An HNM-based Speaker-nonspecific Timbre Transformation Scheme for Speech Synthesis

Hung-Yan Gu, Chen-Lin Cai, and Song-Fong Cai

National Taiwan University of Science and Technology, Taipei, Taiwan
E-mail: {guhy, m9615030, m9615069}@mail.ntust.edu.tw

Abstract

In this paper, the harmonic-plus-noise model (HNM) based speech signal synthesis scheme studied previously is further 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). In addition, a commonly adopted method is frequency axis scaling (FAS). Both methods have been integrated into our HNM speech synthesis scheme, and a real-time synthesis system is implemented according to this scheme. The perception test 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 for obtaining more manful timbre.

1. Introduction

To obtain natural synthetic speech, a large amount of utterances from a speaker must be recorded, labeled, and segmented in order to train relevant models, e.g., prosodic model or HMM (hidden Markov model) [1]. However, just one specific timbre of the speaker who utters the training sentences can be synthesized. A speech synthesis system will be more versatile if it can provide several synthetic timbres, e.g., a male or female adult, a girl, or a boy, for the user to select. A simple approach to achieve this goal is to record utterances from each speaker who provides a timbre and then train each timbre's speech models. However, duplicated efforts and money would be spent to record and process training sentences for each new speaker. Therefore, we are motivated to study a more economical approach. The approach is to synthesize distinct timbres of nonspecific speakers in terms of timbre transformation that uses 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 formant frequencies of a male are lower than those of a female [2]. To transform a female's timbre into a male's timbre, we can scale down the frequency axis of a synthetic speech's spectrum to lower formant frequencies. However, there are relevant factors that must be considered simultaneously besides frequency-axis scaling (FAS). For example, speaking rate and pitch-contour should be independently controlled when speech signal is synthesized.

In the past, a well known method than can provide independent control of FAS and speaking rate is the technique of phase vocoder [2, 3]. However, it does not provide a precise and flexible mechanism for modifying pitch-contour. On the other hand, we had proposed a time-domain speech synthesis method [4] that is a variant of PSOLA (pitch synchronous overlap and add) [5], and can support independent control of the three factors. However, the synthetic speech has some drawbacks in signal quality. Reverberation is perceivable and SNR (signal to noise ratio) is significantly degraded.

To synthesize speech with high signal quality, some methods may be considered. However, it must be checked if the selected method can support convenient and independent control of the three factors mentioned. Also, another factor of computation burden is important because we intend to build a speech synthesis system capable of real-time timbre transformation and speech synthesis. Therefore, we decide here to study an HNM (harmonic-plus-noise model) [6] based and extended scheme for timbre transformation and speech synthesis.

The method of FAS is effective for timbre transformation and is often used in phase vocoder based signal processing methods [3, 7]. Besides adopting FAS, we have also developed another method, piece-wise linear frequency mapping (PLFM), to do timbre transformation.

2. Timbre transformation methods

To transform timbre in the frequency domain, one signal model at least must be used to model the magnitude spectrum of a signal frame. Here, we adopt HNM to model the magnitude spectrum of a speech frame. HNM was proposed by Y. Stylianou [6]. In HNM, an MVF (maximum voiced frequency) detection method is provided to divide a speech frame's spectrum into lower and higher frequency parts. The lower-frequency part is modeled as a sum of harmonic partials as in sinusoidal [8] or sinusoids-plus-noise model [9]. In contrast, the higher-frequency part is roughly modeled with a smoothed spectral envelope that is represented with some cepstrum coefficients. A figure that shows the division of magnitude spectrum into two parts is drawn in Fig. 1. In this figure, the pulses at the left side represent the harmonic partials while the smooth curve at the right side represents the spectral envelope of the high frequency noise.

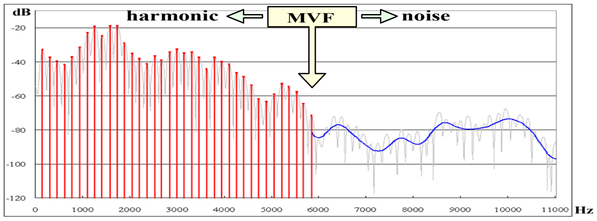

Fig. 1. Spectrum partitioned to harmonic and noise parts.

Here, we denote the frequency, amplitude, and phase of the i-th partial with f_i, a_i, θ_i for a source spectrum (i.e., before timbre transformation). As for the noise part, the spectral envelope across entire frequency range is represented with 20 cepstrum coefficients although in Fig. 1, only the envelope after the MVF is drawn. 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 spectrum model of HNM, two timbre transformation methods, FAS and PLFM, are studied.

2.1. Frequency axis scaling

As indicated by the name FAS, this method just multiply the frequencies, f_i and g_j, with a scaling factor, α, and keeps the amplitude and phase values intact. That is, let the frequency value of the i-th harmonic partial be $f_i' = \alpha \cdot f_i$ after timbre transformation. Similarly, let $g_j' = \alpha \cdot g_j$. If α is smaller than 1, the transformed spectrum would have the formant frequencies lowered, which is equivalent to lengthen the vocal-track. Then, a more manful timbre can be synthesized by using the transformed spectrum. An example spectrum obtained by transforming the spectrum in Fig.1 with FAS and $\alpha = 0.8$ is shown in Fig. 2. From this figure, it can be seen that the spectral envelopes for the harmonic and noise parts are both shrunk in frequency axis, and hence formant frequencies and MVF are all lowered. Also, another observable effect is that the spectrum for the frequency range, from nearby 9,000Hz to the Nyquist frequency 11,025Hz, become empty. Therefore, in our synthesis program, we do not synthesize noise signals within such frequency range.

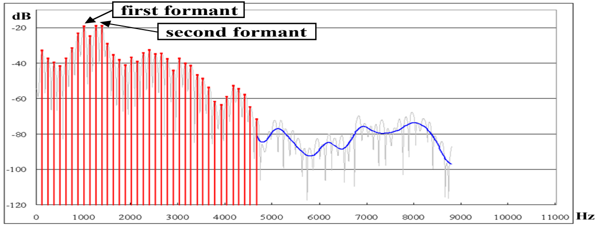

Fig. 2. Transformed spectrum with FAS under $\alpha = 0.8$.

In contrast, if α is greater than 1, the transformed spectrum would have the formant frequencies raised, which is equivalent to shorten the vocal-track. This raising of formant frequencies can be seen from the example transformed spectrum in Fig. 3. Also, another effect that can be seen is the envelope curve for the noise part is cut out partially for those bins with frequencies, g_j', greater than the Nyquist frequency.

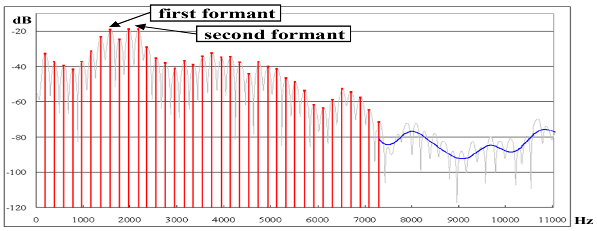

Fig. 3. Transformed spectrum with FAS under $\alpha = 1.25$.

2.2. Piece-wise linear frequency mapping

We develop this method by applying the acoustic knowledge about vowel production [2]. This method needs not to know the voice content in advance, and is simple to implement. In this method, the formant frequencies, F_1, F_2, and F_3, of the vowels /u, a, i/ uttered by the female who records the synthesis units for our speech synthesis system 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 select a male whose timbre is manful enough to record the vowels for analyzing formant frequencies.

Then, we can similarly obtain another set of reference frequencies, U_1, U_2, and U_3. Next, the two sets of reference frequencies are associated one to one, and 5 frequency pairs are formed, 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. By using this mapping function, a source spectrum's formant frequencies that are nearby R_1, R_2, or R_3 will be mapped to the frequencies nearby U_1, U_2, or U_3. In general, the frequency of the i-th partial, f_i, in the source spectrum is mapped to $M(f_i)$. Similarly, the frequency, g_j, of a bin in the noise part is mapped to $M(g_j)$. Take the spectrum in Fig. 1 as an example source spectrum. When this spectrum is transformed by using the method PLFM, the resulted spectrum would be the one drawn in Fig. 4. From this figure, it can be seen that the spacing between two adjacent harmonic partials is incrementally increased from the low frequency end to the MVF.

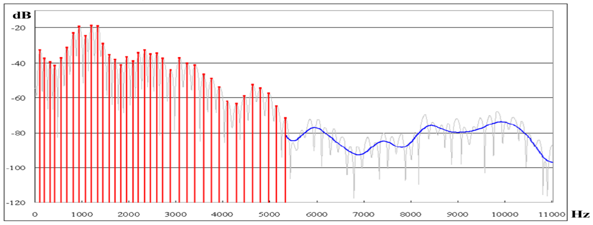

Fig. 4. Transformed spectrum with PLFM.

As 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. HNM based transform.-synthesis scheme

Here, we continue to adopt syllable as the unit for synthesizing speech signal. This is because we will base on the previously developed HNM-based syllable-signal synthesis scheme [10] to consider how the function of timbre-transformation can be integrated into the scheme. As a result, the integrated scheme is the one shown in Fig. 5.

According to the scheme in Fig. 5, the function of timbre transformation is executed in Block (c). As to the functions executed in Block (a) and (b), they may be replaced if HMM is adopted. Suppose HMM is used to generate the spectral parameters for a control point. Spectral parameters may be MFCC (mel frequency cepstrum coefficient) [2] or DCC (discrete cepstrum coefficient) [11]. Anyway, a continuous spectral envelope can be obtained in terms of the spectral parameters. For a continuous spectral envelope, we can sample it in frequency 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 be followed.

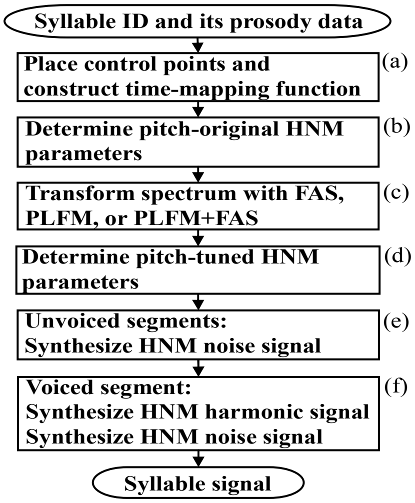
Syllable ID and its prosody data
Place control points and construct time-mapping function
(a)
Determine pitch-original HNM parameters
(b)
Transform spectrum with FAS, PLFM, or PLFM+FAS
(c)
Determine pitch-tuned HNM parameters
(d)
Unvoiced segments: Synthesize HNM noise signal
(e)
Voiced segment: Synthesize HNM harmonic signal Synthesize HNM noise signal
(f)
Syllable signal

Fig. 5. HNM based scheme for timbre transformation and speech synthesis

Currently, we do not adopt HMM to generate spectral parameters. One reason is the quantity of training sentences collected is not large enough. Therefore, we recorded each of the Mandarin syllables in isolation for analyzing HNM harmonic and noise parameters. In terms of the analyzed HNM parameters for a syllable, the scheme in Fig. 5 can then be followed step by step to synthesize timbre-transformed syllable signal.

3.1. Control points and time-mapping function

Here, "control point" is distinguished from analysis frame. This is because the HNM parameters for a control point located at voiced segment are obtained by interpolating the HNM parameters from two corresponding analysis frames. In synthesizing voiced segment, adjacent control points are always placed 100 sample points (4.5ms) apart. A fixed pace, 100 sample points, is adopted because a more accurate control of spectrum progressing is intended. However, when synthesizing unvoiced segment, we just copy an analysis frame's HNM parameters into its corresponding control point.

To find 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, we have previously studied a method to plan phoneme durations and use phonemes' durations to construct a piece-wise linear time mapping function [10].

3.2. Determine pitch-orignl. HNM parameters

Let the mapped time be t_m and $n=\lfloor t_m \rfloor$ for a voiced control point. Then, the n-th and (n+1)-th analysis frames' HNM parameters are taken to interpolate out the HNM parameters for the control point. Interpolation is done here in a linear manner, and the phase values from the two analysis frames must be unwrapped beforehand [10]. Let $\bar{A}_i$, $\bar{F}_i$, and $\bar{\theta}_i$ denote the amplitude, frequency, and phase of the i-th harmonic partial after interpolation. Then, for example, $\bar{A}_i$ is interpolated as

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

Because the pitch height defined by $\bar{F}_i$ is the original pitch predetermined in recording time, the obtained parameters here are 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. However, the timbre must be kept consistent across the control points that have their pitches tuned. One principle is to keep the spectral envelope of each control point unchanged. Therefore, the spectral envelope of a control point must be estimated from the pitch-original harmonic partials before computing the pitch-tuned harmonic partials' amplitude, frequency, and phase values.

Spectral envelope can be estimated with a global [11] or local approximation method. Here, for directly making use of HNM parameters and computation-efficiency consideration, we adopt a Lagrange interpolation based local approximation. Let $\tilde{F}_k$ and $\tilde{A}_k$ denote the frequency and amplitude of the k-th tuned harmonic partial. To compute the value of $\tilde{A}_k$, 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, $\bar{F}_{j\text{-}1}$, $\bar{F}_j$, $\bar{F}_{j+1}$, and $\bar{F}_{j+2}$, are used to perform order three Lagrange interpolation. That is,

$$\tilde{A}_k = \sum_{m=j-1}^{j+2} \bar{A}_m \cdot \prod_{\substack{h=j-1 \\ h\neq m}}^{j+2} \frac{\tilde{F}_k - \bar{F}_h}{\bar{F}_m - \bar{F}_h} \qquad (2)$$

If Lagrange interpolation of Equation (2) is applied to the partials in the harmonic part of Fig. 1, the estimated spectral envelope would be the curve shown in Fig. 6.

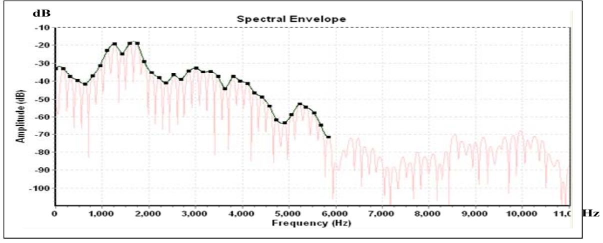

Fig. 6. Spectral envelope by Lagrange interpolation

3.4. Synthesis of speech signal with HNM

For the harmonic signal, $H(t)$, between the n-th and (n+1)-th control points, its sample values are computed with these equations (rewritten by us):

$$H(t) = \sum_{k=0}^{L} a_k^n(t) \cos\left(\phi_k^n(t)\right) \quad , \quad t = 0,1,\ldots,99 \ , \qquad (3)$$

$$a_k^n(t) = \tilde{A}_k^n + \frac{t}{100}(\tilde{A}_k^{n+1} - \tilde{A}_k^n) \ , \qquad (4)$$

$$\phi_k^n(t) = \phi_k^n(t-1) + 2\pi f_k^n(t) / 22{,}050 \ , \ \phi_k^n(0) = \hat{\theta}_k^n \ , \qquad (5)$$

$$f_k^n(t) = \tilde{F}_k^n + \frac{t}{100}(\tilde{F}_k^{n+1} - \tilde{F}_k^n) \ , \qquad (6)$$

where L is number of harmonic partials, 100 is the number of samples between 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, $f_k^n(t)$ is the time-varying frequency for the k-th partial, and $\hat{\theta}_k^n$ is unwrapped phase of $\tilde{\theta}_k^n$ versus $\hat{\theta}_k^{n-1}$. In Equations (4) and (6), linear interpolation is used, which seems enough according to perception tests.

For the noise signal, $N(t)$, we decide to synthesize it as a summation of sinusoidal signal components [6]. Then, the values of the noise-signal samples between the n-th and (n+1)-th control points can be computed as a summation of sinusoidal components whose amplitudes and frequencies are linearly varied as in Equations (4) and (6).

4. Perception tests

Our Mandarin speech synthesis system is developed in previous studies [10, 12]. 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 original HNM based syllable signal synthesis scheme [10] is extended, and an HNM based timbre-transformation and synthesis scheme as shown in Fig. 5 is obtained.

To conduct perception tests, we prepare 6 synthetic speech files beforehand, which are denoted as *AA*, *AB*, *AC*, *AD*, *AX*, and *AY*, respectively. In synthesizing *AA*, *AB*, *AC*, and *AD*, the method, FAS, is used to transform their timbres. The scaling factor, α, is set to 0.9, 0.8, 0.7, and 0.6, respectively. On the other hand, *AX* is synthesized by using PLFM and *AY* is synthesized by using PLFM+FAS, i.e. FAS with scaling factor, 0.9, is executed after executing PLFM. The 6 speech files can be downloaded from the web page, http:// guhy.csie.ntust.edu.tw/TmbrHNM/F2M.html.

Here, 12 persons are invited to participate in the tests. For each person, we allow him to play each of the 6 files again and again. Then, he is asked which timbre of *AA*, *AB*, *AC*, and *AD* is most similar to the timbre of *AX*. Similarly, he is asked which of the four timbres is most similar to that of *AY*. As a result, 11 of the 12 persons recognize that *AB* is most similar to *AX*, and 9 persons recognize that *AC* is most similar to *AY*. Based on these results, we consequently ask each of the participants to compare *AB* with *AX*, and give scores about which is more manful and which is more intelligible. Here, the score, 2 (or -2), is defined as that *AX* is significantly more (or less) manful or intelligible than that of *AB*. If *AX* is just slightly more (or less) manful or intelligible than *AB*, the score, 1 (or -1), is defined. Otherwise, the score, 0, should be given to indicate that they cannot be distinguished. Similarly, we also ask each of the participants to compare *AC* with *AY*, and give scores about timbre-manfulness and intelligibility.

After analyzing the scores given by the participants, we obtain the averaged scores and standard deviations as shown in Table 1. According to the averaged scores, 0.75 and 0.67, it is seen that the transformation method, PLFM, can provide more manfulness 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 less degradation in intelligibility than the method FAS.

Table 1. Averaged scores and standard deviations

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

5. Conclusions

In this paper, we propose a speaker-nonspecific timbre transformation scheme. This scheme is based on the HNM syllable-signal synthesis scheme studied previously. That is, we integrate the two transformation methods, PLFM and FAS, into this scheme. According to the extended scheme, we built a Mandarin speech synthesis and timbre transformation system. By using this system, some speech files are synthesized with timbre transformation, and used to conduct perception tests. The results show that the method PLFM is better than FAS for obtaining more manful timbre. Also, the transformed speech by PLFM is slightly more intelligible than that by FAS.

6. References

[1] K. Tokuda, H. Zen, and A. W. Black, “An HMM-based Speech Synthesis System Applied To English”, IEEE Workshop on Speech Synthesis, Santa Monica, CA, pp. 227-230, 2002.

[2] D. O’Shaughnessy, *Speech Communications: Human and Machine*, IEEE Press, Piscataway, NJ, 2000.

[3] F. R. Moore, *Elements of Computer Music*, Prentice-Hall, Englewood Cliffs, NJ, 1990.

[4] H. Y. Gu and W. L. Shiu, “A Mandarin-Syllable Signal Synthesis Method with Increased Flexibility in Duration, Tone and Timbre Control”, *Proceedings of the National Science Council ROC(A)*, Vol. 22, No. 3, pp. 385-395, 1998.

[5] E. Moulines and E Charpentier, “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones”, *Speech Communication*, Vol. 9, pp. 453-467, Dec. 1990.

[6] Y. Stylianou, “Modeling speech based on harmonic plus noise models”, in *Nonlinear Speech Modeling and Applications*, eds. G. Chollet *et al*., Springer-Verlag, Berlin, pp. 244-260, 2005.

[7] M. Tang, C. Wang, S. Seneff, “Voice Transformations: From Speech Synthesis to Mammalian Vocalizations”, European Conference on Speech Communication and Technology, Aalborg, Denmark, pp. 353-356, 2001.

[8] T. F. Quatieri, *Discrete-Time Speech Signal Processing*, Prentice-Hall, New Jersey, 2002.

[9] X. Serra, “Musical Sound Modeling with Sinusoids plus Noise”, in *Musical Signal Processing*, eds. C. Roads *et al*., Swets & Zeitlinger Publishers, 1997.

[10] H. Y. Gu and Y. Z. Zhou, “An HNM Based Scheme for Synthesizing Mandarin Syllable Signal”, International Journal of Computational Linguistics and Chinese Language Processing, Vol. 13, No. 3, pp. 327-341, 2008.

[11] O. Cappe and E. Moulines, "Regularization Techniques for Discrete Cepstrum Estimation", IEEE Signal Processing Letters, Vol. 3, No. 4, pp. 100-102, April 1996.

[12] H. Y. Gu and C. Y. Wu, “Model Spectrum-progression with DTW and ANN for Speech Synthesis”, accepted by Int. conf. ECTI-CON 2009, Pattaya, Thailand, 2009.