Mandarin Syllable Signal Synthesis Using an HNM Based Scheme

Hung-Yan Gu and Yen-Zuo Zhou
Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology, Taipei, Taiwan
E-mail: guhy@mail.ntust.edu.tw

Abstract

In this paper, HNM (harmonic plus noise model) is enhanced and used to design a scheme for synthesizing Mandarin syllable signals. Each syllable is recorded once only and used to synthesize syllable signals with diverse prosodic characteristics without suffering significant signal-quality degradation. For a control point on the synthetic syllable's time axis, two corresponding analysis frames' HNM parameters are interpolated to derive the HNM parameters for the control point. Furthermore, for pitch-contour tuning, another timbre-reserving interpolation is performed for the HNM parameters on a control point. Then, signal samples are synthesized with the HNM synthesis equations rewritten here. According to the result of the perception tests, the HNM based scheme proposed here can indeed be used to synthesize syllable signals with consistent timbre and high signal clarity.

1. Introduction

Since the introducing of PSOLA (pitch synchronous overlap and add) [1], 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 speech units are considerably changed [2]. Note that considerably changing the height and contour shape of a speech unit's pitch-contour is required when a signal-model based (i.e. not corpus-based) approach is adopted. Therefore, we started to search for a signal synthesis technique to replace PSOLA. Recently, we find that HNM (harmonic plus noise model) is a good base because it can provide much higher signal quality and can be further enhanced to support the synthesis of Mandarin syllables.

HNM is proposed by Y. Stylianou to model speech signal to obtain higher signal quality [3, 4]. It may be viewed as improving the sinusoidal model [5] to better model the noise signal components in the higher frequency band of speech signal. In HNM, a MVF (maximum voiced frequency) detection method is provided to divides a speech frame's spectrum into lower and higher frequency parts. The lower-frequency part is modeled as the sum of harmonic partials as in sinusoidal model. And the higher-frequency part is modeled with a smoothed spectrum envelope that is represented with some cepstrum coefficients.

Many languages are syllable prominent, e.g. Chinese, Japanese, Korean, etc. The structure of a syllable is assumed to be, C_xVC_n, here. The initial, C_x, may be null, a voiced consonant, or an unvoiced consonant while the final, C_n, may be null or a nasal as /n/ or /ng/. Besides, the nucleus, V, may be a vowel, diphthong, or triphthong. If HNM is to be applied to synthesize syllable signals of these languages, some issues may be found that are not clearly explained or solved in the literature on HNM.

The first issue is how to keep the timbre of synthetic syllables consistent. The detailed implementation method is not given and explained in original HNM [3, 4]. Note that we intend to record each syllable's signal once only, and then modify the prosodic characteristics of the syllable to satisfy the requests from the prosody unit. Therefore, to realize a given pitch-contour, the parameter values of a syllable's 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 [6, 7] placed at the synthetic time axis (of a synthetic syllable). It is needed to adjust a recorded syllable's duration to satisfy the duration request from the prosody unit. When a control point at the synthetic time axis is mapped to a time point between two analysis frames of the recorded syllable, then some way of interpolation is needed to determine the HNM parameter values for the control point.

In this paper, the two issues mentioned above are studied. Then, a syllable signal synthesis scheme is designed. The main processing flow of this scheme is drawn in Fig. 1. When a syllable's signal is to be synthesized, its prosodic parameters' values are ready provided by the prosody unit. Hence, in the first block of Fig. 1, control points are uniformly placed on the time axis of a synthetic syllable first. Then, HNM parameters' values for each control point are determined. In the following blocks, three types of signals are classified and synthesized separately. Here, when C_x a short unvoiced (e.g. /b/), its signal is directly copied from the recorded to the synthesized. But when C_x a long unvoiced (e.g. /p/), its signal is synthesized as noise signal components in HNM. Otherwise, when C_x is a voiced consonant or null, we treat the initial part of C_x or V (when C_x is null) as starting with a short-unvoiced. Then, C_x is considered together with the remaining parts, VC_n, and their signals are synthesized as the sum of both the harmonic and noise signal components.

When a syllable is started with a short unvoiced, the time length of the short unvoiced is planned as the corresponding segment length in the recorded syllable. But when started with a long unvoiced, the length of the long unvoiced is planned by multiplying a proportion factor *Fu*. But the value of *Fu* is restricted to within the range from 0.6 to 1.4. After the length of the unvoiced part is determined, the length of the voiced part can then be determined.

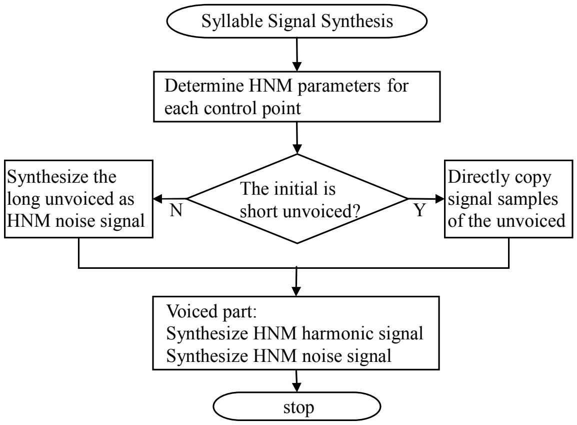
Syllable Signal Synthesis
Determine HNM parameters for each control point
Synthesize the long unvoiced as HNM noise signal
N
The initial is short unvoiced?
Y
Directly copy signal samples of the unvoiced
Voiced part:
Synthesize HNM harmonic signal
Synthesize HNM noise signal
stop

Fig. 1 Main flow of the HNM based synthesis scheme

2. HNM Parameter Determination

2.1 Control Point Placement

In this paper, the source syllables are recorded with the sampling rate, 22,050Hz. In analyzing HNM parameters, frame size is set to 512 sample points (23.2ms), and frame shift is 256 sample points. However, in signal synthesis processing, the concept of a control point is adopted, which is commonly adopted in computer music synthesis [6, 7]. The term "control point" is used instead of "frame" because the HNM parameter for a control point located at voiced part are obtained by interpolating the parameters from two corresponding analysis frames. Also, adjacent control points are always placed 100 sample points (4.5ms) apart. But in synthesizing 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 Fig. 2.

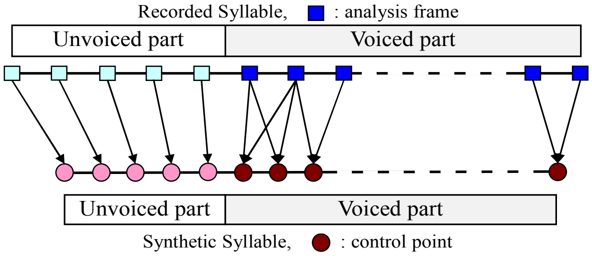

Fig. 2 Analysis frame to control point mapping

2.2 Pitch-original HNM parameters

Suppose a control point's time position, t_s, on the synthetic time axis is mapped to t_r on the recorded-syllable time axis. Then, the HNM parameters analyzed from the two frames numbered $\lfloor t_r \rfloor$ and $\lfloor t_r \rfloor +1$ are used to interpolate out HNM parameters for the control point. Here, we study to perform the interpolation in a linear way. That is,

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

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

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

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, phase unwrapping is done as

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

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

In original HNM, noise signal's spectrum envelope is 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.

2.3 Pitch-tuned HNM Parameters

On a control point, the parameters, $\tilde{A}_k$, $\tilde{F}_k$, and $\tilde{\theta}_k$, of pitch-tuned harmonic partials should be determined 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. Suppose the pitch defined by the harmonic frequencies, $\bar{F}_i$, is 100Hz, and a pitch-height of 150Hz is required. Apparently, a simple way to define the values of $\tilde{A}_k$, $\tilde{F}_k$, and $\tilde{\theta}_k$ is to set $\tilde{F}_k = \bar{F}_k \cdot 150/100$, $\tilde{A}_k = \bar{A}_k$, and $\tilde{\theta}_k = \bar{\theta}_k$. This is illustrated in Fig. 3. The pitch can indeed be tuned from 100Hz to 150Hz. But the first formant frequency is also shifted from 240Hz to 360Hz. Such shifting of formant frequencies will cause the timbre be distinctly changed. As a result, the timbre of a synthetic syllable will vary with the scaling factors (e.g. 150/100) computed in different control points.

To keep the timbre consistent while tuning the pitch-height of a control point, a principle is to keep the spectral envelope unchanged [6]. This implies that the amplitude $\tilde{A}_k$ of the pitch-tuned harmonic partial located at frequency $\tilde{F}_k$ must be interpolated according to the spectral envelope defined by 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,

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

A figure that illustrates this wav of pitch tuning without changing spectral envelope is drawn in Fig. 4. In this figure, the pitch is promoted by a factor of 1.25 but the envelope is kept unchanged.

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.

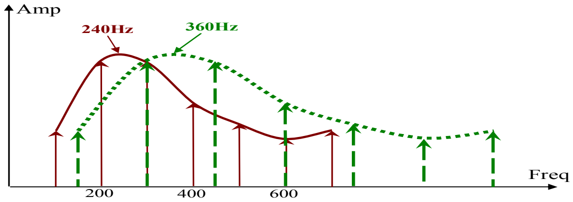

Fig. 3. Pitch tuning that scales spectral envelope

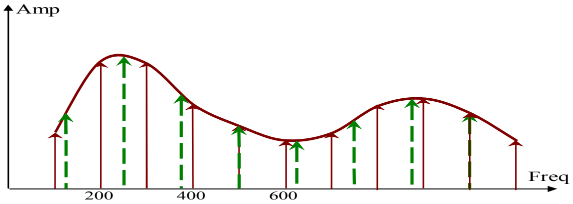

Fig. 4. Pitch tuning that preserves spectral envelope

3. Signal Waveform Synthesis

For the synthetic voiced part in Fig. 2, the synthetic signal, $S(t)$, is consisted of harmonic and noise components. That is, $S(t) = H(t) + N(t)$ where $H(t)$ represents the summation of harmonic partials and $N(t)$ represents the summation of noise signal components. The synthesis of $H(t)$ and $N(t)$ are described in the following subsections.

3.1 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 the rewritten equations,

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

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

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

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

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 test.

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 (t=0 and t=100). This kind 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 this kind 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 to the 100 sample points between two adjacent control points. Here, the amount of mismatched phase is computed 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 $\phi_k^n(100)$ can be directly computed as

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

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

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

3.2 Noise Signal Synthesis

For the noise signal, $N(t)$, we decide to synthesize it as a summation of sinusoidal signal components [3]. Let G_k be the frequency of the k-th sinusoid, and let G_k =100·k (Hz), i.e. G_k need not be distinguished for different control points. 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 = $\lceil$MVF(n) / 100$\rceil$. But the end value is always a fixed value, K_e=$\lfloor$11,025 / 100$\rfloor$.

Besides, 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 are first appended with zeros 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,…,512. 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 noise-signal samples between the n-th and (n+1)-th control points can then be computed with the rewritten equations,

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

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

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

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.

4. Signal Synthesis Experiments

Mandarin has only 408 different syllables if the superimposed tones are not distinguished. Hence, we decide to record and save each of these syllables once for analyzing HNM parameters. Each of these syllables is isolatedly uttered in level tone by a female in a sound proof room. After recording, these syllable signals are analyzed to obtain their HNM parameters. The analysis program is written by us but following Stylianou's method [3]. However, some improvements are made. For example, the frequency values of harmonic peaks in a spectrum are more precisely estimated with parabolic interpolation. And an analysis frame's MVF is more strictly defined as that its following five harmonic candidates are check to be not harmonic peaks.

4.1 Experiment for Timbre Consistency

In this experiment, the two ways for synthesizing pitch-tuned syllable signals as illustrated in Fig. 3 and 4 are considered. First, the two ways are coded into executable programs. Then, the two programs are used to synthesize the syllable pair /lin da/. The pitch contour for /lin/ is assumed to be linear ascending from 145Hz to 195Hz while the contour for /da/ is assumed to be linear descending from 195Hz to 145Hz. Note that the average pitch for the two recorded syllables /lin/ and /da/ are respectively 170Hz and 160Hz. After analyzing the synthetic signal with the software package, WaveSurfer, we obtain the pitch contours and spectrograms as shown in Fig. 5 and 6.

From Fig. 5, it is found that the formant frequencies for /lin/ would go upward as the pitch go higher. Also, the formant frequencies for /da/ would go downward as the pitch go lower. Such pitch-dependent varying of formant frequency values will result in inconsistent timbre within a syllable and between syllables. In perception test, we observed that the timbre for /lin/ and /da/ in Fig. 5 are distinctly different. On the contrary, the timbre for /lin/ and /da/ in Fig. 6 are perceived to be consistent. Note that the heights and shapes of the formant traces in Fig. 6 are kept the same as those of the recorded syllables although the pitch contours are different from those of the recorded. Therefore, heights and shapes of formant traces are very important factors for timbre perception. For demonstration, we have set up a web page, http://guhy.csie.ntust.edu.tw/trhnm/syllable.html, to provide the synthetic signal files corresponding to Fig. 5 and 6, and the recorded signal files for the syllables, /lin da/.

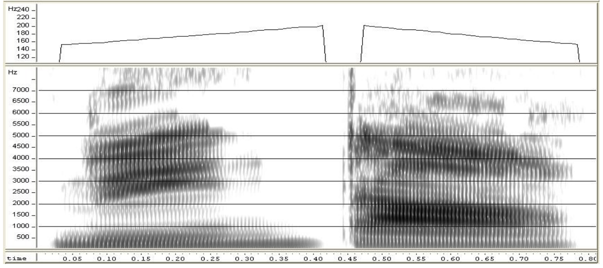

Fig. 5. Pitch and spectrogram using the simple way

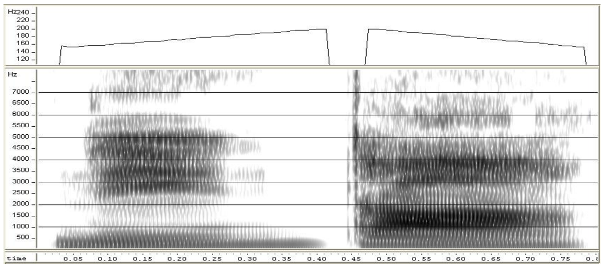

Fig. 6. Pitch and spectrogram using the timbre reserving way

4.2 Experiment for Signal Clarity

In this experiment, the clarity of the speech signal synthesized by the HNM based scheme is considered. Since PSOLA is a popular speech synthesis method, we will compare the speech synthesized by the HNM based scheme with that synthesized by a PSOLA based scheme. First, the two schemes are programmed. The program modules of text analysis and prosodic parameter generation are shared among the two schemes. Then, a same article is fed as the input to obtain two synthetic speech signal files for the two schemes, respectively. As to the speed of signal synthesis, it is found that the HNM based scheme can run in 5 times of real-time on a 2.6GHz Intel Pentium based PC. That is, the ratio of the spent CPU time to the signal file's time length is 20%. And the PSOLA based scheme can run in a much higher speed, i.e. 20 times of real-time or CPU time ratio of 5%.

Afterward, the two synthetic speech files are played in random order to 15 participants to do perception tests. A score of 0 is defined if the clarity of the two 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. According to the scores given by the participants, the average score is computed to be 1.53. That is, the HNM based scheme is apparently better than the PSOLA based scheme in signal clarity. For demonstration, the two speech files synthesized can be browsed and downloaded from the web page, http://guhy.csie.ntust.edu.tw/trhnm/sentence.html.

5. Concluding Remarks

In this paper, HNM is studied and enhanced. Then it is used to design a Mandarin syllable signal synthesis scheme. For the problem of syllable duration lengthening or shortening, the method of linear interpolation plus phase unwrapping as explained with Equations (1), (2), (3), and (4) is proposed to determine pitch-original HNM parameters. Although the method is simple, the synthetic speech signal is perceived to be clear and natural enough. As to the problem of timbre consistency, we have proposed a detailed Lagrange interpolation based method as explained around Equation (5). According to the experiments performed in Sections 4.1 and 4.2, the timbre can indeed be kept consistent although each syllable is only recorded and saved once. In addition, we have based on the enhanced HNM to design a scheme for synthesizing Mandarin syllables' signals as illustrated in Fig. 1. With this scheme, syllable signals of diverse prosodic characteristics can be synthesized from a source syllable's HNM parameters without significant signal quality degradation.

References

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

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

[3] Stylianou, Yannis, *Harmonic plus Noise Models for Speech, Combined with Statistical Methods, for Speech and Speaker Modification*, Ph.D. thesis, Ecole Nationale Supèrieure des Télécommunications, Paris, France, 1996.

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

[5] Quatieri, T. F., *Discrete-Time Speech Signal Processing*, Prentice-Hall, NJ, USA, 2002.

[6] Dodge, C. and T. A. Jerse, *Computer Music: Synthesis, Composition, and Performance*, second edition, Schirmer Books, New York, 1997.

[7] Moore, F. R., *Elements of Computer Music*, Prentice-Hall, 1990.