SPEECH-TO-SINGING SYNTHESIS: CONVERTING SPEAKING VOICES TO SINGING VOICES BY CONTROLLING ACOUSTIC FEATURES UNIQUE TO SINGING VOICES

*Takeshi Saitou,** *Masataka Goto,*

National Institute of Advanced Industrial Science and Technology (AIST)
1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan
{saitou-t,m.goto}[at]aist.go.jp

Masashi Unoki, and Masato Akagi

School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
{unoki,akagi}[at]jaist.ac.jp

ABSTRACT

This paper describes a speech-to-singing synthesis system that can synthesize a singing voice, given a speaking voice reading the lyrics of a song and its musical score. The system is based on the speech manipulation system *STRAIGHT* and comprises three models controlling three acoustic features unique to singing voices: the fundamental frequency (F0), phoneme duration, and spectrum. Given the musical score and its tempo, the F0 control model generates the F0 contour of the singing voice by controlling four types of F0 fluctuations: overshoot, vibrato, preparation, and fine fluctuation. The duration control model lengthens the duration of each phoneme in the speaking voice by considering the duration of its musical note. The spectral control model converts the spectral envelope of the speaking voice into that of the singing voice by controlling both the singing formant and the amplitude modulation of formants in synchronization with vibrato. Experimental results show that the proposed system can convert speaking voices into singing voices whose naturalness is almost the same as actual singing voices.

1. INTRODUCTION

The goal of this research is to synthesize natural singing voices by controlling the acoustic features unique to them. Most previous research approaches [1, 2, 3] have focused on *text-to-singing (lyrics-to-singing) synthesis*, which generates a singing voice from scratch like speech is generated in text-to-speech synthesis. On the other hand, our approach focuses on *speech-to-singing synthesis*, which converts a speaking voice reading the lyrics of a song to a singing voice given its musical score. Research on the speech-to-singing synthesis is important for investigating the acoustic differences between speaking and singing voices. It will also be useful for developing practical applications for computer-based music productions where the pitch of singing voices is often manipulated (corrected or intentionally modified) [4] but their naturalness is sometimes degraded. Our research will make it possible to manipulate singing voices while keeping their naturalness. In addition, speech-to-singing synthesis itself is interesting for end users because even if the original speaker of a speaking voice is not good at singing, end users, including the speaker, can listen to the converted good singing voice having the speaker's voice timbre.

Although many studies have investigated the acoustic features unique to singing voices [5, 6] and their perceptual effects [7, 8, 9, 10], few have investigated the acoustic differences between speaking and singing voices [7, 11]. For example, by modifying (deteriorating) one of the two main acoustic features (the F0 contour [8, 10] and the spectrum [7]) of singing voices, the perceptual effect of each feature has been individually investigated, but there has been no comparison between those two features in terms of their perceptual contributions. Although Ohishi et al. [11] developed a method for automatically discriminating between speaking and singing voices, they did not attempt speech-to-singing synthesis. In our preliminary study [7], we found that a speaking voice could potentially be converted to a singing voice by manually controlling its three acoustic features: the F0, phoneme duration, and spectrum. In that work, we hand-tuned those control parameters by trial and error; there were no acoustic-feature control models except for the F0 control model [8]. In addition, the naturalness of the converted singing voice was not evaluated.

We therefore propose an automatic speech-to-singing synthesis system that integrates acoustic-feature control models for the F0, phoneme duration, and spectrum. Section 2 describes the three models having experimentally optimized control parameters. Section 3 shows experimental results indicating that converted singing voices are natural enough compared to actual singing voices and that the perceptual contribution of the F0 control is stronger than that of the spectral control. Finally, Section 4 summarizes the contributions of this research.

2. SPEECH-TO-SINGING SYNTHESIS SYSTEM

A block diagram of the proposed speech-to-singing synthesis system is shown in Fig 1. The system takes as the input a speaking voice reading the lyrics of a song, the musical score of a singing voice, and their synchronization information in which each phoneme of the speaking voice is manually segmented and associated with a musical note in the score. This system converts the speaking voice to the singing voice in six steps by: (1) decomposing the speaking voice into three acoustic parameters — F0 contour, spectral envelope, and aperiodicity index (AP) — estimated by using the analysis part of the speech manipulation system STRAIGHT [12]; (2) generating the continuous F0 contour of the singing voice from discrete musical notes by using the F0 control model; (3) lengthening the duration of each phoneme by using the duration control model; (4) modifying the spectral envelope and AP by using the spectral control model 1: (5) synthesizing the singing voice by using the synthesis part of the STRAIGHT; and (6) modifying the amplitude of the synthesized voice by using the spectral control model 2.

2.1. F0 control model

When converting a speaking voice to a singing voice, the F0 contour of the speaking voice is discarded and the target F0 contour

*This research was supported in part by CrestMuse, CREST, JST.

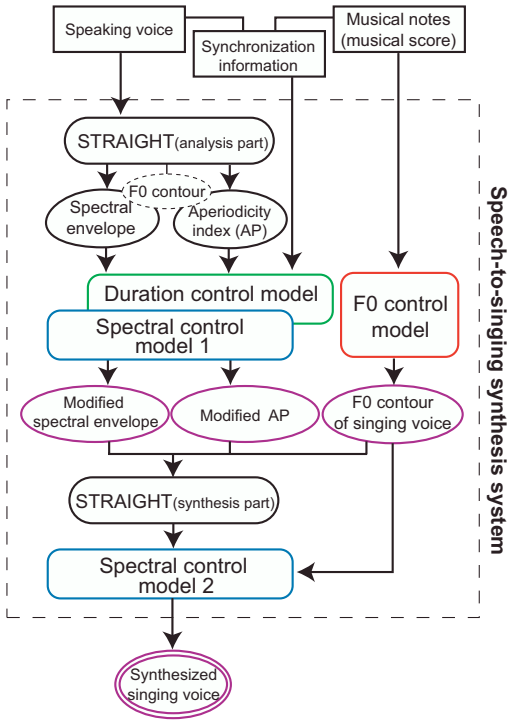

Figure 1: *Block diagram of the speech-to-singing synthesis system.*

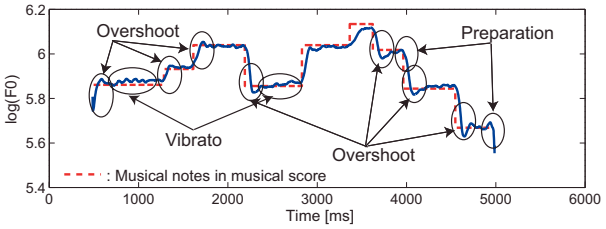

Figure 2: *Examples of F0 fluctuations in the singing voice of an amateur singer.*

of the singing voice is generated by using the musical notes of a song. The target F0 contour should have the following characteristics: (a) global F0 changes that correspond to the musical notes and (b) local F0 changes that include F0 fluctuations. There are four types of F0 fluctuations, which are defined as follows:

1. *Overshoot*: a deflection exceeding the target note after a note change [13].
2. *Vibrato*: a quasi-periodic frequency modulation (4-7 Hz) [14].
3. *Preparation*: a deflection in the direction opposite to a note change observed just before the note change.
4. *Fine fluctuation*: an irregular frequency fluctuation higher than 10 Hz [15].

Figure 2 shows examples of these fluctuations. Our previous study [8] confirmed that all of the above F0 fluctuations are contained in various singing voices and affect the naturalness of singing voices.

Figure 3 shows a block diagram of the proposed F0 control model [8]. This model can generate the target F0 contour by adding the four types of F0 fluctuations to a score-based melody contour, which is the input of this model as shown in Fig. 3. The melody contour is described by the sum of consecutive step functions, each corresponding to a musical note.

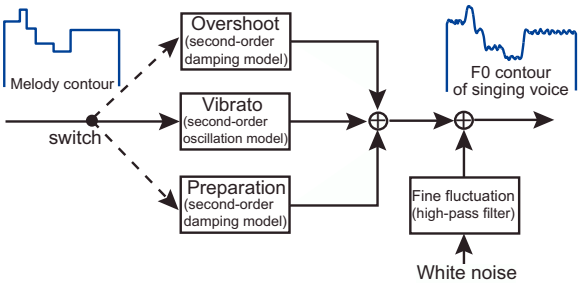

Figure 3: *Block diagram of the F0 control model for singing voices.*

The overshoot, vibrato, and preparation are added by using the transfer function of a second-order system represented as

$$H(s) = \frac{k}{s^2 + 2\zeta\omega s + \omega^2}, \tag{1}$$

where ω is the natural frequency, ζ is the damping coefficient and k is the proportional gain of the system. Here, the impulse response of $H(s)$ can be obtained as

$$h(t) = \begin{cases} \frac{k}{2\sqrt{\zeta^2-1}}(\exp(\lambda_1\omega t) - \exp(\lambda_2\omega t)), & |\zeta| > 1 \\ \frac{k}{\sqrt{1-\zeta^2}}\exp(-\zeta\omega t)\sin(\sqrt{1-\zeta^2}\omega t), & 0 < |\zeta| < 1 \\ kt\exp(-\omega t), & |\zeta| = 1 \\ \frac{k}{\omega}\sin(\omega t), & |\zeta| = 0 \end{cases} \tag{2}$$

where $\lambda_1 = -\zeta + \sqrt{\zeta^2 - 1}$, $\lambda_2 = -\zeta - \sqrt{\zeta^2 - 1}$. The above three fluctuations are represented by Eq. (2) as follows:
1. *Overshoot*: the second-order damping model ($0 < |\zeta| < 1$).
2. *Vibrato*: the second-order oscillation model ($|\zeta| = 0$).
3. *Preparation*: the second-order damping model ($0 < |\zeta| < 1$).

Characteristics of each F0 fluctuation are controlled by the system parameters ω, ζ, and k. In this study, the system parameters (ω, ζ, and k) were set to (0.0348 [rad/ms], 0.5422, 0.0348) for overshoot, (0.0345 [rad/ms], 0, 0.0018) for vibrato, and (0.0292 [rad/ms], 0.6681, 0.0292) for preparation. These parameter values were determined using the nonlinear least-squared-error method [16] to minimize errors between the generated F0 contours and actual ones.

The fine fluctuation is generated from white noise. The white noise is first high-pass-filtered and its amplitude is normalized. It is then added to the generated F0 contour having the other three F0 fluctuations. In this study, the cut off frequency of the high-pass filter was 10 Hz, its damping rate was -20 dB/oct, and the amplitude was normalized so that its maximum is 5 Hz.

2.2. Duration control model

Because the duration of each phoneme of the speaking voice is different from that of the singing voice, it should be lengthened or shortened according to the duration of the corresponding musical note. Note that each phoneme of the speaking voice is manually segmented and associated with a musical note in the score in advance. The duration of each phoneme is determined by the kind of musical note (e.g., crotchet or quaver) and the given local tempo.

Figure 4 shows a schema of the duration control model. This model assumes that each segmented boundary between a consonant and a succeeding vowel consists of a consecutive combination of a consonant part, a boundary part, and a vowel part. The boundary part occupies a region ranging from 10 ms before the boundary to 30 ms after the boundary, so its duration is 40 ms. The three parts are controlled as follows:

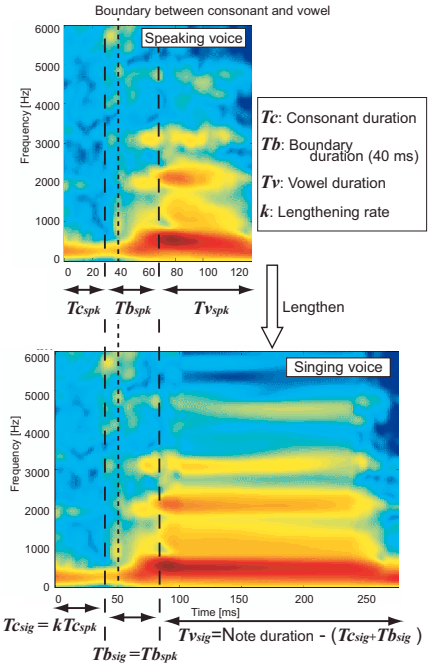

Figure 4: *Schema of the duration control model.*

1. The consonant part is lengthened according to fixed rates that were determined experimentally by comparing speaking and singing voices (1.58 for a fricative, 1.13 for a plosive, 2.07 for a semivowel, 1.77 for a nasal, and 1.13 for a /y/).
2. The boundary part is not lengthened.
3. The vowel part is lengthened so that the duration of the whole combination corresponds to the note duration.

2.3. Spectral control model

To generate the spectral envelope of the singing voice, the spectral envelope of the speaking voice is modified by controlling the spectral characteristics unique to singing voices as reported in the previous works [9, 17]. Sundberg [9] showed that the spectral envelope of a singing voice has a remarkable peak called the "singing formant" near 3 kHz. Oncley [17] reported that the formant amplitude of a singing voice is modulated in synchronization with the frequency modulation of each vibrato in the F0 contour. Figure 5 shows examples of the singing formant, and Fig. 6 shows an example where the formant amplitude in the lower panel as well as the amplitude envelope in the upper panel is modulated in synchronization with the frequency modulation of the F0 contour. Our previous study [7] also confirmed that these two types of acoustic features are contained in various kinds of singing voices and that they affect how a singing voice is perceived.

As shown in Fig. 1, the spectral envelope of the speaking voice is modified by two spectral control models (1 and 2) corresponding to the two acoustic features. The spectral control model 1 adds the singing formant to the speaking voice by emphasizing the peak of the spectral envelope and the dip of the aperiodicity index (AP) at about 3 kHz during vowel parts of the speaking voice. The peak of the spectral envelope can be emphasized by the following equation:

$$S_{sg}(f) = W_{sf}(f)S_{sp}(f), \tag{3}$$

where $S_{sp}(f)$ and $S_{sg}(f)$ are the spectral envelopes of the speaking and singing voices, respectively. $W_{sf}(f)$ is a weighting function for emphasizing the formant in $S_{sp}(f)$ and represented as

$$W_{sf}(f) = \begin{cases} (1+k_{sf})(1-\cos(2\pi\frac{f}{F_b+1})), & |f-F_s| \le \frac{F_b}{2} \\ 1, & \text{otherwise} \end{cases} \tag{4}$$

where F_s is the frequency of the peak in $S_{sp}(f)$ near 3 kHz, F_b is the bandwidth for the emphasis, and k_{sf} is the gain for adjusting the degree of emphasis. In this study F_b was set to 2000 Hz, and k_{sf} was set to emphasize F_s by 12 dB. These values were determined by analyzing the characteristics of singing formants in several singing voices [15]. The dip of AP can also be emphasized in the same way.

After synthesizing the singing voice, the spectral control model 2 adds the corresponding AM to the amplitude envelope of the synthesized singing voice. During each vibrato in the generated F0 contour, the AM is added as follows:

$$E_{sg}(t) = (1 + k_{am}\sin(2\pi f_{am}t))E_{sp}(t), \tag{5}$$

where $E_{sp}(f)$ and $E_{sg}(f)$ are the amplitude envelopes of the speaking and singing voices, respectively. f_{am} is the rate (frequency) of AM and k_{am} is the extent (amplitude) of AM. In this study, f_{am} and k_{am} were set to 5.5 Hz and 0.2, respectively. These values were determined by considering the characteristics of the vibrato generated by the F0 control model.

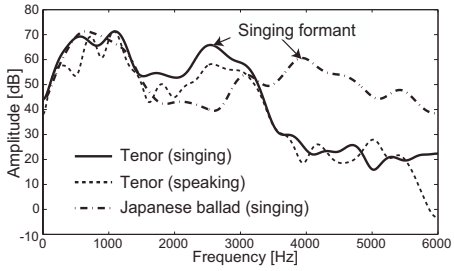

Figure 5: *Examples of singing formant near 3 kHz.*

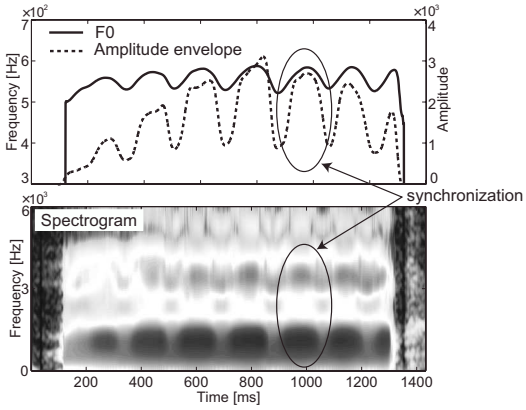

Figure 6: *Example of formant amplitude modulation (AM) in synchronization with vibrato of the F0.*

3. EVALUATION

We examined the performance of the proposed speech-to-singing synthesis system by evaluating the quality of synthesized singing voices in a psychoacoustics experiment. In this experiment, perceptual contributions of the F0 control and the spectral control were also investigated.

3.1. Singing voice conversion from speaking voice

Speaking voices taken as the input of the speech-to-singing synthesis system were recorded by letting two speakers (one female and one male) read the first phrase /karasunazenakuno/ of a Japanese children's song *"Nanatsunoko"*. The duration of each speaking voice was about 2 s. The speaking voices were digitized at 16 bit/48 kHz.

In addition to the original speaking voice and a reference singing voice provided by the same speaker, we prepared four different synthesized singing voices by disabling different control models:

SPEAK: speaking voice reading the phrase /karasunazenakuno/.

SING-BASE: singing voice synthesized using only the duration control model without the F0 and spectral control models (The F0 contour is the melody contour without any F0 fluctuations).

SING-F0: singing voice synthesized using the F0 and duration control models.

SING-SP: singing voice synthesized using the duration and spectral control models.

SING-ALL: singing voice synthesized using the proposed system with all the control models.

SING-REAL: real (actual) singing voice sung by the speaker of SPEAK.

3.2. Psychoacoustic experiment

Scheffe's method of paired comparison (Ura's modified method) [20] was used to evaluate the naturalness of synthesized singing voices. Ten subjects, all graduate students with normal hearing ability, listened to paired stimuli through a binaural headphone (Sennheiser HDA200) at a comfortable sound pressure level and rated the naturalness of the singing on a seven-step scale from "-3 (The former stimulus is very natural in comparison with the latter" to "+3 (The latter stimulus is very natural in comparison with the former". Paired stimuli having either female or male voices were randomly presented to each subject.

Figure 7 shows the experimental result. The numbers under the horizontal axis indicate the degree of the naturalness of the synthesized singing voices. The result of the F-test confirmed that there are significant differences amongst all stimuli at the 5% critical rate. This shows that the naturalness of the synthesized singing voices can be increased by controlling acoustic features unique to singing voices (by adding either the F0 or spectral control model; SING-F0 or SING-SP), and is almost the same with that of actual singing voices (SING-REAL) when using all the control models (SING-ALL). Moreover, the SING-F0 result was better than the SING-SP result, indicating that the perceptual contribution of the F0 fluctuations was greater than that of the spectral characteristics.

4. CONCLUSIONS

This paper proposed a speech-to-singing synthesis system that can convert speaking voices to singing voices by adding acoustic features of singing voices to the F0 contour and spectral envelope and lengthening each phoneme duration. The experimental results showed that our system can synthesize singing voices whose naturalness is close to that of actual singing voices and that the F0 fluctuations are more dominant acoustic cues than the spectral characteristics in the perception of singing voices.

Although we tested the system for conversion from speaking voices to singing voices, the same system can be used for conversion from singing voices to singing voices, i.e., changing the characteristics of singing voices that have already been recorded. Since our system can independently control the F0 control and spectral envelope, we plan to use it as a singing-voice effects processor for music productions. Future work will also include research on investigating acoustic features that affect perceptions of the singer's individuality and singing styles and extending our system to express them.

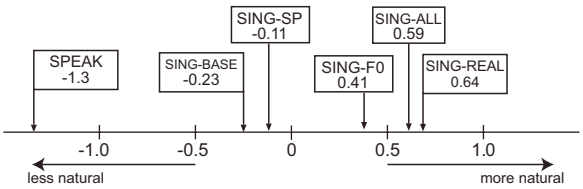

Figure 7: *Result of the psychoacoustic experiment: the degree of the naturalness of speaking voices (SPEAK), actual singing voices (SING-REAL), singing voices synthesized by our system (SING-ALL), and singing voices synthesized by disabling control models (SING-BASE, F0, and SP).*

5. REFERENCES

[1] J. Bonada, *et al*., "Synthesis of the Singing Voice by Performance Sampling and Spectral Models," IEEE Signal Processing Magazine, Vol. 24, Iss.2, pp. 67-79, 2007.

[2] YAMAHA Corporation, Vocaloid: New Singing Synthesis Technology, http://www.vocaloid.com/en/index.html

[3] K. Saino, *et al*., "HMM-based singing voice synthesis system," Proc. ICSLP06, pp.1141-1144, 2006.

[4] Antares Audio Technologies, Auto-Tune 5: Pitch Correcting Plug-In, http://www.antarestech.com/products/auto-tune5.shtml

[5] J. Sundberg, "The Science of Singing Voice," Northern Illinois University Press, 1987.

[6] A. B. Meribeth, "Dynamics of the Singing Voice," Springer, 1997.

[7] T. Saitou, *et al*., "Analysis of acoustic features affecting "singingness" and its application to singing voice synthesis from speaking voice," Proc. ICSLP2004, Vol. III, pp. 1929-1932, 2004.

[8] T. Saitou, *et al*., "Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis," Speech Commun., Vol. 46, pp. 405-417, 2005.

[9] J. Sundberg, "Articulatory Interpretation of the 'Singing Formant'," J. Acoust. Soc. Am., Vol. 55, pp. 838-844, 1974.

[10] H. B. Rothman, *et al*., "Acoustic variability in vibrato and its perceptual significance," J. Voice, Vol.1, no.2, pp.123-141, 1987.

[11] Y. Ohishi, *et al*., "On the human capability and acoustic cues for discriminating the singing and the speaking voices," Proc. ICMPC2006, pp. 1831-1837, 2006.

[12] H. Kawahara, *et al*., "Restructuring speech representations using a pitch adaptive time-frequency smoothing and an instantaneous- frequency based on F0 extraction: Possible role of a repetitive structure in sounds," Speech Commun., Vol. 27, pp. 187-207, 1999.

[13] G. de Krom, *et al*., "Timing and accuracy of fundamental frequency changes in singing," Proc. ICPhS 95, Vol. I, pp. 206-209, 1995.

[14] C. E. Seashore, "The Vibrato," University of Iowa Studies in the Psychology of Music, Vol. I, 1932.

[15] M. Akagi, *et al*., "Perception of synthesized singing-voices with fine-fluctuations in their fundamental frequency fluctuations," Proc. ICSLP2000, Vol. 3, pp.458-461, 2000.

[16] W. H. Press, *et al*., "Numerical Recipes in C," Cambridge University Press, Cambridge, 1988.

[17] P. B. Oncley, "Frequency, Amplitude, and Waveform Modulation in the Vocal Vibrato," J. Acoust. Soc. Am., Vol. 49, Issue 1A pp. 136, 1971.

[18] I, Nakayama, "Comparative Studies on Vocal Expression in Japanese Traditional and Western Calssical-style Singing, Using a Common Verse," Proc. ICA, Mo4. C1. 1, 2004.

[19] N. Minematsu, *et al*., "Prosodic Modeling of Nagauta Singing and Its Evaluation," Proc. Speech Prosody 2004, pp. 487-490, 2004.

[20] S. Ura, *et al*., "Sensory Evaluation Handbook (in Japanese)," JUSE Press Ltd., pp. 366-384, 1973.