllable (or phone) HMMs for speech synthesis. Nevertheless, these training sentences have been used to train prosody models. In addition, to analyze the HNM parameters for each Mandarin syllable, we asked the invited female to utter each Mandarin syllable in isolation, and recorded each syllable's signal. The HNM parameters of a signal frame include the frequency, amplitude, and phase values of each harmonic partial in the harmonic part, and the cepstrum coefficients for representing the noise part's spectral envelope. In terms of the analyzed HNM parameters for each frame of a syllable, the scheme in Figure 6 can then be followed step by step to synthesize a timbre-transformed syllable signal. In the following subsection, the operations executed in each block of Figure 6 except Block (c) will be briefly explained. If these explanations are not satisfactory, the original paper of the adopted speech synthesis scheme (Gu and Zhou 2008) may be referred to.

3.1 Place control points and construct time-mapping function

Here, "control point" is distinguished from analysis frame. In analysis stage, a syllable's signal is sliced into a sequence of overlapped frames. The frame width and shift are 512 and 256 sample points, respectively, under the sampling rate 22,050 Hz. Nevertheless, in synthesis stage, control points are placed along the planned duration of a syllable to be synthesized. The HNM parameters for a control point located at the voiced segment are obtained by interpolating the HNM parameters from its two corresponding analysis frames. Within the voiced segment, two adjacent control points are always placed 100 sample points (4.5 ms) apart. A fixed pace, 100 sample points, is adopted because a more accurate control of spectrum progressing is intended. Nevertheless, when synthesizing the unvoiced segment, HNM parameters of an analysis frame are directly copied and used for its corresponding

control point. That is, the pace between adjacent control points is not a constant when synthesizing the unvoiced segment.

To find the two corresponding analysis frames for a voiced control point, a time axis mapping function from the synthetic syllable to the source syllable is required. To obtain higher perceived fluency, a method is used to plan the phoneme durations of a synthetic syllable and the phonemes' durations are used to construct a piece-wise linear time mapping function (Gu and Zhou 2008).

3.2 Determine pitch-original HNM parameters

By using the time mapping function constructed, a control point's time location can be mapped to a time location on the source syllable's time axis. Suppose that the mapped time is t_m and $n=\lfloor t_m \rfloor$. Then, the n-th and (n+1)-th analysis frames' HNM parameters are taken to interpolate the HNM parameters for the control point. Interpolation is done in a linear manner but the phase values from the two analysis frames must be unwrapped beforehand (Gu and Zhou 2008). Let $\bar{A}_i$, $\bar{F}_i$, and $\bar{\theta}_i$ denote the amplitude, frequency, and phase of the i-th harmonic partial for the concerned control point, and A_i^n, F_i^n, and θ_i^n denote the amplitude, frequency, and phase of the i-th harmonic partial for the n-th analysis frame. Then, for example, $\bar{A}_i$ is interpolated as

$$\bar{A}_i = (t_m - n)(A_i^{n+1} - A_i^n) + A_i^n \,. \tag{1}$$

As for phase unwrapping, the formula is

$$\hat{\theta}_i^{n+1} = \theta_i^{n+1} - R \cdot 2\pi \,, \tag{2}$$

$$R = \left\lfloor \frac{1}{2\pi}\left(\theta_i^{n+1} - \theta_i^n + \theta_c\right) \right\rfloor, \quad \theta_c = \begin{cases} \pi, & \text{if } \theta_i^{n+1} \geq \theta_i^n \\ -\pi, & \text{otherwise} \end{cases},$$

where $\hat{\theta}_i^{n+1}$ denotes the unwrapped phase of θ_i^{n+1} versus θ_i^n. Because the pitch height defined by $\bar{F}_i$, i=1, 2, 3, …, is the original pitch predetermined in recording time, the parameters, $\bar{A}_i$, $\bar{F}_i$, and $\bar{\theta}_i$, obtained here are thus called pitch-original HNM parameters.

3.3 Determine pitch-tuned HNM parameters

The pitch-height of a voiced control point must be tuned in order to follow the pitch contour defined by the prosody unit. Nevertheless, the timbre must be kept consistent across a sequence of control points that have their pitches tuned. One principle is to keep the spectral envelope of each control point unchanged while the frequency values of the newly defined harmonic partials may be varied considerably. Therefore, the spectral envelope of a control point must be estimated from the pitch-original harmonic partials (i.e. those obtained in Block (b) of Figure 6) beforehand. Then, the estimated spectral envelope is used to determine the amplitude and phase values of the pitch-tuned (or newly defined) harmonic partials. The two steps, estimating spectral envelope and determining new harmonic partials' parameter values, if executed in order will obtain a timbre consistent to the source speaker. Nevertheless, if we intend to obtain a transformed timbre, the frequency values of the pitch-original harmonic partials must be mapped through Block (c) of Figure 6 before they are used to estimate the spectral envelope.

A spectral envelope can be estimated either with a global approximation method such as discrete cepstrum (Cappe and Moulines 1996) or with a local approximation method such as cubic-spline interpolation. Here, the pitch-original HNM parameters are directly taken to perform Lagrange interpolation based local approximation because the synthetic timbre is perceived as accurate enough and the computation burden can be decreased. In details, let $\tilde{F}_k$, $\tilde{A}_k$, and $\tilde{\theta}_k$ denote the frequency, amplitude, and phase of the k-th pitch-tuned harmonic partial. The value of $\tilde{F}_k$ can be directly set to k times of the pitch frequency generated by the prosody module and assigned to the current control point. In contrast, to compute the values of $\tilde{A}_k$ and $\tilde{\theta}_k$, we first find a pitch-original harmonic

frequency $\bar{F}_j$, from $\bar{F}_1$, $\bar{F}_2$, $\bar{F}_3$, …, that is nearest to and less than $\tilde{F}_k$. Then, the four pitch-original partials of the frequencies, $\bar{F}_{j-1}$, $\bar{F}_j$, $\bar{F}_{j+1}$, and $\bar{F}_{j+2}$, are used to perform order three Lagrange interpolation. That is,

$$\tilde{A}_k = \sum_{m=j-1}^{j+2} \bar{A}_m \cdot \prod_{\substack{h=j-1 \\ h\neq m}}^{j+2} \frac{\tilde{F}_k - \bar{F}_h}{\bar{F}_m - \bar{F}_h} \quad . \tag{3}$$

Similarly, the value of $\tilde{\theta}_k$ can be computed with Equation (3). If the Lagrange interpolation of Equation (3) is applied to the partials in the harmonic part of Figure 1, the estimated spectral envelope would be the curve shown in Figure 7.

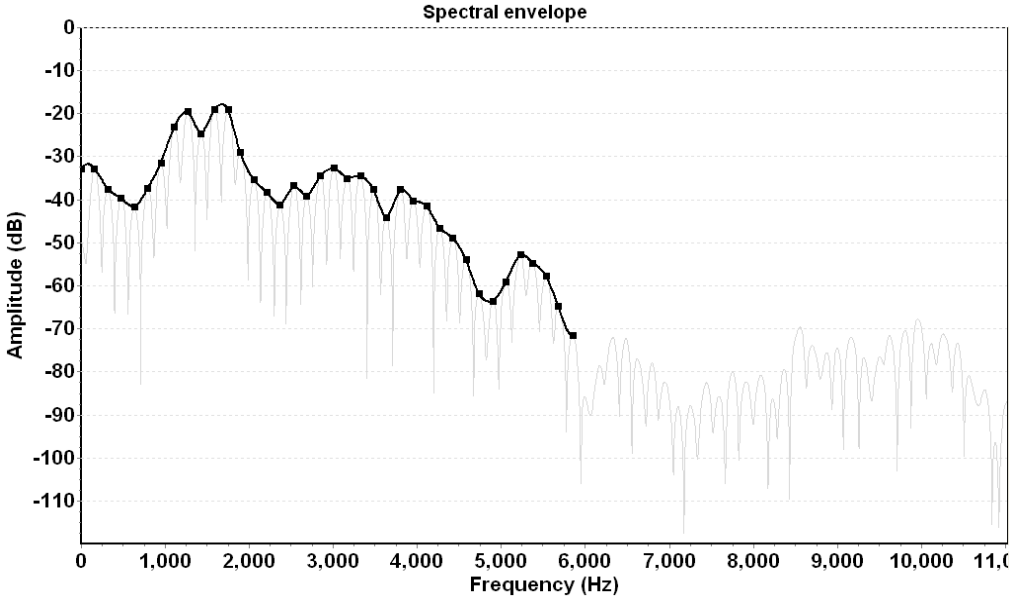

Figure 7. Estimated spectral envelope with Lagrange interpolation.

3.4 Synthesis of HNM harmonic signal

As the basic concept of HNM, the speech signal is synthesized as the noise signal $D(t)$ alone for the unvoiced segment of a syllable. In contrast, the speech signal is synthesized as the harmonic signal $H(t)$ plus the noise signal $D(t)$ for the voiced segment. Here, for the harmonic signal, $H(t)$, between the n-th and $(n+1)$-th control points, its sample values are computed with the following equations:

$$H(t) = \sum_{k=0}^{L} a_k^n(t) \cos\left(\phi_k^n(t)\right) \quad , \quad t = 0,1,...,99 \,, \tag{4}$$

$$a_k^n(t) = \tilde{A}_k^n + \frac{t}{100}(\tilde{A}_k^{n+1} - \tilde{A}_k^n) \,, \tag{5}$$

$$\phi_k^n(t) = \phi_k^n(t-1) + 2\pi f_k^n(t) / 22{,}050 \quad , \quad \phi_k^n(0) = \hat{\theta}_k^n \,, \tag{6}$$

$$f_k^n(t) = \tilde{F}_k^n + \frac{t}{100}(\tilde{F}_k^{n+1} - \tilde{F}_k^n) \,, \tag{7}$$

where L is number of harmonic partials, 100 is the number of samples between two adjacent control points, 22,050 is the sampling rate, $a_k^n(t)$ is the time-varying amplitude of the k-th partial at time t from the start of the n-th control point, $\phi_k^n(t)$ is the cumulated phase for the k-th partial at time t, $f_k^n(t)$ is the time-varying frequency for the k-th partial at time t, and $\hat{\theta}_k^n$ is the unwrapped phase of $\tilde{\theta}_k^n$ versus $\hat{\theta}_k^{n-1}$. In Equations (5) and (7), linear interpolation is used, which seems good enough according to perception tests.

Note that, when Equation (4) is used to synthesize signal samples, the cumulated phase, $\phi_k^n(t)$, is generally not continued at the boundary time points, i.e. t=0 or t=99. To avoid such discontinuity, the

amount of mismatched phase at the boundary point must be computed beforehand, and shared to the 100 sample points between two adjacent control points.

3.5 Synthesis of HNM noise signal

For the noise signal, $D(t)$, we decided to synthesize a summation of sinusoidal signal components as proposed in Stylianou's thesis (Stylianou 1996). Let G_k be the frequency of the k-th sinusoidal and G_k =100·k (Hz). The index k of G_k is not started from 1. Its starting value, K_s^n, is determined by the MVF of the n-th control point, *i.e.* $K_s^n = \lceil \mathrm{MVF}(n) / 100 \rceil$. In addition, let B_k^n be the amplitude of the k-th sinusoid on the n-th control point. To determine the value of B_k^n, the 20 cepstrum coefficients, of the n-th control point, representing the noise spectral envelope are first appended with zeros and inversely Fourier transformed to the spectral domain. Then, the value B_k^n can be computed by linearly interpolating the two adjacent bins whose frequencies surround G_k.

The values of the noise-signal samples between the n-th and (n+1)-th control points are computed with the following equations:

$$D(t) = \sum_{k=K_s}^{K_e} b_k^n(t) \cos\left(\gamma_k^n + t \cdot 2\pi G_k / 22{,}050\right), \quad t = 0,1,...,99, \tag{8}$$

$$b_k^n(t) = B_k^n + \frac{t}{100}(B_k^{n+1} - B_k^n), \tag{9}$$

$$\gamma_k^n = \gamma_k^{n-1} + 100 \cdot 2\pi G_k / 22{,}050, \tag{10}$$

where K_S is set to the lesser of K_s^n or K_s^{n+1}, and γ_k^n is the initial phase for the k-th sinusoid on the n-th control point. In Equation (9), the time-varying amplitude, $b_k^n(t)$, is just linearly interpolated.

4. System construction and experiments

Our Mandarin speech synthesis system was implemented by adopting the techniques proposed in the works (Gu and Wu 2009, Gu and Zhou 2008). This system can be subdivided into three components, i.e., text analysis, prosody parameter generation, and signal waveform synthesis. Here, we integrate the timbre transformation methods, FAS and PLFM, into the component of signal waveform synthesis. That is, the HNM based syllable signal synthesis scheme (Gu and Zhou 2008) is adopted and extended to the one shown in Fig. 6.

The processing of the components, text analysis and prosody parameter generation, will be briefly described in Subsection 4.1. By connecting the two components with the signal synthesis component whose processing steps are shown in Figure 6, we have constructed a Mandarin speech synthesis system capable of timbre-transformation. Then, this system is fed a common input text to synthesize speech signals with different timbre-transformation methods. In terms of the synthetic speech files, perception tests are then conducted.

4.1 Text analysis and prosody parameter generation

When the text of a Chinese sentence is inputted, the text-analysis component will parse it into a sequence of words by looking up a word dictionary. While the dictionary is being checked, the pronunciation syllables and tones of each searched word's comprising characters are also obtained. Next, tone sandhi rules are applied. Then, the sequence of syllables and tones is fed to the prosody parameter generation component to generate prosody parameters for each syllable.

The prosody parameters of a syllable include pitch contour, duration, and intensity. Here, we trained a separate artificial neural network (ANN) for each of the three prosody parameters. The reason for not building a combined ANN for the three parameters is that we have only 375 recorded training sentences which are not sufficient. Nevertheless, by appropriately grouping the values of the contextual data items for a concerned syllable, the generated values of the prosody parameters can

still present an acceptable naturalness level. As for the structure of the ANN and the details of the adopted contextual data items, readers are referred to relevant works (Chen *et al.* 1998, Gu and Wu 2009, Lin *et al.* 2004).

Since the pitch contour ANN was trained with a female's recorded sentences, the generated pitch contour needs to be shifted to synthesize the timbre of an intended individuality (e.g., male or female adult). Let *Pa* be the average pitch, in logarithmic Hz scale, of the source speaker who utters the training sentences, and *Qa* be the defined average pitch, in logarithmic Hz scale, for the intended individuality. Then, the generated pitch contour, $P(t)$, is shifted to the target pitch contour, $Q(t)$, according to the simple formula,

$$\log\left(Q(t)\right) = \log\left(P(t)\right) - Pa + Qa\,, \quad t = 0,1,2,...,N-1\,. \tag{11}$$

Where N is the number of control points in the voiced segment of a syllable. As to the value of *Pa*, it can be computed in advance by averaging the estimated pitch frequencies, in logarithmic Hz scale, of all voiced frames.

4.2 Perception tests of timbre-transformed synthetic speech

In the first type of perception tests, the method of FAS was used to transform the source timbre into the timbres of a male adult, a boy, and a girl. To obtain the timbre of a male adult, we set the scaling factor, α, of FAS to 0.8, and set the average pitch for pitch-contour generation to 120Hz. On the other hand, we set the scaling factor of FAS to 1.2 to obtain the timbres of a boy and a girl. As to the average pitches for boy and girl, they were set to 140 Hz and 280 Hz, respectively. By using these settings, three speech files were synthesized, respectively, which can be listened to by accessing the web page, http://sites.google.com/site/fungriam/ . Then, these speech files were played to each of 12 invited participants. Each participant was asked if he (or she) agreed that the three synthetic speech files' timbres are male adult, boy, and girl, respectively. As a result, the 12 participants all agreed on the synthetic timbres' individualities, i.e. a male adult, a boy, and a girl.

In the second type of perception tests, we prepared 6 synthetic speech files beforehand, which were denoted as *AA*, *AB*, *AC*, *AD*, *AX*, and *AY*, respectively. In synthesizing *AA*, *AB*, *AC*, and *AD*, the method of FAS was used to transform their timbres. The scaling factor, α, was set to 0.9, 0.8, 0.7, and 0.6, respectively. On the other hand, *AX* was synthesized by using PLFM with the mapping function shown in Figure 4, and *AY* synthesized by using PLFM+FAS, i.e. FAS with scaling factor, 0.9, was executed after executing PLFM. The 6 speech files can also be listened to by accessing the same web page mentioned.

Here, the participants are the same 12 persons as mentioned above. During a person's perception test, we allowed him to play each of the 6 files again and again. Then, he was asked which timbre of *AA*, *AB*, *AC*, and *AD* is most similar to the timbre of *AX*. Similarly, he was also asked which of the four timbres is most similar to that of *AY*. As a result, eleven of the twelve persons recognized that *AB* is most similar to *AX*, and nine persons recognized that *AC* is most similar to *AY*. Based on these results, we next asked each of the participants to compare *AB* with *AX*, and give scores about which was more masculine and which was more intelligible. Here, the score, 2 (or -2), was defined as *AX* being significantly more (or less) masculine or intelligible than *AB*. If *AX* was just slightly more (or less) masculine or intelligible than *AB*, the score, 1 (or -1), was defined. Otherwise, the score, 0, should be given to indicate that they cannot be distinguished. Similarly, we also asked each of the participants to compare *AC* with *AY*, and give scores about timbre-masculinity and intelligibility.

After analyzing the scores given by the participants, we obtained the averaged scores and standard deviations as shown in Table 1. In detail, the averaged scores, 0.75 and 0.67, were obtained in the comparisons of timbre-masculinity between *AB* and *AX* and between *AC* and *AY*, respectively. Their standard deviations were 0.92 and 0.62, respectively. As to intelligibility, the averaged scores were 0.33 and 0.25 for the comparisons between *AB* and *AX* and between *AC* and *AY*, respectively. Their standard deviations were 0.94 and 0.92, respectively. According to the averaged scores, 0.75 and 0.67, it is seen that the transformation method, PLFM, can provide more masculinity in timbre

than the method, FAS. Also, according to the averaged scores, 0.33 and 0.25, it seems that the method PLFM will induce slightly less degradation in intelligibility than the method FAS.

Table 1. Averaged scores and standard deviations for the perception tests.

	AX vs. *AB*	*AY* vs. *AC*
Masculinity	0.75 (0.92)	0.67 (0.62)
Intelligibility	0.33 (0.94)	0.25 (0.92)

Why will PLFM provide more masculinity than FAS in their transformed timbres? The major reason, we think, is that linearly lengthening the vocal tract will result in nonlinear shrinking of the formant frequencies for real persons. This is because the vocal tract is comprised of the mouth and pharynx cavities and there are some interactions between the two cavities (O'Shaughnessy 2000). This nonlinear shrinking of formant frequencies can somehow be simulated by PLFM. In contrast, linearly scaling down frequency (equivalent to linearly lengthening the vocal tract) by FAS will just obtain linearly shrunken formant frequencies.

4.3 Timbre similarity measure

As mentioned in Section 4.2, eleven of the twelve persons recognize that *AB* is most similar to *AX*, and nine persons recognize that *AC* is most similar to *AY*. Hence, it is interesting whether an automatic timbre-similarity measure can be designed such that the measured distances will reflect the perceived timbre similarity. Since MFCC are the commonly adopted features for speech recognition, we are thus motivated to use 13 MFCC analyzed from a frame as the features for measuring timbre similarity. Two parallel synthetic speech files are first sliced into a sequence of frames with the same frame length, 512 points, and the same frame shift, 128 points. Then, MFCC analyzed from each pair of corresponding voiced frames are taken to compute a geometric distance. Next, these distances collected from different frames are averaged to define the measured timbre similarity distance. According to this definition, the measured timbre similarity distances between *AX* (or *AY*) and AA, AB, AC, and AD are listed in Table 2. From the first row of this table, it can found that the distance, 6.946, between *AX* and *AB* is not the smallest among the four distances although *AB* is perceived to be most similar to *AX*. Similarly, it can be found from the second row that the distance, 6.950, between *AY* and *AC* is also not the smallest among the four distances although *AC* is perceived to be most similar to *AY*. Therefore, we fail to propose an automatic timbre similarity measure. We think an automatic voice-timbre similarity measure that can reflect the perceived timbre similarity is still a research issue and needs more studies.

Table 2. Measured timbre-similarity distances with MFCC.

	AA	*AB*	*AC*	*AD*
AX	4.357	6.946	8.924	9.796
AY	2.475	4.448	6.950	8.757

5. Concluding remarks

In this paper, we have proposed a speaker-nonspecific timbre transformation method, PLFM, based on HNM. This method and another commonly used method, FAS, were integrated to the HNM based syllable-signal synthesis scheme to form a timbre-transformation capable speech synthesis scheme. Although FAS is a commonly used method for timbre transformation, it can be combined with PLFM to define a new timbre transformation method. The advantages of the two timbre transformation methods, FAS and PLFM, include at least the two points. First, they can provide distinct transformed timbres while the speech content need not be known and no complicated timbre transformation models need to be trained in advance. Secondly, they just consume small amounts of computation time and are hence convenient for being used in implementing a real-time speech synthesis system.

According to the extended speech synthesis scheme, we have built a practical Mandarin speech synthesis system capable of timbre transformation. The timbre transformed synthetic speech files were used to conduct perception tests. The results show that the improved and extended scheme is very effective in timbre transformation. That is, the source timbre of a female adult can indeed be transformed into the timbre of a male adult, boy, or girl. In addition, the method PLFM is shown to be

better than FAS for obtaining masculine timbre. Also, the speech transformed by PLFM is slightly more intelligible than that by FAS.

Acknowledgment

The financial support provided by National Science Council, Taiwan, under the grant, NSC 98 – 2221 – E – 011 – 116, is gratefully acknowledged.

Nomenclature

A_i^n , F_i^n , θ_i^n	amplitude, frequency, and phase of the *i*-th harmonic partial after analyzing the *n*-th analysis frame
$\bar{A}_i$, $\bar{F}_i$, $\bar{\theta}_i$	linear interpolated amplitude, frequency, and phase for the *i*-th pitch-original harmonic partial
$\tilde{A}_k$, $\tilde{F}_k$, $\tilde{\theta}_k$	Lagrange interpolated amplitude, frequency, and phase for the *k*-th pitch-tuned harmonic partial
$a_k^n(t)$, $f_k^n(t)$, $\theta_k^n(t)$	linear interpolated amplitude, frequency, and phase for the *k*-th harmonic partial at time *t* which is between the *n*-th and (*n*+1)-th control points
B_k^n , G_k , γ_k^n	amplitude, frequency, and initial phase of the *k*-th sinusoid of the noise signal on the *n*-th control point
$b_k^n(t)$	linear interpolated amplitude for the *k*-th sinusoid of the noise signal at time *t* which is between the *n*-th and (*n*+1)-th control points
K_s^n	starting value of the index *k* for synthesizing the noise signal
a_i , b_i	amplitudes of the *i*-th harmonic partial and the *i*-th frequency bin
f_i , g_i	frequencies of the *i*-th harmonic partial and the *i*-th frequency bin
$D(t)$	noise-signal sample synthesized at time point *t*
$F1$, $F2$, $F3$	first, second, and third formant frequencies
$H(t)$	harmonic-signal sample synthesized at time point *t*
$P(t)$, $Q(t)$	generated and target pitch-contours as functions of time *t*
Pa, Qa	average pitches of the source speaker and the intended individual
R_1, R_2, R_3	first set of reference frequencies analyzed from a female's utterances
U_1, U_2, U_3	second set of reference frequencies analyzed from a male's utterances
α	frequency-axis scaling factor
θ_i	phase of the *i*-th harmonic partial

Reference

Bonada, J. and Serra X., 2007. Synthesis of the singing voice by performance sampling and spectral models. *IEEE signal processing magazine*, 24(2), 67-79.

Cappe, O. and Moulines E., 1996. Regularization techniques for discrete cepstrum estimation. *IEEE signal processing letters*, 3(4), 100-102.

Chen, S.H., Hwang S.H., and Wang Y.R., 1998. An RNN-based prosodic information synthesizer for Mandarin text-to-speech. *IEEE transaction on speech and audio processing*, 6(3), 226-239.

Gu, H.Y. and Wu C.Y., 2009. Model spectrum-progression with DTW and ANN for speech synthesis. *The sixth international conference of electrical engineering / electronics, computer, telecommunications and information technology*, 6-9 May Pattaya, Thailand. Piscataway, NJ: IEEE Operation Center, 1010-1013.

Gu, H.Y. and Zhou Y.Z., 2008. An HNM based scheme for synthesizing Mandarin syllable signal. *International journal of computational linguistics and Chinese language processing*, 13(3), pp. 327-341.

Gu, H.Y. and Shiu W.L., 1998. A Mandarin-syllable signal synthesis method with increased flexibility in duration, tone and timbre control. *Proceedings of the National Science Council ROC(A)*, 22(3), 385-395.

Hsia, C.C., Wu C.H., and Wu J.Y., 2010. Exploiting prosody hierarchy and dynamic features for pitch modeling and generation in HMM-based speech synthesis. *IEEE transaction on audio, speech, and language processing*, 18(8), 1994-2003.

Kawahara, H., Masuda-Katsuse I., and de Cheveigne A., 1999. Restructuring speech representations using a pitch adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. *Speech communication*, 27(3), 187-207.

Lin, C.T., *et al.*, 2004. A novel prosodic-information synthesizer based on recurrent fuzzy neural network for the Chinese TTS system. *IEEE transaction on systems, man, and cybernetics*, 34(1), 309-324.

Moore, F.R., 1990. *Elements of computer uusic*. Englewood Cliffs, NJ: Prentice-Hall.

Moulines, E. and Charpentier E., 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication*, 9(5), 453-467.

Mousa, A., 2010. Voice conversion using pitch shifting algorithm by time stretching with PSOLA and re-sampling. *Journal of electrical engineering*, 61(1), 57-61.

O'Shaughnessy, D., 2000. *Speech communications: human and machine.* Piscataway, NJ: IEEE Press.

Quatieri, T.F., 2002. *Discrete-time speech signal processing*, Upper Saddle River, NJ: Prentice-Hall.

Stylianou, Y., 2005. Modeling speech based on harmonic plus noise models. *In*: G. Chollet *et al*., eds. *Nonlinear speech modeling and applications.* Berlin: Springer-Verlag, 244-260.

Stylianou, Y., 1996. *Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification.* Thesis (PhD), Ecole Nationale Supèrieure des Télécommunications.

Tang, M., Wang C., and Seneff S., 2001. Voice transformations: from speech synthesis to mammalian vocalizations. *In*: *European conference on speech communication and technology*, 3-7 September Aalborg, Denmark. International Speech Communication Association, 353-356.

Tokuda, K., Zen H., and Black A.W., 2002. An HMM-based speech synthesis system applied to English. *In*: *IEEE workshop on speech synthesis*, 11-13 September Santa Monica, CA. Piscataway, NJ: IEEE Operation Center, pp. 227-230.