Computational Linguistics and Chinese Language Processing
Vol. 13, No. 3, September 2008, pp. 327-342

An HNM Based Scheme for Synthesizing Mandarin Syllable Signal

Hung-Yan Gu* **and Yan-Zuo Zhou***

Abstract

In this paper, an HNM based scheme is developed to synthesize Mandarin syllable signals. With this scheme, a Mandarin syllable can be recorded just once, and diverse prosodic characteristics can be synthesized for it without suffering significant signal-quality degradation. In our scheme, a synthetic syllable's duration is subdivided to its comprising phonemes and a piece-wise linear mapping function is constructed. With this mapping function, a control point on a synthetic syllable can be mapped to locate its corresponding analysis frames. Then, the analysis frames' HNM parameters are interpolated to obtain the HNM parameters for the control point. Furthermore, for pitch-height adjusting, another timbre-preserving interpolation is performed on the HNM parameters of a control point. Thereafter, signal samples are synthesized according to the HNM synthesis equations rewritten here. This HNM based scheme has been programmed to synthesize Mandarin speech. According to the perception tests, our HNM based scheme is found to be apparently better than a PSOLA based scheme in signal clarity, *i.e.* much clearer and no reverberation.

Keywords: Speech Synthesis, Harmonic-plus-noise Model, Voice Timbre, Pitch Contour.

1. Introduction

Since the introduction of PSOLA (pitch synchronous overlap and add) [Moulines *et al.* 1900], it has been widely used to synthesize speech signal. However, the signal quality of the synthetic speech by PSOLA is not stable. The quality will be degraded a lot if the pitch-contours or durations of the recorded syllables are considerably changed [Dutoit 1997]. Here, signal quality actually means signal clarity, *i.e.* a signal that is less reverberant and less noisy is better in quality. It may be argued that the prosodic characteristics of a syllable need

* Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, 43 Keelung Rd., Sec. 4, Taipei, Taiwan
E-mail: {guhy, M9315058}@mail.ntust.edu.tw

[Received December 28, 2007; Revised August 25, 2008; Accepted August 27, 2008]

only be slightly changed in a corpus-based approach [Chou 1999; Chang 2005]. This argument will hold only if a sufficiently large quantity of speech data is recorded and used. Otherwise, pitch contours between some adjacent syllables may not be smoothly connected and the speaking rate may not be kept constant within a synthetic sentence. Then, pitch-contours and durations will still need to be changed considerably. In addition, the potential for economically transferring a speech synthesis scheme from Mandarin to another language (*e.g.*, Min-nan or Hakka) is an important consideration factor for us. Therefore, we tend not to adopt an expensive approach, such as corpus-based re-sequencing.

Mandarin is a tonal language, and the distinction of the five tones of Mandarin mainly relies on the height and shape of a syllable's pitch-contour. When a signal-model based approach is adopted, the pitch-contour and duration of a syllable inevitably needs considerable change. Thus, the synthesis method, PSOLA, will not be adequate for use, and another suitable technique should be found or developed. Recently, we have found that HNM (harmonic-plus-noise model) is a good base because it can be improved to synthesize Mandarin syllable signals with much higher signal quality.

HNM was proposed by Y. Stylianou to model speech signals to retain high signal quality after such processing as coding and synthesis [Stylianou 1996; Stylianou 2005]. It may be viewed as improving the sinusoidal model [Quatieri 2002] to better model the noise signal components in the higher frequency band of speech signal. 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 model. In contrast, the higher-frequency part is modeled with a smoothed spectrum envelope that is represented with some cepstrum coefficients.

When applying HNM to synthesize Mandarin syllables, we find some issues that are not clearly explained or solved in the literature on HNM. The first issue (not clearly explained) is how to keep the timbre of synthetic syllables consistent, *i.e*. the timbre consistent issue. Note that we intend to record each of the 408 different Mandarin syllables just once then modify the height and shape of a recorded syllable's pitch-contour to that of a different tone's. When the pitch-contour of a syllable to be synthesized is given, the parameter values of the harmonic partials should be adjusted in a way that the timbre can be kept consistent. The second issue is how to determine the HNM parameter values for a control point [Dodge 1997; Moore 1990] placed at the synthetic time axis (of a synthetic syllable), *i.e.* the parameter determination issue. In speech synthesis, one must adjust a recorded syllable's duration to meet the duration requirement given by the prosodic parameter generation unit. When a control point at the synthetic time axis is mapped to a time point between two analysis frames of a recorded syllable, some method of interpolation is needed to determine the HNM parameter values for the control point. In addition, the third issue is how to warp the time axis of a synthetic

syllable in order that more fluent syllables and sentences can be synthesized, *i.e*. the time warping issue. This issue is more relevant to speech synthesis than HNM. When a syllable's duration needs to be lengthened or shortened, a simple time warping method, *i.e*. linear warping, will usually result in lower perceived fluency.

In this paper, the three issues mentioned above are investigated, and equations for signal synthesis with HNM are rewritten in a clearer notation. In addition, a system based on the extensions and rewritten equations for HNM signal synthesis is developed to synthesize Mandarin syllable signal. The main processing flow of the system is drawn in Figure 1. When a syllable's signal is to be synthesized, its prosodic parameters' values are readily determined by the prosody unit. Hence, in the first block of Figure 1, a synthetic syllable's time length can be planned and subdivided to its comprising phonemes. For example, the syllable /man/ has three phonemes, /m/, /a/, and /n/. Then, a piece-wise linear time mapping function is constructed to map the synthetic phonemes to their corresponding phonemes in the recorded syllables. In the second block of Figure 1, control points are uniformly placed on the synthetic time axis. Then, HNM parameters' values for each control point are determined. In the following blocks, three types of signals are classified and synthesized separately. Here, the signal of a short unvoiced syllable-initial is directly copied from the recorded to the synthesized. The signal of a long unvoiced syllable-initial is synthesized as noise signal components in HNM while the signals of voiced initial and syllable-final are synthesized as the sum of both the harmonic and noise signal components.

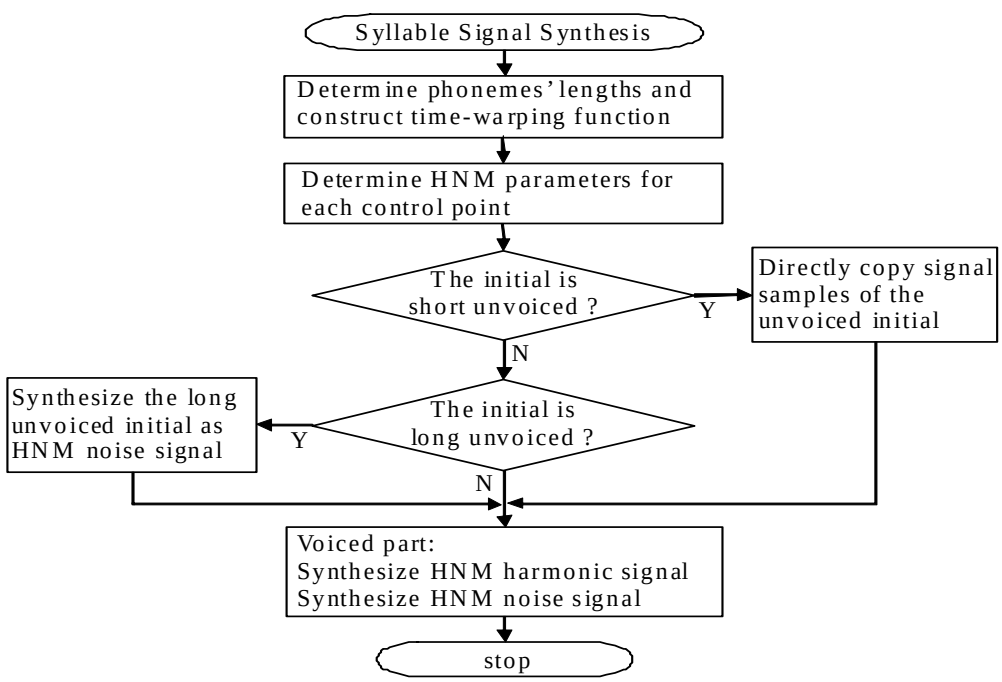

Figure 1. Main processing flow of the HNM based syllable-signal synthesis scheme.

2. Phoneme Duration Planning and Time Axis Mapping

The issues of duration planning and time-axis mapping are not mentioned in the literature on HNM [Stylianou 1996; Stylianou 2005]. Mandarin syllables have the structure, C_xVC_n. The component, C_x, may be null, a voiced consonant, or an unvoiced consonant while the component, C_n, may be null or a nasal /n/ or /ng/. Also, the component, *V*, may be a vowel, diphthong, or triphthong. When C_x is an unvoiced consonant, we classify it as a short-unvoiced (*e.g*. /b/, not aspirated) or long-unvoiced (*e.g*. /p/, aspirated). For a short-unvoiced, its signal will be directly copied from the initial part of the recorded syllable to the initial part of the synthetic syllable. This processing is indicated in the block at the right side of Figure 1. However, for a long-unvoiced, its signal will be synthesized as the sum of noise signal components with HNM. This processing is indicated in the block at the left side of Figure 1. In addition, C_x is a voiced consonant or null, and it will be synthesized together with the syllable final, VC_n, as the sum of both the harmonic and noise signal components.

When a syllable is started with a short-unvoiced consonant, *e.g*. /bau/, the time length of the consonant is planned as the corresponding consonant's length in the recorded syllable. In contrast, when started with a long-unvoiced consonant, the length of the consonant is planned by multiplying its original length with a factor, *Fu*. The value of *Fu* is first computed as the synthetic syllable's length divided by its corresponding recorded syllable's length. However, the value, *Fu*, is restricted to the range from 0.6 to 1.4, *i.e*. set to 1.4 when larger than 1.4 and set to 0.6 when smaller than 0.6. After the length of the unvoiced part, *Du*, is determined, the length of the voiced part, *Dv*, is apparently the synthetic syllable's length minus *Du*.

To plan the lengths of the phonemes within the voiced part, consider the example syllable, /man/. Suppose that in the recorded signal of /man/, the three phonemes, /m/, /a/, and /n/, occupy *Rm*, *Ra*, and *Rn* seconds, respectively, and $Rv = Rm + Ra + Rn$. Also, suppose that *Dm*, *Da*, and *Dn* represent the time lengths of the three phonemes within the synthetic syllable, and $Dv = Dm + Da + Dn$. Note that *Dm* (or *Rm*) is used here to denote the time length of the initial voiced consonant of a syllable, *Da* (or *Ra*) denotes the time length of the vowel nucleus, and *Dn* (or *Rn*) denotes the time length of the final nasal consonant. In this study, the values of *Dm*, *Da*, and *Dn* are planned according to an observation. That is, the consonant-to-vowel duration ratio, $(Rm + Rn) / Rv$, will become smaller when the syllable is uttered within a sentence instead of uttered in isolation. The planning procedure is as below.

```
r = 0.85;
while ( r >= 0.1 ) {
    Dm = (Rm/Rv) * r * Dv;
    Dn = (Rn/Rv) * r * Dv;
    Da = Dv – Dm – Dn;
    if (Da > Dv*0.5) break;
    r = r – 0.05;
}
Db = Dm + Dn;
if (Dm > 0 && Dm/Db < 0.35) { Dm = 0.35*Db;  Dn=Db-Dm; }
if (Dn > 0 && Dn/Db < 0.35) { Dn = 0.35*Db;  Dm=Db-Dn; }
```

In this procedure, the value of *Dm* is planned by multiplying a duration reduction rate, *r*, with the time ratio (*Rm* / *Rv*) of its counterpart, *Rm*, in the recorded syllable. In the same way, the value of *Dn* is planned. By trying to decrease the value of *r* iteratively, the values of *Dm* and *Dn* are decreased gradually, and the value of *Da* finally becomes sufficiently large. As to the initial value of *r*, *i.e.* 0.85, and the vowel duration threshold, *i.e.* 0.5, they are set according to analyzing some real spoken sentences.

If the structure of a syllable is same as /san/ or /an/, *i.e.* without voiced initial consonant, then the values of *Rm* and *Dm* can be set to zero directly. Similarly, if the structure of a syllable is the same /ma/, *i.e.* without an ending nasal, then the values of *Rn* and *Dn* can be set to zero directly. After the values of *Dm*, *Da*, and *Dn* are determined, a mapping function from the phonemes in the synthetic syllable to their corresponding phonemes in the recorded syllable can be established and used in the second block of Figure 1. The mapping function adopted here is as depicted in Figure 2. That is, it is a piece-wise linear function. Although a simple mapping function is adopted here, we think the fluency level of the synthetic speech can still be improved a lot. In the future, we will study the mapping problem between the source and synthetic syllables with a more systematic method.

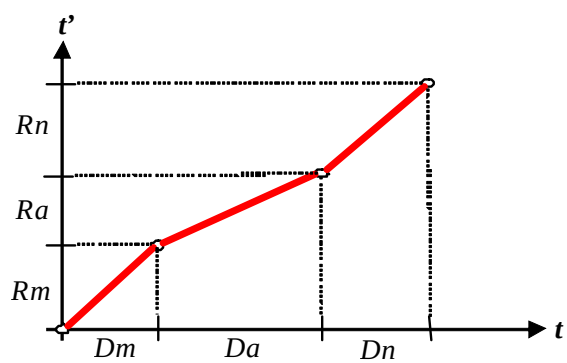

Figure 2. A piece-wise linear mapping function.

3. Control Points and HNM Parameter Determination

3.1 Control Point Placement

In this paper, the source syllables are recorded at a sampling rate of 22,050Hz. In analyzing HNM parameters, frame size is set to 512 sample points (23.2ms), and frame shift is set to 256 sample points. However, in signal synthesis processing, the concept of a "control point" is adopted, which is commonly used in computer music synthesis [Dodge 1997; Moore 1990]. The term "control point" is used instead of "frame" because, in our scheme, the HNM parameters for a control point located at voiced part are obtained by interpolating the parameters from two corresponding analysis frames, *i.e.* not directly copying parameters from a frame into a control point (note that original HNM uses only direct copying). However, in synthesizing a long-unvoiced part, the HNM parameters of an analysis frame located at the unvoiced part are directly copied and used for a control point corresponding to it. These different manners of HNM parameter determination for voiced and long-unvoiced parts are illustrated in Figure 3.

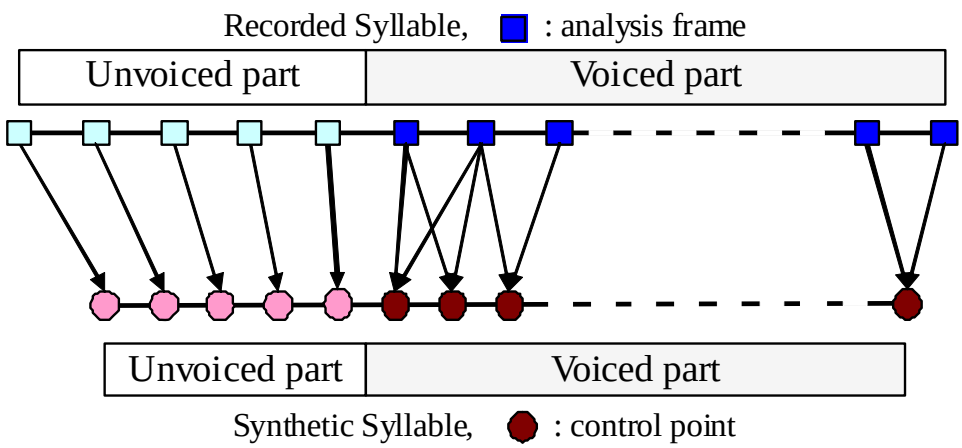

Figure 3. Analysis frame to control point mapping.

From Figure 3, it can be seen that the number of control points in the synthetic unvoiced part is same as the number of analysis frames in the recorded unvoiced part. Hence, the time axis is simply linearly shortened or lengthened. However, in the synthetic voiced part, adjacent control points are always placed 100 sample points (4.5ms) apart. Thus, the number of control points depends on the time length planned. Here, a fixed pace, 100 sample points, is adopted because an accurate control of spectrum progressing within the synthetic voiced part is intended.

3.2 Pitch-original HNM Parameters

To determine the HNM parameter values for a control point within the synthetic voiced part, the first step is to do time-position mapping according to the constructed mapping function as shown in Figure 2. Suppose the control point's time position, t_s, on the synthetic time axis is mapped to t_r on the recorded-syllable time axis. Then, we use the HNM parameters analyzed from the two frames numbered $\lfloor t_r \rfloor$ and $\lfloor t_r \rfloor+1$ to interpolate out HNM parameters for the control point. Currently, we do the interpolation in a linear way. That is:

$$\bar{A}_i = (1-w)\cdot A_i^n + w\cdot A_i^{n+1},\ n = \lfloor t_r \rfloor,\ w = t_r - n \tag{1}$$

$$\bar{F}_i = (1-w)\cdot F_i^n + w\cdot F_i^{n+1} \tag{2}$$

$$\bar{\theta}_i = w\cdot(\hat{\theta}_i^{n+1} - \theta_i^n) + \theta_i^n \tag{3}$$

where A_i^n, F_i^n, and θ_i^n denote the amplitude, frequency, and phase of the *i*-th harmonic partial in the *n*-th analysis frame, and $\bar{A}_i$, $\bar{F}_i$, and $\bar{\theta}_i$ denote the amplitude, frequency, and phase of the *i*-th harmonic partial for the control point. Note that in Equation (3), $\hat{\theta}_i^{n+1}$ represents the unwrapped phase of θ_i^{n+1} versus θ_i^n, i.e. $\hat{\theta}_i^{n+1} = puw(\theta_i^{n+1}, \theta_i^n)$. The phase θ_i^{n+1} is unwrapped in order that the phase difference is within the range from -π to π. Here, our modified phase unwrapping is done as:

$$\hat{\theta}_i^{n+1} = puw(\theta_i^{n+1}, \theta_i^n) = \theta_i^{n+1} - M\cdot 2\pi \tag{4}$$

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

In original HNM, the noise signal components are represented with 10 cepstrum coefficients. Therefore, for each control point, 10 cepstrum coefficients should be derived. Here, the cepstrum coefficients from the two mapped analysis frames are linearly interpolated to derive the cepstrum coefficients for the control point.

3.3 Pitch-tuned HNM Parameters

On a control point, after the parameters, $\bar{A}_i$, $\bar{F}_i$, and $\bar{\theta}_i$, for pitch-original harmonic partials are computed, the parameters, $\tilde{A}_k$, $\tilde{F}_k$, and $\tilde{\theta}_k$, for pitch-tuned harmonic partials should be computed next. Note that the pitch-height defined by $\bar{F}_i$ is the original pitch predetermined in recording time. Thus, the pitch-height of a control point must be tuned in order to follow the pitch contour given by the prosody unit. For example, let the pitch defined by the harmonic frequencies, $\bar{F}_i$, be 100Hz, and a pitch-height of 150Hz is needed according to the assigned pitch-contour. Apparently, a simple tuning method is to set the values of $\tilde{A}_k$, $\tilde{F}_k$, and $\tilde{\theta}_k$ as $\tilde{F}_k = \bar{F}_k \cdot 150/100$, $\tilde{A}_k = \bar{A}_k$, and $\tilde{\theta}_k = \bar{\theta}_k$. This is illustrated in Figure 4. From this figure, it can be seen that the pitch can indeed be tuned from 100Hz to 150Hz. However,

the formant frequencies are also scaled up. For example, the first formant is shifted from 240Hz to 360Hz in Figure 4. The shifting of formant frequencies will cause the timbre be distinctly changed. As a result, the timbre of a synthetic syllable will not be consistent and will vary with the scaling factors (*e.g.* 150/100) set for different control points.

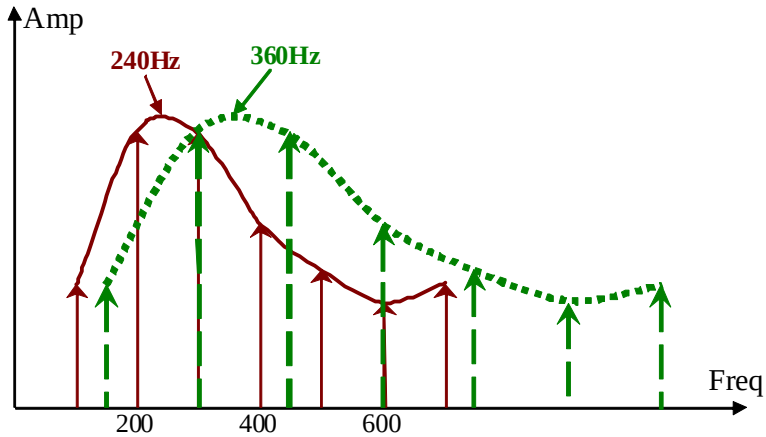

Figure 4. Pitch tuning with spectral envelope scaled simultaneously.

To preserve the timbre while tuning the pitch-height of a control point, one principle is to keep the spectral envelope unchanged [Dodge 1997]. This implies that the amplitude $\tilde{A}_k$ of the pitch-tuned harmonic partial located at frequency $\tilde{F}_k$ must be computed according to an estimated spectral envelope. Here, considering both factors of efficient processing and sufficient accuracy, we estimate the spectral envelope by Lagrange interpolating the sequence of pairs, ($\bar{F}_i$, $\bar{A}_i$). In details, for the *k*-th harmonic frequency $\tilde{F}_k$, 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 to compute the value of $\tilde{A}_k$. That is:

$$\tilde{A}_k = \sum_{m=j-1}^{j+2} \bar{A}_m \cdot \prod_{\substack{h=j-1 \\ h \neq m}}^{j+2} \frac{\tilde{F}_k - \bar{F}_h}{\bar{F}_m - \bar{F}_h} \tag{5}$$

A figure that illustrates this method of pitch tuning without changing spectral envelope is shown in Figure 5. In this figure, the pitch is scaled up by a factor of 1.25 but the timbre is preserved. Similarly, the phase $\tilde{\theta}_k$ of the pitch-tuned harmonic partial located at frequency $\tilde{F}_k$ can also be interpolated with the four pitch-original partials of frequencies, $\bar{F}_{j-1}$, $\bar{F}_j$, $\bar{F}_{j+1}$, and $\bar{F}_{j+2}$. However, the phases of the four partials, $\bar{\theta}_{j-1}$, $\bar{\theta}_j$, $\bar{\theta}_{j+1}$, and $\bar{\theta}_{j+2}$, must be unwrapped beforehand to prevent phase discontinuities. That is, the unwrapped phases, $\hat{\theta}_{j-1} = \bar{\theta}_{j-1}$, $\hat{\theta}_j = puw(\bar{\theta}_j, \hat{\theta}_{j-1})$, $\hat{\theta}_{j+1} = puw(\bar{\theta}_{j+1}, \hat{\theta}_j)$, and $\hat{\theta}_{j+2} = puw(\bar{\theta}_{j+2}, \hat{\theta}_{j+1})$, are used instead in the interpolation processing.

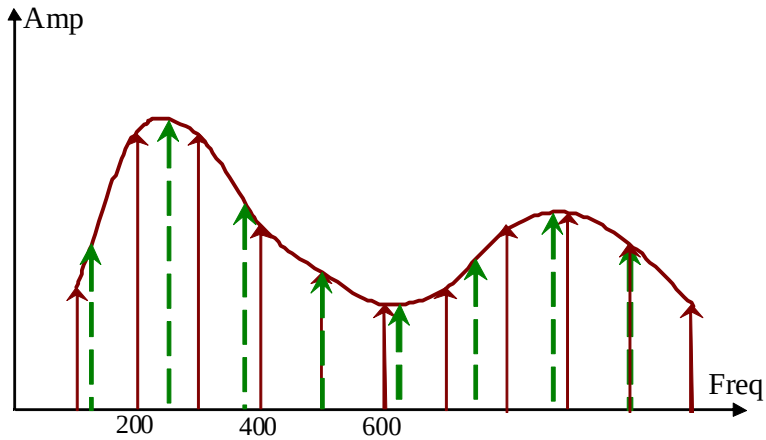

Figure 5. Pitch tuning without changing spectral envelope.

4. Signal Waveform Synthesis

For the synthetic voiced part of Figure 3, the synthetic signal, $S(t)$, consists of harmonic and noise components. That is, $S(t) = H(t) + N(t)$ where $H(t)$ represents the summation of the harmonic partials and $N(t)$ represents the summation of the noise signal components. The synthesis methods for $H(t)$ and $N(t)$ are described in details in the following subsections.

4.1 Harmonic Signal Synthesis

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):

$$H(t) = \sum_{k=0}^{L} a_k^n(t)\cos\left(\phi_k^n(t)\right),\ t = 0,1,\ldots,99, \tag{6}$$

$$a_k^n(t) = \tilde{A}_k^n + \frac{t}{100}\left(\tilde{A}_k^{n+1} - \tilde{A}_k^n\right), \tag{7}$$

$$\phi_k^n(t) = \phi_k^n(t-1) + 2\pi f_k^n(t)/22{,}050,\ \phi_k^n(0) = \hat{\theta}_k^n, \tag{8}$$

$$f_k^n(t) = \tilde{F}_k^n + \frac{t}{100}(\tilde{F}_k^{n+1} - \tilde{F}_k^n), \tag{9}$$

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 $\hat{\theta}_k^n = puw(\tilde{\theta}_k^n, \hat{\theta}_k^{n-1})$, *i.e.* unwrapped phase of $\tilde{\theta}_k^n$ versus $\hat{\theta}_k^{n-1}$. In Equations (7) and (9), linear interpolation is used, which seems enough according to perception tests.

Note that, when using Equation (6) 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=100. These kinds of discontinuities, *i.e.* $\phi_k^n(100) \neq \phi_k^{n+1}(0)$, will induce amplitude discontinuities to signal waveform, and cause clicks to be heard. To prevent these kinds of discontinuities, the amount of mismatched phase, ξ_k^n, at the boundary point, t=100, must be computed beforehand. Then, this amount can be divided and shared among the 100 sample points between two adjacent control points. Accordingly, the phases of the signal samples (especially those around the boundary point) will advance smoothly. Here, we compute the amount of mismatched phase as:

$$\xi_k^n = puw\left(\phi_k^n(100), \phi_k^{n+1}(0)\right) - \phi_k^{n+1}(0) \tag{10}$$

where the phase unwrapping function, $puw(x, y)$, is as defined in Equation (4), and according to our derivation $\phi_k^n(100)$ can be directly computed as

$$\phi_k^n(100) = \phi_k^n(0) + \frac{\pi}{22,050}\left(101\tilde{F}_k^{n+1} + 99\tilde{F}_k^n\right) \tag{11}$$

The formula in Equation (11) is obtained by recursively evaluating Equations (8) and (9). Then, by dividing and sharing ξ_k^n to the samples between two control points, Equation (6) is modified to:

$$H'(t) = \sum_{k=0}^{L} a_k^n(t)\cos\left(\phi_k^n(t) - \frac{t}{100}\cdot\xi_k^n\right),\ t = 0,1,\ldots,99, \tag{12}$$

Let L_n be the number of harmonic partials on the n-th control point. The value of L_n is computed as dividing the MVF by the pitch frequency, *i.e.* L_n = MVF(n) / $\tilde{F}_1^n$. In general, L_n may not be equal to L_{n+1}. Hence, we set the value of L, *i.e.* the number of partials, in Equations (6) and (12) to the greater of L_n and L_{n+1}. Suppose here that L_n is less than L_{n+1}. Then, the parameter values for the extended partials on the n-th control point must be defined. Here, from the consideration of signal-waveform continuity, we simply let $\tilde{A}_k^n$ =0, $\tilde{F}_k^n = \tilde{F}_k^{n+1}$, $\tilde{\theta}_k^n = \tilde{\theta}_k^{n+1}$, for k = 1+L_n, 2+L_n, …, L_{n+1}.

4.2 Noise Signal Synthesis

For the noise signal, $N(t)$, we decide to synthesize it as a summation of sinusoidal signal components [Stylianou 1996]. Let G_k be the frequency of the k-th sinusoid. As G_k does not change with time, we need not to distinguish G_k for different control points. Here, we let G_k =100·k (Hz). However, for the n-th control point, the index k of G_k is not started from 1 and its starting value, K_s^n, is determined by the MVF of this control point, *i.e.* $K_s^n = \lceil \text{MVF(n)} / 100 \rceil$. In contrast, the end value of the index k is always a fixed value, $K_e = \lfloor 11{,}025 / 100 \rfloor$, because G_k cannot be greater than half of the sampling frequency.

On the other hand, let B_k^n be the amplitude of the k-th sinusoid on the n-th control point.

To determine the values of B_k^n, the 10 cepstrum coefficients, of the *n*-th control point, representing the noise spectral envelope are first appended with zero values and inversely transformed (inverse discrete Fourier transform) to the spectral domain. Then exponentiation is taken to obtain the corresponding spectral magnitude coefficients, X_j, j=0,1,…,2047. According to X_j, the value of B_k^n can be obtained by linearly interpolating the two adjacent X_i whose frequencies indicated by the index, *i*, surround the frequency of G_k.

When the values of K_s^n and B_k^n for the *n*-th control point are known, the values of the noise-signal samples between the *n*-th and (*n*+1)-th control points can be computed as:

$$N(t) = \sum_{k=K_s}^{K_e} b_k^n(t)\cos\left(\gamma_k^n + t \cdot 2\pi G_k / 22,050\right),\ t = 0,1,\ldots,99, \tag{13}$$

$$b_k^n(t) = B_k^n + \frac{t}{100}(B_k^{n+1} - B_k^n), \tag{14}$$

$$\gamma_k^n = \gamma_k^{n-1} + 100 \cdot 2\pi G_k / 22,050\ , \tag{15}$$

where K_s is set to the lesser of K_s^n and 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 (14), the time-varying amplitude, $b_k^n(t)$, is only linearly interpolated.

For the synthesis of the long unvoiced part in Figures 1 and 3, the Equations (13), (14) and (15) can still be used to generate signal samples. However, the lower bound of the summation index, *k*, in Equation (13) will now be fixed to 1. This is equivalent to setting all the MVF values to the constant, 0Hz, for all the control points within the unvoiced part.

5. Signal Synthesis Experiment and Perception Test

Several years ago, we proposed a synthesis method called TIPW (time-proportioned interpolation of pitch waveform) [Gu *et al.* 1998] that is an improved variant of PSOLA. Therefore, we intend to compare the three synthesis methods, based on PSOLA, TIPW, and HNM respectively, in signal clarity. A synthetic speech signal is considered to have better clarity if it is less noisy and less reverberant. Since signal clarity is the primary concern here, we use the same text analysis unit and prosody parameter generation unit for the three methods [Gu *et al.* 2000; Gu *et al.* 2007]. When run on a personal computer with an Intel Pentium 2.6 GHz CPU, the three methods can all be executed in real-time. However, the execution speeds are very different. In detail, the CPU time consumed by the HNM based method is 19.4% of the time length of the synthetic speech file, *i.e.* the speed is about 5 times real-time. On the other hand, the CPU time consumed by the TIPW and PSOLA based methods are as little as 3.5% and 4.2% of the time length of the synthetic speech file, *i.e.* the speeds are about 28 and 24 times real-time.

For comparison of signal clarity, the Mandarin short sentence, /syuen-2 zhuan-3 li-4/

(旋轉力, rotating power), is taken as an example and used to synthesize speech signals with the three methods. The spectrogram in Figure 6 is obtained by analyzing the signal synthesized by the HNM based method while the spectrograms in Figures 7 and 8 are obtained respectively by analyzing the signals synthesized by the TIPW and PSOLA based methods. By comparing Figure 6 with Figures 7 and 8, we find that more fragments exist in Figures 7 and 8 than in Figure 6, and the traces of the harmonic partials in the lower frequency band in Figure 6 are more continuous and concrete (less vibrating) than those in Figures 7 and 8. Therefore, the signal synthesized by the HNM based method should be clearer than the signal synthesized by the TIPW and PSOLA based methods.

In addition, we have used the three methods to synthesize an article and obtain three speech signal files. The article selected is a simple composition, with a total of 132 syllables, by an elementary school student. Then, the three synthetic speech files are played to 15 participants for perception tests. A score of 0 is defined if the clarity of two compared synthetic speech files cannot be distinguished. A score of 1 (or -1) is defined if the latter (or former) played is slightly better. Besides, a score of 2 (or -2) is defined if the latter (or former) played is sufficiently better. Each participant is requested to do two comparisons and give two scores. One comparison is to compare the signal clarity of the two files synthesized respectively by PSOLA and HNM based methods. And the other comparison is to compare the two files synthesized respectively by PSOLA and TIPW based methods. According to the scores given by the participants, the averaged scores are 1.2 for the first comparison and 0.33 for the second comparison. That is, the HNM based method is significantly better than the PSOLA based method in signal clarity, but the two methods, PSOLA and TIPW, are difficult to distinguish. For demonstration, we have set up a web page, http://guhy.csie.ntust.edu.tw/hmtts/hnm-demo.html. It can be browsed to listen to the synthesized Mandarin speeches using the three methods.

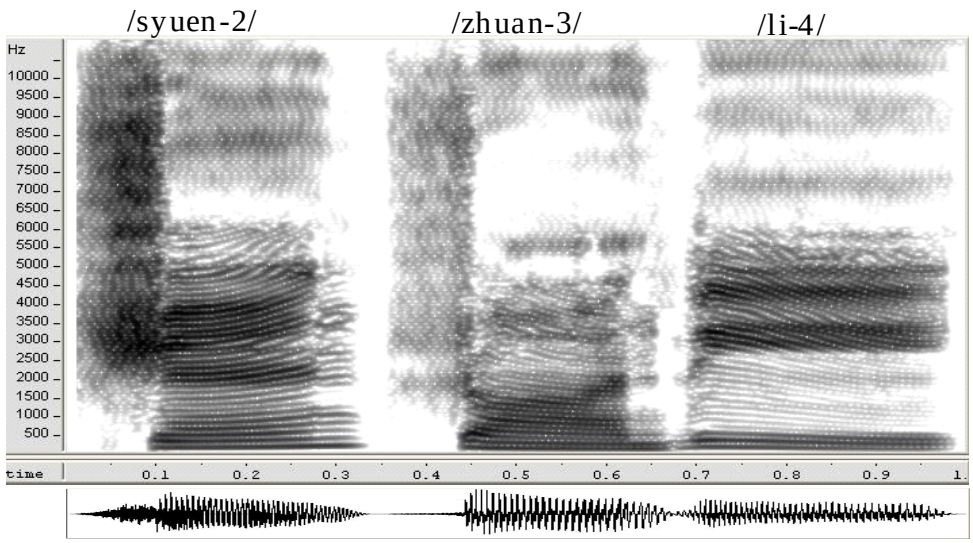

Figure 6. Spectrogram of the signal synthesized by the HNM based method.

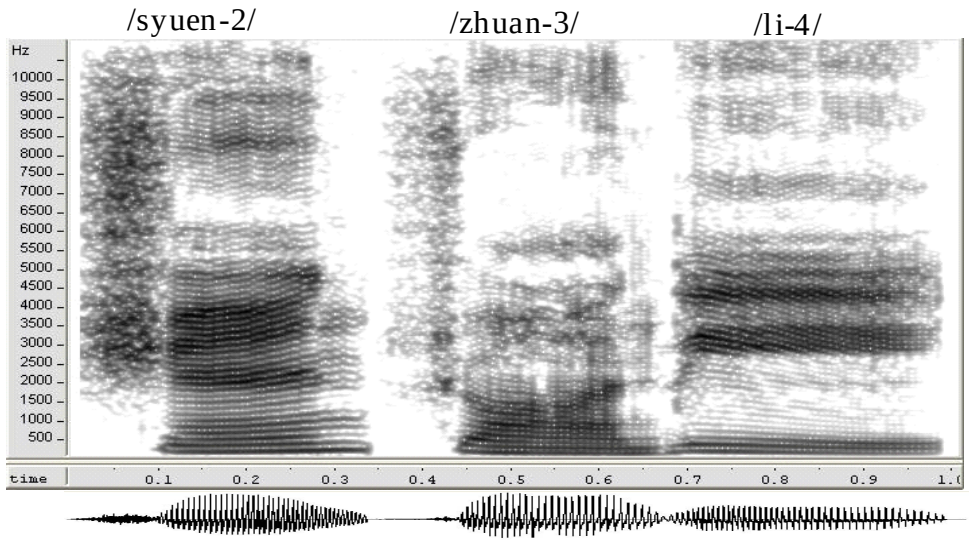

Figure 7. Spectrogram of the signal synthesized by the TIPW based method.

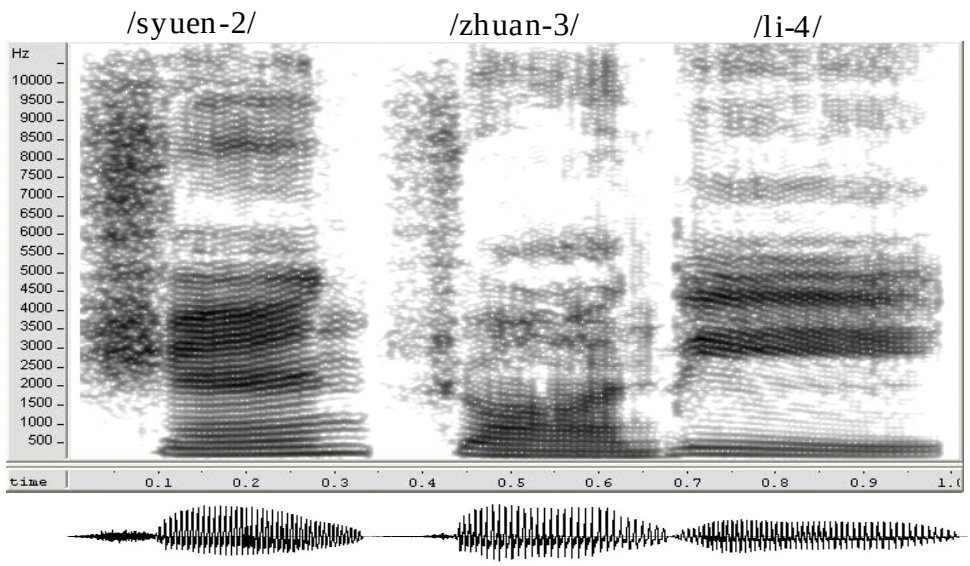

Figure 8. Spectrogram of the signal synthesized by the PSOLA based method.

6. Concluding Remarks

In this study, we used HNM to develop a scheme for synthesizing a Mandarin syllable's signal. Each Mandarin syllable needs only be recorded once. With this scheme, diverse prosodic characteristics can still be synthesized for a syllable without suffering significant signal-quality degradation. Three relevant issues are investigated. That is, (a) the

determination of the HNM parameters for the control points that are placed with a fixed pace on the time axis of a synthetic syllable (note that pace widths are varied in original HNM); (b) keeping timbre consistent when the HNM parameters of a control point are adjusted to have a different pitch height (implementation method is not clearly explained in original HNM); (c) the construction of a time warping function to map between the two time axes of a synthetic syllable and its corresponding source syllable in order to synthesize more fluent syllable signal (this issue is not mentioned in original HNM). For these three issues, we have proposed feasible solutions (considering both signal quality and implementation practice). With these solutions, our scheme is therefore called an HNM based and extended syllable signal synthesis scheme (HNMES).

To test the signal clarity of the synthetic speech, the HNMES scheme has been programmed and integrated with the other units of text analysis and prosodic parameter generation that were developed earlier. Since signal clarity is the primary concern here, the same units of text analysis and prosodic parameter generation are also used for the PSOLA and TIPW based methods. According to spectrogram inspecting and perception test results, we conclude that the HNMES scheme can significantly outperform the PSOLA and TIPW based schemes in signal clarity (much clearer and no reverberation). Therefore, the HNMES scheme is recommended for synthesizing speech signal of not only Mandarin but also other syllable-prominent languages.

Note that in this study, signal clarity is the primary concern and prosodic parameters are generated with only simple rules. Therefore, the synthetic speeches are not natural and felt of a machine tongue when listening to the example synthetic speeches. In the future, we will study to construct a more powerful prosodic parameter generation unit. Then, we will combine it with the syllable signal synthesis scheme, HNMES.

Acknowledgments

This study is partially supported by National Science Council under the contract number, NSC 96-2221-E-011-163.

Reference

Chang, T. Y., *A Mandarin Text-to-speech System Using a Large Number of Words as Synthesis Units*, Master thesis, National Chung Hsing University, Taichung, Taiwan, 2005. (in Chinese)

Chou, F. C., *Corpus-based Technologies for Chinese Text-to-Speech Synthesis*. PhD thesis, National Taiwan University, Taipei, Taiwan, 1999.

Dodge, C., and T. A. Jerse, *Computer Music: Synthesis, Composition, and Performance*, 2nd edition, Schirmer Books, New York, 1997.

Dutoit, T., *An Introduction to Text-to-Speech Synthesis*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997.

Gu, H. Y., 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)*, 22(3), 1998, pp. 385-395.

Gu, H. Y., and C. C. Yang, "A Sentence-Pitch-Contour Generation Method Using VQ/HMM for Mandarin Text-to-speech," *International Symposium on Chinese Spoken Language Processing*, 2000, Beijing, China, pp. 125-128.

Gu, H. Y., Y. Z. Zhou, and H. L. Liau, "A System Framework for Integrated Synthesis of Mandarin, Min-nan, and Hakka Speech," *International Journal of Computational Linguistics and Chinese Language Processing*, 12(4), 2007, pp. 371-390.

Moore, F. R., *Elements of Computer Music*, Prentice-Hall, New Jersey, 1990.

Moulines, E., and E Charpentier, "Pitch-synchronous Waveform Processing Techniques for Text-to-speech Synthesis Using Diphones," *Speech Communication*, 9(5), 1990, pp. 453-467.

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

Stylianou, Y., *Harmonic plus Noise Models for Speech, Combined with Statistical Methods, for Speech and Speaker Modification*, PhD thesis, Ecole Nationale Supèrieure des Télécommunications, Paris, France, 1996.

Stylianou, Y., "Modeling Speech Based on Harmonic Plus Noise Models," *Nonlinear Speech Modeling and Applications*, Springer-Verlag, Germany, 2005.