# Singing Voice Synthesis: Singer-dependent Vibrato Modeling and Coherent Processing of Spectral Envelope

*S. W. Lee and Minghui Dong*

Human Language Technology Department, Institute for Infocomm Research, A*STAR, Singapore 138632

{swylee, mhdong}@i2r.a-star.edu.sg

## Abstract

Pleasant singing voice is often ornamented by vibrato. This pitch fluctuation acts as a distinctive feature for singing and promotes voice quality. Nevertheless, independent pitch processing in singing voice synthesis does not guarantee the output quality. The spectral envelope actually varies with pitch during human voice production. This paper proposes a modeling technique for singers' vibratos, followed by a joint processing on vibrato and spectral envelope, such that these attributes are consistent. The performance of the proposed processing has been verified by subjective listening test. The synthetic singing outputs are found to have similar quality as the human singing.

**Index Terms**: singing voice synthesis, vibrato, spectral envelope

## 1. Introduction

Singing voice synthesis has been one of the emerging and popular research topics recently [1]-[3]. There is a growing number of applications, especially for pop songs, ranging from music production, entertainment development to computer-assisted vocal training [4]-[7]. Majority of the singing voice synthesis methods are derived from two approaches. The first approach directly generates a singing voice from lyrics, similar to the conventional text-to-speech systems [3], [4], [8]. In this approach, the output singing voice either mimics the vocal features of stored recordings such as lyrics, tones, durations, etc. or is built by concatenating recording segments. Consequently, the output vocal characteristics are inherited and fixed.

The second approach converts a lyrics-reading speech input into singing voice [9], [10]. Based on the given melody, the input speech spectrum and pitch are modified accordingly. Some attributes unique to singing, for example, vibrato, singing formant, preparation, etc., may be added to the output singing voice. Compared with the former approach, the output spectra resulted from the latter approach vary case-by-case, depending on the input speech. In other words, the input vocal characteristics are often preserved. This essentially improves the flexibility of the output singing voice. Pleasant singing voice can thus be generated from any individual's recorded speech; even he or she is not good at singing.

In the following study, our analysis perspective is based on this speech-to-singing synthesis approach. In particular, we focus on how to manipulate the pitch contour and the spectra to generate pleasant singing voice.

Pleasant singing voice is often accompanied by vibrato. This also applies to musical instrument performance. By vibrato, it refers to a periodic fluctuation in pitch of a musical tone [11], [12]. This fluctuation is sometimes sinusoidal. Vibrato is essential for pleasant performance. It has been considered as a sign of the singer's vocal skill. In Western operatic singing, the vocal skill of a singer is often associated to how regular this periodic fluctuation is. Likewise, performers may color a tone, personalize the performance and express their emotions with vibrato. More important, vibrato is found to contribute most to perceived voice quality, compared to various singing features [13].

Despite of these important roles of vibrato, most of the studies of singing voice and vibrato are on Western operatic singing. Analysis of pop song singing is rare. The enormous digital pop song collection is left unexplored. Besides, the characteristics of vibrato are generally considered to be personal, and remain the same for a singer [11], [12]. Individual singers may possess similar or distinct vibrato characteristics. In order to generate pleasant singing voice, the vibrato exhibited in the synthetic singing voice is expected to carry the singer's characteristics. Direct adoption of vibrato findings from Western operatic studies may not be appropriate.

The pitch frequency (F0) during vibrato does not fluctuate in an independent manner. Instead, the spectral envelope changes coherently with vibrato [8], [11]. Perceived spectral envelope and formant frequencies vary with F0. Referring to Figure 1, the vocal tract filters in Scenario A and B are identical, as shown in the top diagrams. The pitch frequencies (and the harmonics) are different. With the inference on the associated harmonics, the perceived spectral envelopes, as shown in the bottom diagrams, differ. Formant frequencies shift and some of the spectral peaks are smoothed out.

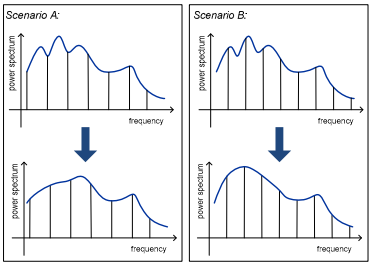


Figure 1: *Demonstration of the coherent interaction between F0 and spectral envelope.*

Natural singing voice always possesses coherent vibrato and spectral envelope. To the contrary, in typical singing voice synthesis systems, vibrato is often simply included in pitch contour without the corresponding change in spectral envelope. This violates the intrinsic relationship between F0 and spectral envelope, degrading the quality of output singing voice.

In this work, a coherent processing of spectral envelope matching the instantaneous F0 is proposed. We start by introducing our feature extraction and singer-dependent vibrato modeling. This vibrato modeling, i.e. sole modification on F0, has been verified to improve the quality and singingness for pop-song speech-to-singing synthesis. It is further enhanced with the relevant modification on spectral

Copyright © 2011 ISCA

envelope. The aforementioned intrinsic relationship is, hence, preserved. Our experiment has shown that this proposed coherent processing of spectral envelope and the vibrato modeling is effective to provide similar quality of singing voice as the real vocal recordings.

## 2. Speech-to-singing synthesis and vibrato characterization

Our speech-to-singing synthesis method is reviewed first. It illustrates how the proposed vibrato modeling and spectral envelope processing are employed in singing voice synthesis. Building singer-dependent vibrato models requires the learning of individual singers' F0 characteristics. We will then briefly describe a feature extraction scheme that we used to characterize a vibrato segment.

### 2.1. Speech-to-singing synthesis

Figure 2 depicts the block-diagram for the speech-to-singing synthesis method used. Given a lyrics-reading speech input $x(n)$, Tandem-STRAIGHT [14] is used to decompose $x(n)$ into sequences of spectral envelope, F0 and aperiodicity. The alignment information of the speech input is then extracted by forced-aligning $x(n)$ with the lyrics. This gives the timing information of each syllables in $x(n)$ and completes the analysis-phase. During the synthesis-phase, the vocal-timing process computes the target sequences of spectral envelope and aperiodicity for the singing output $y(n)$. This is done by locating the corresponding syllables in $x(n)$ with the alignment information and replicating the associated spectral envelopes and aperiodicity functions according to the melody. Typical representations for this melody are score and MIDI. For each melodic note, the consonant part of the corresponding syllable is lengthened, such that the syllable duration matches with the note duration. Based on the melody, the singing F0 contour is found by calculating the note duration and converting the note number to frequency. Other fluctuations, for example, overshoot and preparation, etc., may be included. Finally, the output singing voice is synthesized by Tandem-STRAIGHT with the estimated sequences of spectral envelope, F0 and aperiodicity.

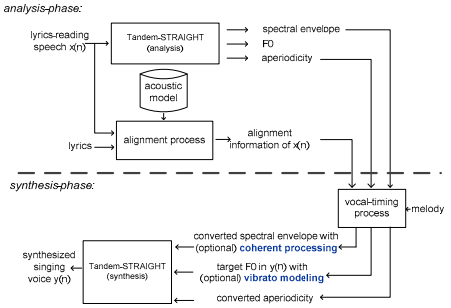


Figure 2: *The block-diagram of the speech-to-singing synthesis method used.*

With the proposed vibrato modeling and spectral envelope processing, target F0 and the spectral envelope will be revised on the outputs of the vocal-timing process before synthesis.

### 2.2. Vibrato feature extraction

Individual singers' vibrato characteristics are first collected by the feature extraction described below. Being a periodic fluctuation in pitch, vibrato is characterized by two parameters: rate and extent [11]. The vibrato rate represents the number of vibrato cycles per second. The extent describes how far the F0 rises and falls from the average pitch of a vibrato segment concerned (in a relative sense). The feature extraction scheme below aims to estimate these two parameters. Traditionally, vibrato is often manually analyzed by professionals [12], whereas some of the recent feature extractions are applicable to sinusoidal vibratos only [15]. As vibrato tone may take various forms, e.g. triangular vibrato in cello, saw tooth-shaped vibrato in flute, etc., our extraction scheme does not assume any specific shape of vibrato.

With a singing voice recording and the associated electroglottograph (EGG) signal, the F0 contour is extracted by Tandem-STRAIGHT. EGG is used for accurate pitch estimation, leading to precise vibrato modeling. In our implementation, the pitch range for Tandem-STRAIGHT is from 80 to 1100 Hz.

Let $p(n)$ be the extracted pitch contour of a vibrato segment. $p(n)$ is normalized with its average value, so as to calculate the vibrato extent. The mean value of normalized pitch contour is further removed. Let $\hat{p}(n)$ and $\tilde{p}(n)$ be the normalized F0 and the zero-mean, normalized F0 respectively. The vibrato rate $r$ is estimated by locating the peaks in the autocorrelation function $R_{\tilde{p}}(m)$ of $\tilde{p}(n)$. Specifically, the non-zero lag $m^*$ ($m^* > 0$) maximizing $R_{\tilde{p}}(m)$ is selected and

$$r = \frac{1}{m^* \cdot \text{F0 sampling period}}. \quad (1)$$

The vibrato extent $e$ is estimated as the mean of the maximum deviations observed in each vibrato period. $\hat{p}(n)$ is first partitioned into finite numbers of periods, based on $m^*$. Let $L$ denote the number of periods in total. For the $l$-th period, the maximum deviation $d_l$ is calculated as

$$d_l = 1200 \times \left| \log_2 \hat{p}_l(n^*) \right|, \quad (2)$$

where $\hat{p}_l(n^*)$ is the normalized F0 value with maximum deviation. $n^*$ denotes where the maximum deviation is. The logarithmic operation and multiplication by 1200 convert the deviation into cents. Finally,

$$e = \frac{1}{L}\sum_{l=1}^{L} d_l \,. \quad (3)$$

The estimates $\{r, e\}$, will be used to built vibrato models in the following section.

## 3. Singer-dependent vibrato modeling

Previous speech-to-singing systems often directly adopt Western operatic vibrato findings and generate identical vibrato patterns. It is not guaranteed that the resultant vibrato matches nearby F0s and the singer's vocal characteristics or not. We propose to build vibrato models, via statistical learning from individuals' pop song data. The following demonstrates our methods for building a vibrato model.

### 3.1. Methods

A collection of solo singing recordings from a female professional singer is recorded and used for the studies below. The data include both audio and EGG waveforms. There are ten Mandarin Chinese pop songs in total. Each song lasts about four mins, summing up to 37 mins 42 secs. These songs were selected by the singer, based on her singing skills, rhythms and pitch ranges, such that she sings well. Besides, the MIDI files for these songs are available and will be used as the melody for speech-to-singing synthesis.

*Vibrato modeling*: After pitch tracking by Tandem-STRAIGHT on the EGG signal, the F0 contour is manually inspected and vibrato segments are located. The vibrato rates and extents $\{r, e\}$ of these vibrato segments are estimated by

the above feature extraction. As singing voice vibrato is generally constant within a singer and difficult to change, a two-dimensional single-mixture Gaussian model (GMM) is used. Neither Hidden Markov model (HMM) nor multiple mixtures is adopted, as single-mixture GMM is suitable for the stationary nature and the results can be directly applied to generate artificial vibrato $v(n)$. This GMM also characterizes the vocal for a particular singer, which may be used for singer identification purpose.

Suppose $\{\bar{r}, \bar{e}\}$ is the GMM distribution mean vector. Artificial vibrato will be incorporated in where long-lasting, constant MIDI F0 segments lie. In our implementation, segments longer than 0.5 secs are replaced $v(n)$. $v(n)$, in the form of sinusoid, is defined as

$$\hat{e} = 2^{\frac{\bar{e}}{1200}} \tag{4}$$

$$v(n) = m(n)\left[(\hat{e}-1)\sin\left(\frac{2\pi\bar{r}n}{f_{ps}}\right)+1\right] \tag{5}$$

where $m(n)$ and $f_{ps}$ are the constant pitch value and the F0 sampling frequency (200 Hz in our experiments) respectively. Figure 3 shows an example of the resultant F0 contour with artificial vibrato.

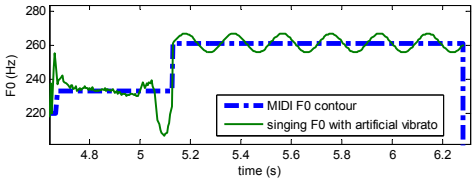


Figure 3: *Resultant F0 with vibrato modeling.*

### 3.2. Experiments and results

A singing voice synthesis method is found to be effective if it generates pleasant, singing-like output. In our experiment, the performance of the singer-based vibrato modeling is evaluated in terms of two indices: (1) singing voice quality; and (2) 'singingness' (the degree of singing). The artificial vibrato is applied on the F0 contour extracted from the reference singing F0 (as shown in Figure 3).

Figure 4 depicts the scatter plot of $\{r, e\}$. The GMM distribution is also shown. $\{\bar{r}, \bar{e}\}$ is found to be {5.19, 36.26}. Comparing with the typical vibrato values ($5 \le r \le 8$ Hz and $30 \le e \le 150$ cents) in Western operatic singing, this singer exhibits a slower and smaller vibrato. The covariance matrix is

$$\begin{bmatrix} 2.17 & -4.38 \\ -4.38 & 779.98 \end{bmatrix}.$$

By calculating the standard deviation of $e$ (27.93 cents), it is shown that this singer has a smaller range of vibrato extent than typical Western operatic singing.

The subjective listening test consists of 30 questions. Each question compares stimuli generated from three methods: (A) recorded singing voice with Tandem-STRAIGHT analysis and synthesis (no modification on singing F0 contour); (B) the proposed vibrato modeling (singing F0 contour with artificial vibrato); and (C) singing F0 contour with a constant F0 value during the instant when artificial vibrato takes place in method B. The constant F0 value in Method C is calculated as the mean value of the recorded singing voice during vibrato period. The spectral envelope and aperiodicity extracted from recorded singing voice remain unchanged for all methods. All stimuli contain one line of lyrics.

Listeners are asked to compare and rate the quality and 'singingness' of the three methods by mean opinion score (MOS). Possible MOS ranges from 1 (bad) to 5 (excellent) for quality. For 'singingness', the scale is also from 1 (weakest) to 5 (strongest). Listeners can play the stimuli as many times as they wish. There are 21 subjects participated in the test, contributing 30 × 21 = 630 responses.

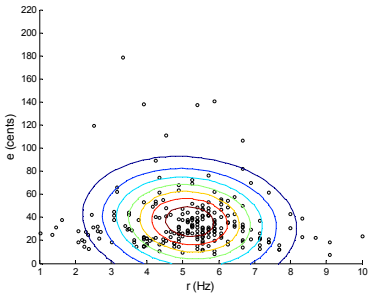


Figure 4: *Scatter plot of* $\{r, e\}$.

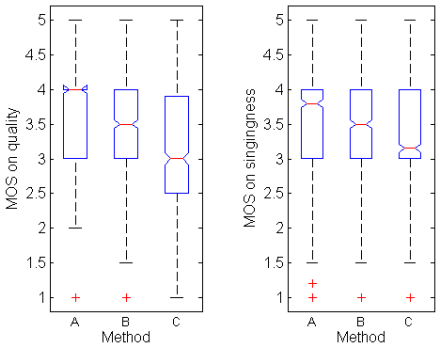


Figure 5: *Box plot of the MOS results. (left) on quality; (right) on 'singingness'.*

Figure 5 depicts the box plots of the MOS results on quality and 'singingness'. On each box, the central mark is the median. The edges are the 25th and 75th percentiles. Outliers are indicated by plus symbols. The experimental results show that the quality and 'singingness' of the proposed vibrato modeling (Method B) are significantly better than the one without vibrato (Method C) with 95% confidence. Compared with the recorded singing voice (Method A), the quality achieved by the proposed vibrato modeling is not as high as the recorded singing. For 'singingness', the proposed vibrato modeling does not bring stronger 'singingness' over the recorded singing. As vibrato is less often present in recorded singing voice, this indicates that there are other essential attributes contributing the feeling of 'singingness' besides vibrato.

## 4. Coherent processing of spectral envelope for vibrato

The following investigates if matched spectral envelope promotes the quality of vibrato-present singing voice or not.

### 4.1. Methods

*Spectral envelope processing*: As shown in Figure 1, the perceived spectral envelope is governed by the power distribution among the harmonics. In the proposed vibrato modeling, the spectral envelope remains unchanged, while the F0 fluctuates. This discrepancy is now handled by estimating the expected spectral envelope with the instantaneous F0. Let $U(k)$ be the power spectrum, in where the spectral envelope is implicitly stored. We first estimate the spectral envelope indicated by $U(k)$ by computing the autocorrelation sequence

$$R_u(m) = \frac{1}{K}\sum_k U(k) e^{\frac{2\pi jkm}{K}} \tag{6}$$

where $K$ is the number of FFT points (which is 4096 in our experiments). The spectral envelope $E(k)$ is found by estimating the $Q$-th order linear prediction coefficients $\{a_q, 1 \le q \le Q\}$ by the Levinson-Durbin recursion [16] and

$$E(k) = \left| G \Big/ \left( 1 + \sum_q a_q e^{-\frac{2\pi jkq}{K}} \right) \right|^2 \quad (7)$$

and $G$ is the scaling gain that matches the power levels of $U(k)$ and $E(k)$. Sample the power distribution at harmonic frequencies $\{bv(n), b = 1, 2, \dots\}$ as

$$E(f)\big|_{f=bv(n)} = \left| G \Big/ \left( 1 + \sum_q a_q e^{-\frac{2\pi jqbv(n)}{f_s}} \right) \right|^2 \quad (8)$$

where $f_s$ is the sampling frequency (48 kHz in our experiments). Finally, the expected spectral envelope matching the instantaneous F0 is found by interpolation with $E(f)|_{f=bv(n)}$.

*Phase-matching*: The phase of vibrato joining the F0 contour contributes to the singing voice quality, besides the vibrato waveform [8]. In (5), zero-phase vibrato is always used, regardless of the boundary phase of the original F0 segment. In the following, the optimal phase of $v(n)$ is found. Hence,

$$v(n) = m(n)\left[ (\hat{e} - 1)\sin\left( \frac{2\pi \bar{r} n}{f_{ps}} + \phi \right) + 1 \right] \quad (9)$$

where $\phi$ is the optimal phase defined as the one which achieves minimum summed pitch change occurring at the beginning and the end of vibrato. The resultant F0 contour will then be used to revise $U(k)$ as in the coherent processing of spectral envelope.

### 4.2. Experiments and results

Informal listening test indicated that the generated synthetic singing voice has similar quality as the recorded singing voice from Method A. Therefore, a pair-wise preference test (PPT) is conducted to evaluate the performance of the proposed spectral envelope processing on vibrato-present singing voice.

This PPT consists of nine questions. Similar to the subjective listening test in Section 3, all stimuli contain one line of lyrics. Vibrato is present in both the original and synthetic singing, such that listeners rate the quality of vibrato, rather than its presence. Listeners are asked to select which stimulus from the methods below achieves better quality. Method A is the recorded singing voice with Tandem-STRAIGHT analysis and synthesis. Method B* is the proposed vibrato modeling with spectral envelope processing and phase-matching. Listeners may select 'same quality' if the two stimuli are with equivalent quality. The aperiodicity extracted from the recorded singing voice is kept intact for both methods.

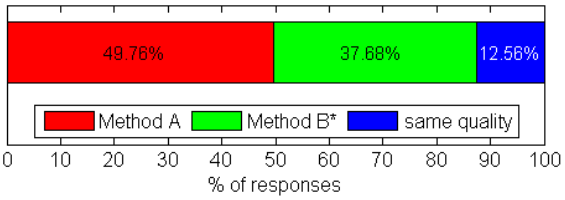


Figure 6: *Pair-wise preference test results on quality.*

There are 23 subjects participated in the PPT, contributing 9 × 23 = 207 responses. Figure 6 shows the PPT results. Half of the responses have indicated that the proposed method with vibrato modeling, spectral envelope processing and phase-matching (Method B*) generates better or same quality as the recorded singing voice (Method A). This shows the satisfactory performance of the proposed method in vibrato-present singing synthesis.

## 5. Conclusions

Vibrato plays an important role in singing voice synthesis. Nevertheless, independent manipulation of vibrato may damage the intrinsic relationships between vibrato and other attributes. This paper introduces a synthesis technique which jointly processes vibrato and the associated spectral envelope. Based on the evaluation results, it is shown that this joint processing generates singing voice with similar quality as human singing. Concerning 'singingness', by using singer-specific model, artificial vibrato is shown to bring stronger 'singingness' over constant pitch contour, but not human singing with less often vibrato. This indicates that there are other features greatly contribute to 'singingness'.

## 6. References


[1] *Synthesis of Singing Challenge (Special Session), Proc. Interspeech*, Aug. 2007.

[2] M. Akagi, "Rule-based voice conversion derived from expressive speech perception model: How do computers sing a song joyfully?" in *Proc. ISCSLP*, Tutorial 01, Nov. 2010.

[3] K. Saino, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "HMM-based singing voice synthesis system," in *Proc. Interspeech*, pp. 1141-1144, Sep. 2006.

[4] H. Kenmochi and H. Ohshita, "VOCALOID – Commercial singing synthesizer based on sample concatenation," in *Proc. Interspeech*, Aug. 2007.

[5] P. Kirn (2009, Oct. 6). iPhone Day: LaDiDa's Reverse Karaoke composes Accompaniment to Singing [Online]. Available: http://createdigitalmusic.com/2009/10/iphone-day-ladidas-reverse-karaoke-composes-accompaniment-to-singing/

[6] P. Hämäläinen, T. Mäki-Patola, V. Pulkki, and M. Airas, "Musical computer games played by singing," in *Proc. Int. Conference on Digital Audio Effects*, pp. 367-371, Oct. 2004.

[7] L. Regnier and G. Peeters, "Singing voice detection in music tracks using direct voice vibrato detection," in *Proc. ICASSP*, pp. 1685-1688, Apr. 2009.

[8] Y. Meron and K. Hirose, "Synthesis of vibrato singing," in *Proc. ICASSP*, pp. 745-748, Jun. 2000.

[9] T. Saitou, M. Goto, M. Unoki, and M. Akagi, "Speech-to-singing synthesis: Converting speaking voices to singing voices by controlling acoustic features unique to singing voices," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 215-218, Oct. 2007.

[10] T. L. Nwe, M. Dong, P. Chan, X. Wang, B. Ma, and H. Li, "Voice conversion: From spoken vowels to singing vowels," in *Proc. ICME AdMIRe Workshop*, pp. 1421-1426, Jul. 2010.

[11] J. Sundberg, *The Science of the Singing Voice*, Illinois: Northern Illinois University Press, 1987.

[12] R. Timmers and P. Desain, "Vibrato: Questions and answers from musicians and science," in *Proc. ICMPC*, Aug. 2000.

[13] T. Saitou and M. Goto, "Acoustic and perceptual effects of vocal training in amateur male singing," in *Proc. Interspeech*, pp. 832-835, Sep. 2009.

[14] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation," in *Proc. ICASSP*, pp. 3933-3936, Mar. 2008.

[15] H. -S. Pang, "On the use of the maximum likelihood estimation for analysis of vibrato tones," *Applied Acoustics*, vol. 65, pp. 101-107, Jan. 2004.

[16] J. Proakis and D. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*, 3rd ed. New Jersey: Prentice Hall, 1996.