MANDARIN SINGING VOICE SYNTHESIS USING ANN VIBRATO PARAMETER MODELS

HUNG-YAN GU, ZHENG-FU LIN

Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology, Taipei, Taiwan
E-MAIL: guhy@mail.ntust.edu.tw, m9415017@ mail.ntust.edu.tw

Abstract:

In this paper, the vibrato parameters of sung syllables are analyzed by using short-time Fourier transform and the method of analytic signal. After the vibrato parameter values for all training syllables are obtained, they are used to train an artificial neural network (ANN) for each type of vibrato parameter. Then, these ANN models are used to generate the values of vibrato parameters. Next, these parameter values and other music information are used together to control a harmonic-plus-noise (HNM) model to synthesize singing voice signals. With the synthetic singing voice, subjective perception tests are conducted. The result show that the singing voice synthesized with the ANN generated vibrato parameters is apparently more natural than the singing voice synthesized with fixed vibrato parameters.

Keywords:

Singing voice; signal synthesis; vibrato parameter; artificial neural network; harmonic-plus-noise model

1. Introduction

Several techniques for the synthesis of singing voice signals had been proposed, including phase vocoder [1, 2], formant synthesis [1, 2], ABS/OLA sinusoidal model [3], PSOLA synthesis [4], and EpR model[5]. Also, we have investigated an HNM (harmonic-plus-noise model) based and enhanced scheme to synthesize the signals of Mandarin singing voice [6]. Nowadays, to synthesize clear and natural signals of singing voice is not difficult. However, the synthesized singing voice is not felt as so expressive as that sung by a real singer. In fact, it may be felt as sung by a mechanical tongue. One major reason is that the factors relevant to the expressing of singing are not adequately modeled and controlled. Such factors include vibrato, marcato, soffocato, rubato, accelerando, ritardando, etc. Among these factors, vibrato is thought to be a very important one. Therefore, in this paper, we study to analyze and model vibrato expressing in order to synthesize Mandarin singing voice that can express natural vibrato.

According to the studies of Horii [7] and Imaizumi, et al. [8], the most noticeable phenomenon due to vibrato is that the pitch-frequency will vibrate periodically. An example is as the solid-lined curve in Figure 1, which is obtained from analyzing a sung syllable. In this figure, the pitch-frequency is seen to vibrate between 265 to 285Hz, and the vibrating rate is about 5Hz. Therefore, to synthesize singing voice with vibrato expression, the pitch-frequency is the major acoustic factor to deal with.

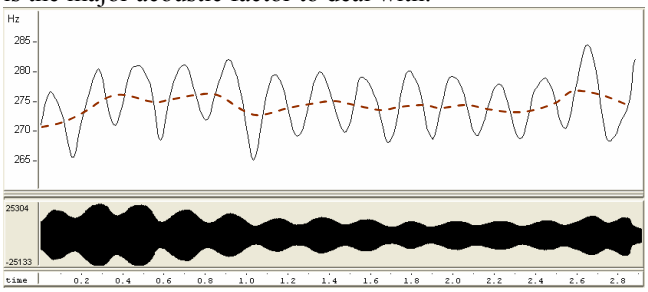

Figure 1. Pitch contour from analyzing a vibrato singing

Although a vibrating pitch contour may be synthesized by applying some rules [1], its perceived naturalness level is however questionable. Hence we deicide to construct ANN based models. The models will not generate a vibrating pitch contour directly but generate its corresponding vibrato parameters. Then, in terms of these parameters, a vibrato-expressing pitch contour can be indirectly generated. Such an approach is realizable because, according to the research results by Sundberg, et al. [9], and Shonle and Horan [10], a vibrating pitch contour can be analyzed and represented with three types of parameters, i.e. intonation, vibrato extent, and vibrato rate. Intonation means smooth and averaged pitch height as the dash-lined curve in Figure 1. Vibrato extent is the extent of vibration (i.e. peak value minus intonation value) and vibrato rate is the changing rate of the pitch-frequency.

After a vibrato-expressing pitch contour is generated with the constructed ANN models, the pitch contour can be

used to determine the pitch-tuned HNM parameter values for each control point placed on the time axis of the singing voice to be synthesized [6]. Then, the singing voice with vibrato expression can be synthesized by using the HNM based synthesis scheme studied previously [6]. HNM is originally proposed by Y. Stylianou [11, 12]. It may be viewed as improving the sinusoidal model to better model the noise signal components in the higher frequency band of voice signal.

2. Vibrato parameter analysis

Before vibrating pitch contours can be generated, vibrato parameters must be analyzed from a real singer's singing voice and then used to train the ANN models. That is, in the training stage, we follow the steps of the flowchart in Figure 2 to do parameter analysis and model training. First, song signals sung by a real singer are recorded. Secondly, the recorded signals are labeled manually with pronunciation symbols and segmented into separate syllable signal files. For each syllable's signal, its IPF (instantaneous pitch frequency) curve is measured. Then, the IPF curve is further analyzed to extract intonation, vibrato extent, and vibrato rate parameters. The processing steps mentioned above will be detailed in the following subsections. But the training of the ANN models is explained in the next section.

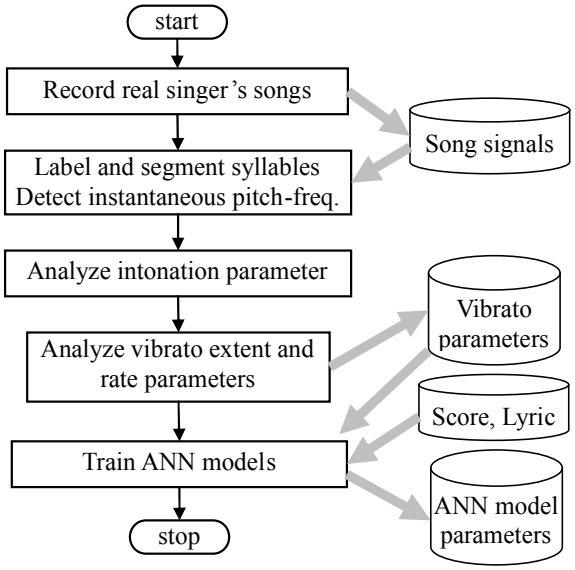

Figure 2. Flow of the works done in the training stage

2.1. Singing voice signal recording

In this study, we invite a male singer to sing songs of popular music in an Acoustic-Systems RE-242 sound-proof room. He followed the midi accompaniment played to his headphone. Hence, the pitch of each lyric should be in tune with the accompaniment basically. Signal recording is made in real-time (i.e. signal samples are directly saved to a computer) and the sampling rate is 22,050Hz. For the 15 songs sung, the number of lyric syllables is totally 2,841, and the tempos of these songs include the slower and quick.

2.2. Instantaneous pitch frequency measuring

In a Mandarin song, usually a lyric syllable has only a music note assigned to it. Hence, syllable is taken as the processing unit. Here, each lyric syllable's signal is labeled manually with the package, WaveSurfer, and then segmented into separate signal files. For a Mandarin syllable, it may be started with an unvoiced initial consonant but its final is always voiced. Therefore, the boundary point of unvoiced and voiced signals is determined first with a pitch detection method. Then, the curve of IPF is measured after the boundary point.

The way of measuring is as the following. First, the part of voiced signal is segmented into a sequence of frames. The length of each frame is 512 sample points but frame shift is only 32 points. For each frame, zero valued samples are appended in order to perform 4,096 points FFT (fast Fourier transform). Then, on the FFT spectrum, five spectral peaks are searched after 0Hz. Let $g(i)$ denote the frequency value of the i-th spectral peak. After $g(i)$ is found, it is divided by i to give an estimate of fundamental frequency. Then, the five estimates are geometrically averaged to give an IPF value for this frame. When IPF values of all frames are obtained, they can be connected to form an IPF curve. This IPF curve, $f(t)$, may be viewed as of the form

$$f(t) = V_d(t) + V_e(t) \cdot cos(\phi(t)) \quad (1)$$

where $V_d(t)$ represents its intonation parameter, $V_e(t)$ represents its vibrato-extent parameter, and its vibrato-rate parameter, $V_r(t)$, can be derived as

$$V_r(t) = \frac{1}{2\pi} \cdot \frac{d\phi(t)}{dt} \quad (2)$$

2.3. Analysis of intonation parameter

A simple idea to obtain the intonation curve, $V_d(t)$, is to low pass the IPF curve, $f(t)$. In practice, low pass filtering may be done in the frequency domain or time domain. When filtered in the frequency domain, we found a serious problem. That is, the difference between the IPF and intonation curves may become large near the two ends. Hence, we decide to use a moving average based filter

finally. Although moving average is simple, it can however prevent the problem mentioned. At a time point t, the IPF values, $f(\tau)$, τ=t-128, t-127, …, t+128, are averaged to get the intonation value, $V_d(t)$.

2.4. Analysis of vibrato extent and rate

To obtain the curves of vibrato extent $V_e(t)$ and vibrato rate $V_r(t)$, the signal, $s(t)$, defined here as $s(t) = V_e(t)\cdot cos(\phi(t))$, is computed first as $f(t) - V_d(t)$ according to Equation 1. Then, by using the analysis method of analytic signal [13], $V_e(t)$ and $\phi(t)$ can be derived consequently.

According to Gabor's definition [13], the analytic signal of $s(t)$ is $z(t)$ and $z(t)$ is composed with the real part, $s(t)$, and the imaginary part, $\hat{s}(t)$. That is,

$$z(t) = s(t) + \mathrm{j}\cdot\hat{s}(t) \ ,$$
$$\hat{s}(t) = H\left[s(t)\right] = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{s(\tau)}{t-\tau}d\tau \ , \tag{3}$$

where $H\left[s(t)\right]$ denotes Hilbert transform. Hilbert transform can rotate the phase angle of the signal with the right amount of π/2 or -π/2. Consequently, we obtain that $\hat{s}(t) = V_e(t)\cdot sin(\phi(t))$ and $z(t) = V_e(t)\cdot exp(\mathrm{j}\cdot\phi(t))$. Then, $V_e(t)$ and $\phi(t)$ can be derived as

$$V_e(t) = \sqrt{s^2(t) + \hat{s}^2(t)} \ ,$$
$$\phi(t) = arc\,tan\left(s(t), \hat{s}(t)\right) \tag{4}$$

In terms of $\phi(t)$, vibrato rate, $V_r(t)$, can be computed according to Equation (2).

3. ANN vibrato parameter models

In this study, ANN models are adopted to learn the singing style of the invited singer in expressing vibrato. Note that four types of parameters are analyzed in last section, i.e. intonation $V_d(t)$, vibrato extent $V_e(t)$, vibrato rate $V_r(t)$, and initial phase $\phi(0)$. Therefore, we decide to train and build an ANN for each of the parameter type. Actually, each ANN is a multi-layer perceptron (MLP) [14]. Here, the adopted learning algorithm is back propagation. The structure of each MLP is as shown in Figure 3. That is, only one hidden layer is placed between the input and output layers. Within each node of the hidden and output layers, the hyperbolic tangent function,

$$f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \ , \tag{5}$$

is adopted as the transformation function because the target values of the vibrato parameters may be negative or positive. The number of nodes in the output layer is 32 for three of the MLPs but the MLP for initial phase needs only one output node. The details for vibrato parameter representation and normalization are given in Subsection 3.1. As to the input layer, it is used to accept the contextual information of the current syllable to be sung. The details of the contextual information adopted here are given in Subsection 3.2. For the number of nodes to be used in the hidden layer, several MLP training experiments have been done to test various settings. The results show that 8 nodes is the best choice.

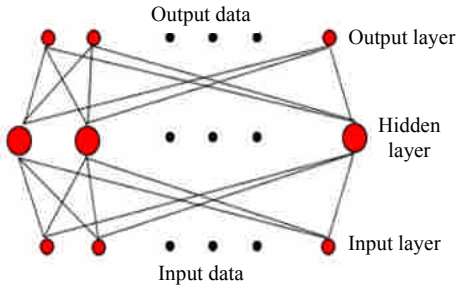

Figure 3. The structure of each MLP

3.1. Vibrato parameter sampling and normalization

Note that the time lengths of the intonation curves (or vibrato extent or rate curves) analyzed from different syllables may vary dramatically under different tempos and beats. Hence, an adequate representation of a curve must be adopted in order to have fixed dimensions of target values for an MLP to learn. Here, a simple representation method is adopted, i.e. sampling a curve at 32 uniformly placed time points. In details, a vibrato parameter's curve, $V_x(t)$, is sampled to $U_x(i) = V_x(T*i/31)$, i=0, 1, …, 31, where T is the time length.

On the other hand, consider the synthesis of a curve when given 32 outputted values, $U_x(i)$, from an MLP and a target time length T. The basic idea is to generate the curve by means of interpolation. Currently, a simple way of piece-wise linear interpolation is adopted, which seems enough. In details, for a sample time point t, the intervals $[T_i, T_{i+1})$, i=0, 1, …, 30, and $T_i = T*i/31$, are searched first to locate the interval $[T_k, T_{k+1})$ that contains t. Then, the value of the interpolated curve, $V_x(t)$, at time t is computed as

$$V_x(t) = U_x(k) + \left(U_x(k+1) - U_x(k)\right)\frac{t - T_k}{T_{k+1} - T_k} \tag{6}$$

In training a MLP, the 32 sampled values, $U_x(i)$, of a curve, are not directly used as the target values for the MLP to learn. This is because the function values of the transformation function defined in Equation (5) are just

ranged from -1 to 1. To suit this value range, the sampled values must be normalized beforehand. When $U_d(i)$ are sampled from an intonation curve, $V_d(t)$, we first define the normalization factor M_d as the geometric mean of those sampled values from the second of the three parts. In details, M_d is defined as

$$M_d = \left(\prod_{i=11}^{20} U_d(i) \right)^{1/10} \quad (7)$$

The first and third parts are not used because their sampled values may be unstable due to portamento. After M_d is determined, the normalized values are computed as

$$\hat{U}_d(i) = \frac{U_d(i)}{M_d} - 1, \quad i = 0,1,...,31 \quad (8)$$

When $U_e(i)$ are sampled from a vibrato extent curve, $V_e(t)$, the way of normalization here is to divide $U_e(i)$ by $U_d(i)$, i.e. $\hat{U}_e(i) = U_e(i) / U_d(i)$. As to the curve of vibrato rate, its sampled values, $U_r(i)$, are normalized here by dividing the constant 20, i.e. $\hat{U}_r(i) = U_r(i) / 20$. In addition, the value of initial phase, $\phi(0)$, is normalized by dividing the constant 5.

3.2. Contextual information and their classification

What factors is the expressing of vibrato affected by? We think the factors include (a) the duration, initial type, and final type of the current syllable, (b) the duration and final type of the previous syllable, (c) the duration and initial type of the next syllable, and (d) the tone differences between the current note and its previous and next notes. Since the number of factors considered is not small, the number of possible combinations of these factors' values will be very huge. However, the training data collected include just 15 songs that are comprised of just 2,841 syllables. Therefore, classification of these factors' values is inevitably needed in order to reduce the number of possible combinations.

Among the three duration factors, the current syllable's duration is thought to be more important than the two adjacent syllables' durations. Therefore, we decide to divide the current syllable's duration into 5 classes but to divide the adjacent syllables' durations into just 3 classes. For current syllable, the 5 classes are defined as 0~0.3 sec., 0.3~0.5 sec., 0.5~0.8 sec., 0.8~1.3 sec., and above 1.3 sec. For adjacent syllables, the 3 classes are defined as 0~0.25 sec., 0.25~0.5 sec., and above 0.5 sec. Thus, 3bits and 2 bits are needed respectively to distinguish these classes.

There are two syllable-final factors. Here, the 39 syllable-final types of Mandarin are divided into 4 classes. That is, single vowels (e.g. /a/), diphthongs (e.g. /ai/), triphthongs (e.g. /iau/), and nasal-ended finals (e.g. /ang/). Also, there are two syllable-initial factors. The 21 syllable-initial types of Mandarin are divided into 3 classes. That is, voiced consonants (e.g. /m, r/), short unvoiced consonants (e.g. /b, z/), and long unvoiced consonants (e.g. /p, s/). Therefore, syllable initial and final classes need respectively 2 bits to distinguish.

As to the two factors of tone differences, 7 classes are defined. Tone difference is computed in semitones. The elements of the 7 classes are as listed in Table 1. To distinguish these classes, 3 bits are used.

Table 1. Classes of tone differences

Class	1	2	3	4	5	6	7
Elements	-6,-7, -8,…	-3,-4, -5	-1,-2	0	1, 2	3, 4, 5	6, 7, 8,…

4. Singing voice synthesis and perception test

Based on the MLP vibrato parameter models, we have constructed a Mandarin singing voice synthesis system. The main processing flow of this system is shown in Figure 4.

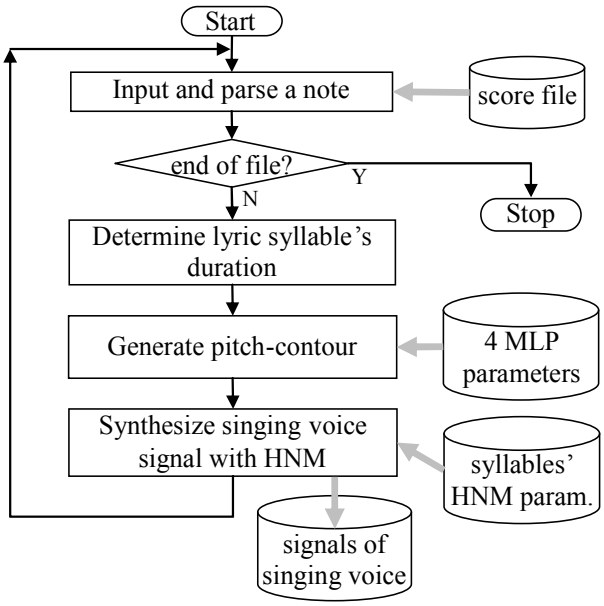

Figure 4. Main flow for Mandarin singing voice synthesis

Each time a music note is inputted from a score file and parsed. The duration of the note's lyric syllable is then computed according to the information of tempo and beats. Next, the contextual information is gathered, and the 4 MLP vibrato parameter models are used to generate a pitch contour with vibrato expressing. The details of the

generating process are explained in Subsection 4.1. Afterward, the last block of Figure 4 uses the inputted pitch contour to adjust the lyric syllable's HNM parameters, and synthesize the corresponding singing voice signal with an HNM based method. This method is explained in Subsection 4.2.

4.1. Pitch contour generating

When the contextual information of the current lyric syllable is fed to the 4 MLPs, the normalized and sampled vibrato parameters can then be obtained from the output-layers of the MLPs. Next, inverse normalizations are performed respectively according to the formula mentioned near Equation (8) to obtain the sampled parameters, $U_d(i)$, $U_e(i)$, and $U_r(i)$, and the initial phase, $\phi(0)$ in correct scales.

To generate an intonation curve, the pitch frequency (in Hz), F, of the current note is needed, which can be computed from its tone symbol (e.g. "G3"). Also, the duration, T, of the lyric syllable is needed, which is already computed in the second block of Figure 4. By replacing M_d in Equation (8) with F, the sampled intonation parameters, $U_d(i)$, can be computed as $U_d(i)=(\hat{U}_d(i)+1)\cdot F$, and they would have the correct pitch. Next, by performing interpolation according to Equation (6) and the duration T, an intonation curve, $V_d(t)$, can then be generated.

When the sampled intonation parameters, $U_d(i)$, is ready, the sampled vibrato extent parameters, $U_e(i)$, can be computed as $U_e(i)=\hat{U}_e(i)\cdot U_d(i)$. Then, the vibrato extent curve, $V_e(t)$, can be generated by interpolation according to Equation (6) and the duration T. Similarly, the vibrato rate curve, $V_r(t)$, can also be generated by interpolating the sampled parameters, $U_r(i)$.

After the curves, $V_d(t)$, $V_e(t)$, and $V_r(t)$, are generated, the phase curve, $\phi(t)$, is next computed with $V_r(t)$ as

$$\phi(t)=\phi(t-1)+2\pi\cdot V_r(t)\cdot\frac{1}{22{,}050},\quad t=1,2,...,T-1 \qquad (9)$$

where 22,050 is the sampling rate. Then, the pitch contour, $P(t)$, can be generated as

$$P(t)=V_d(t)+V_e(t)\cdot cos\left(\phi(t)\right),\quad t=0,1,...,T-1 \qquad (10)$$

4.2. Singing voice signal synthesis

Note that Mandarin is a syllable prominent language and there are only 408 different syllables when the lexical tones are not distinguished. Therefore, we recorded and saved each syllable just once for HNM parameter analyzing. Here, the 408 syllables are uttered by a female in the same RE-242 sound-proof room. Apparently, the syllables recorded for singing voice synthesis and the songs recorded for training MLPs are provided by different persons.

Since each syllable has only one utterance, there is no chance to do unit selection. That is, diverse combinations of pitch height and duration length must all be synthesized from a same syllable's analyzed HNM parameters. To accomplish this goal, the timbre must be kept consistent when the pitch is tuned to a pitch contour as generated by Equation (10). Also, the synthetic syllable signal must be made as fluent as possible when the syllable duration is lengthened or shortened. These two problems are already investigated and feasible solution methods are proposed in our previous study [6]. Besides, the original synthesis methods proposed by Stylianou [11, 12] may also be referred. Therefore, the method of the HNM based signal synthesis will not be detailed here.

In brief, the singing voice signal is synthesized as the harmonic signal, $H(t)$, plus the noise signal, $N(t)$. $H(t)$ is synthesized as

$$H(t)=\sum_{k=0}^{L}a_k^n(t)\,cos\left(\phi_k^n(t)\right),\quad t=0,1,...,T^n, \qquad (11)$$

where L is the number of harmonic partials, T^n is the number of samples between the n-th and $(n+1)$-th control points, $a_k^n(t)$ is the time-varying amplitude of the k-th partial at time t, and $\phi_k^n(t)$ is the cumulated phase for the k-th partial. Here, the harmonic partials occupy the lower frequency band and their frequencies are time varying. On the other hand, $N(t)$ is also synthesized as a summation of sinusoidal components here. But these sinusoids occupy the higher frequency band and their frequencies do not change with time.

4.3. Perception test

A same song score is used here to synthesize three singing voice files. The first file denoted with SA is synthesized with no vibrato. This can be accomplished by setting $U_d(i) = F$ and $U_e(i) = 0$. The second file denoted with SB is synthesized with fixed vibrato. That is, set $U_d(i) = F$, $U_e(i) = F*3/100$, and $U_r(i) = 4$. As to the third file, it is denoted with SC and its vibrato parameters are generated by the 4 MLPs. Then, the three files are played in the order, SA, SB, SC, to 15 participants to do perception tests. Each participant is requested to give two scores of naturalness comparison, i.e. comparing SA with SB and comparing SA with SC. A score of 0 is defined if the naturalness level between two synthetic singing voice 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 averaged scores are 0.63 for comparing SA with SB and 1.29 for comparing SA with SC. These average scores show that using MLPs to generate vibrato parameters can indeed help synthesizing more natural singing voice. For demonstration, the web page, http://guhy.csie.ntust.edu.tw/vibrato/, is prepared which can be browsed to download the three synthetic singing voice files.

5. Conclusions

Vibrato is commonly found in singing voice as a way for expressing. Therefore, vibrato style modeling and vibrato parameter generating are important issues for a computer to synthesize natural and expressive singing voice. In this paper, we study to analyze three types of vibrato parameters. For a vibrato parameter curve, 32 uniformly sampled data are adopted to represent it. Then, the sampled data are normalized by using the methods studied here. In addition, we propose to use an MLP to model each type of vibrato parameter by training the MLP with the sampled and normalized data.

According to the experiment results, short-time Fourier transform based instantaneous pitch frequency detection and the analysis method of analytic signal are found to be feasible for vibrato parameter analyzing. Also, according to the result of perception tests, the singing voice synthesized with MLP generated vibrato parameters is apparently more natural than the singing voice synthesized with just fixed vibrato parameters. Currently, the quantity of recorded syllables for training the MLPs is not sufficient. In the future, we will record more songs to study the performance of the MLP vibrato parameter models.

Acknowledgements

This paper is supported by National Science Council of Taiwan under the Grant number NSC-96-2218-E-011-002.

References

[1] F. R. Moore, Elements of computer music, Prentice-Hall, 1990.

[2] C. Dodge and T. A. Jerse, Computer music: synthesis, composition, and performance, second edition, Schirmer Books, New York, 1997.

[3] M. W. Macon, L. Jensen-Link, J. Oliverio, M.A. Clements and E.B. George, "A singing voice synthesis system based on sinusoidal modeling", Proceedings of IEEE ICASSP, Vol. 1, pp. 435-438, 1997.

[4] N. Schnell, G. Peeters, S. Lemouton, P. Manoury, and X. Rodet, "Synthesizing a choir in real-time using pitch synchronous overlap add", Proceedings of Int. Computer Music Conference., pp. 102-108, 2000.

[5] J. Bonada and A. Loscos, "Sample-based singing voice synthesizer by spectral concatenation", Proceedings of the Stockholm Music Acoustics Conference, Stockholm, Sweden, Aug. 2003.

[6] H. Y. Gu and H. L. Liau, "Mandarin singing voice synthesis using an hnm based scheme", to appear in Proceedings of International Congress on Image and Signal Processing, Sanya, China, May 2008.

[7] Y. Horii, **"**Acoustic analysis of vocal vibrato: a theoretical interpretation of data**"**, J. Voice, **3**, pp. 36–43. 1989.

[8] S. Imaizumi, H. Saida, Y. Shimura, and H. Hirose, **"**Harmonic analysis of the singing voice:—Acoustic characteristics of vibrato**"**, Proceedings of the Stockholm Music Acoustics Conference, Royal Swedish Academy of Music, Stockholm, pp. 197–200. 1994.

[9] J. Sundberg, E. Prame, and J. Iwarsson, **"**Replicability and Accuracy of Pitch Patterns in Professional Singers**"**, Vocal Fold Physiology, edited by P. J. Davis and N. H. Fletcher, Singular, San Diego, 1996.

[10] J. I. Shonle and K. E. Horan, "The pitch of vibrato tones", J. Acoust. Soc. Am., Vol. 67, pp. 246–252. 1980.

[11] Y. Stylianou, 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.

[12] Y. Stylianou, "Modeling speech based on harmonic plus noise models", Nonlinear speech modeling and applications, Springer-Verlag, Germany, 2005.

[13] H. G. Feichtinger and T. Strohmer, Gabor analysis and algorithms: theory and applications, Birkhauser, Boston, 1998.

[14] K. Gurney, An introduction to neural networks, UCL Press, 1997.