Speech-to-Singing Synthesis System: Vocal Conversion from Speaking Voices to Singing Voices by Controlling Acoustic Features Unique to Singing Voices *

Takeshi SAITOU[1], Masataka GOTO[1], Masashi UNOKI[2], and Masato AKAGI[2]
(1. National Institute of Advanced Industrial Science and Technology (AIST), 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan; 2. School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan)

Introduction: This paper introduces a speech-to-singing synthesis system, called *SingBySpeaking*, which 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 is comprised of four models controlling three acoustic parameters: 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 phoneme in the speaking voice by taking into consideration 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. SingBySpeaking enables us to synthesize natural singing voices merely by reading the lyrics of a song and to better understand differences between speaking and singing voices.

Key word: singing voice synthesis; STRAIGHT; vocal conversion; singing voice perception

1. Introduction

Singing songs is one of the most familiar ways of enjoying music, simultaneously being an important way of expressing both linguistic and nonlinguistic information in human communication. Research on singing voice synthesis is therefore not only important for developing practical music applications but also for understanding the mechanism underlying the perception and production of human singing voices.

For decades, many research studies on singing voice synthesis have been done to produce operatic singing voices. These traditional studies have been based on several approaches, such as vocal tract physical models and formant-based methods of synthesis, and its aims have been to understand the acoustic characteristics of operatic singing voices and the mechanism underlying the production of operatic singing [1, 2]. Recently, many research approaches [3–5] have focused on *text-to-singing (lyrics-to-singing)* synthesis, which generates a singing voice from scratch just like speech is generated in text-to-speech synthesis. Since most of these synthesis systems have been based on corpus-based methods, such as wave concatenation synthesis and hidden Markov model (HMM) synthesis, they have been more practical than traditional systems. Vocaloid2 [4], for example, has easily enabled end users to produce synthesized singing voices.

We, on the other hand, have pursued research on constructing a system to synthesize singing voices based on an approach that converts a speaking voice to a singing voice. We called this approach *speech-to-singing synthesis*. Through research on speech-to-singing synthesis, we have aimed at understanding the perceptual mechanism unique to the singing voice by investigating differences between singing and speaking voices and at constructing novel singing voice synthesis applications enabling end users to produce and listen to their own singing voice merely by reading the lyrics of songs.

*This research was supported in part by CrestMuse, CREST, JST.
Author information: Takeshi SAITOU, Ph.D.（1977-），male（Japanese）。Post-doctoral research scientist
Corresponding author: Takeshi SAITOU, *E-mail address*: saitou-t[at]aist.go.jp

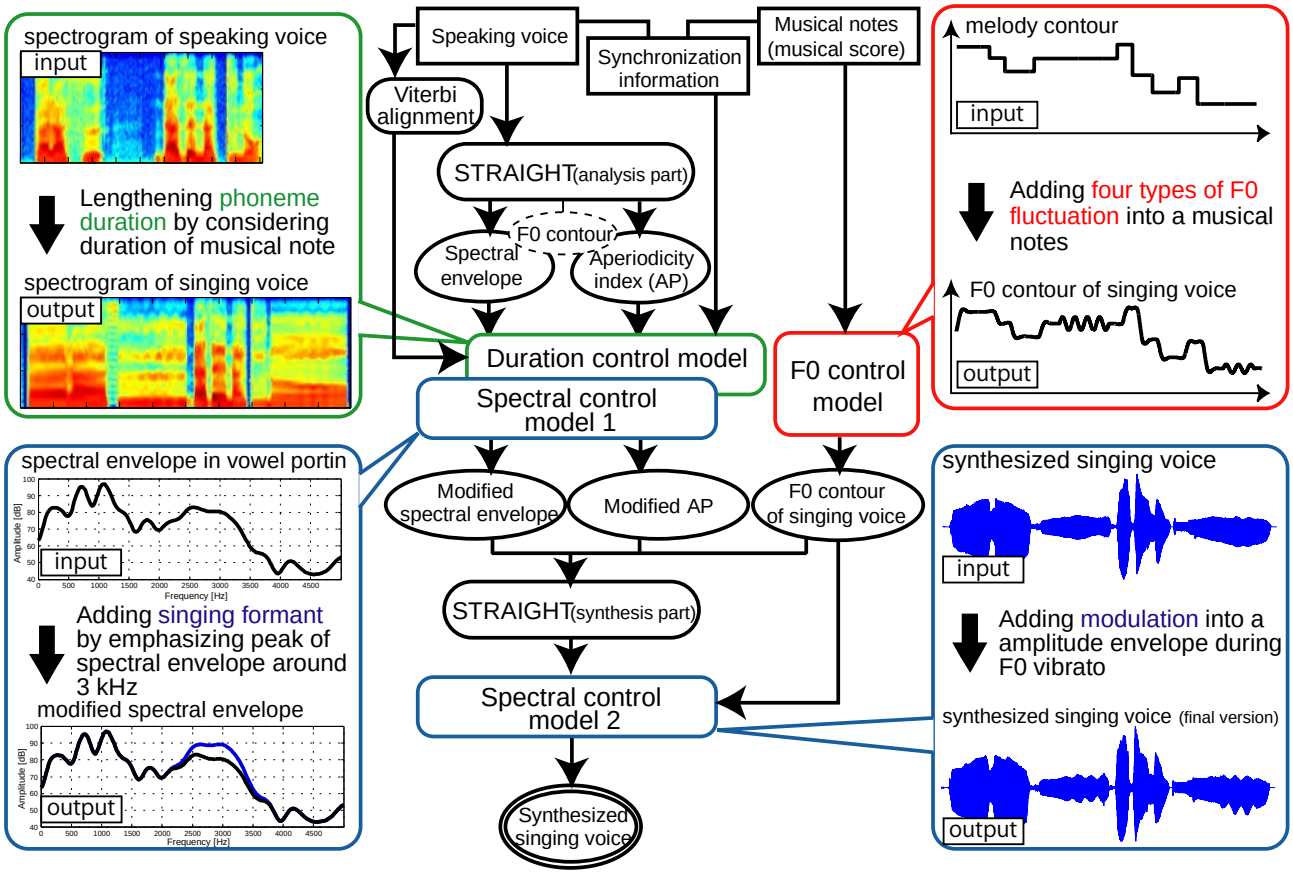

Figure 1: ***Block diagram of SingBySpeaking and examples of processes in four control models.***

This paper introduces a speech-to-singing synthesis system, called *SingBySpeaking*, which we have developed since 2004 [6–10]. SingBySpeaking, as shown in Fig. 1, can synthesize a singing voice, given a speaking voice reading the lyrics of a song and its musical score. This system is based on the speech manipulation system *STRAIGHT* [11] and is comprised of four models controlling acoustic features unique to singing voices in three acoustic parameters: the fundamental frequency (F0), phoneme duration, and spectrum. This paper also introduces the acoustic features, and models for controlling these features in this paper.

2. Outline of SingBySpeaking

Figure 1 overviews SingBySpeaking. 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 automatically associated with a musical note in the score. This system converts the speaking voice to a singing voice in six steps by:

(1) Decomposing the speaking voice into three acoustic parameters – the F0 contour, spectral envelope, and aperiodicity index (AP) – estimated by using the analysis component of the speech manipulation system STRAIGHT,

(2) Generating the continuous F0 contour of the singing voice from discrete musical notes by using the F0 control model,

(3) Segmenting the speaking voice into phonemes by using Viterbi alignment method with a phoneme-level HMM model, and then lengthening the duration of each phoneme by using the duration control model,

(4) Modifying the spectral envelope and AP by using spectral control model 1,

(5) Synthesizing the singing voice by using the synthesis component of STRAIGHT, and

(6) Modifying the amplitude of the synthesized voice by using spectral control model 2.

3. F0 characteristics and its control model

3.1. F0 fluctuations

It is well known that the F0 contours of singing voices have two characteristics: (a) global F0 changes that correspond to the musical notes and (b) local F0 changes that include F0 fluctuations unique to singing

voices. There are four types of F0 fluctuations, which are defined as:

Overshoot: A deflection exceeding the target note after a note change [6, 7, 12].

Vibrato: A quasi-periodic frequency modulation (4–7 Hz) [13].

Preparation: A deflection in the direction opposite to a note change observed just before the note change [6, 7].

Fine fluctuation: An irregular frequency fluctuation higher than 10 Hz [14].

Figure 2 shows examples of these fluctuations. Our previous study [6, 7] confirmed that all four F0 fluctuations were contained in various singing voices.

3.2. F0 control model

When converting a speaking voice into a singing voice with SingBySpeaking, the F0 contour of the speaking voice is discarded and the target F0 contour of the singing voice is generated by the F0 control model [6, 7]. This model, as shown in Fig. 1, can generate the target F0 contour by adding the four F0 fluctuations to a score-based melody contour. The melody contour is described by the sum of consecutive step functions, each corresponding to a musical note.

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

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

where ω is the natural frequency, ζ is the damping coefficient, and k is the proportional gain of the system. Overshoot and preparation are represented with a 2nd order damping model, and vibrato is represented with a 2nd order oscillation (no-loss) model. The characteristics of each F0 fluctuation are controlled by system parameters ω, ζ, and k. When generating the F0 contour of the singing voice, the system parameters (ω, ζ, and k) are set to (0.0348 [rad/ms], 0.5422, and 0.0348) for overshoot, (0.0345 [rad/ms], 0, and 0.0018) for vibrato, and (0.0292 [rad/ms], 0.6681, and 0.0292) for preparation. Note that the characteristics of each fluctuation can be controlled by changing these three system parameters.

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. The cut-off frequency of the high-pass filter was 10 Hz, its damping rate was -20 dB/oct, and its amplitude was normalized to a maximum of 5 Hz.

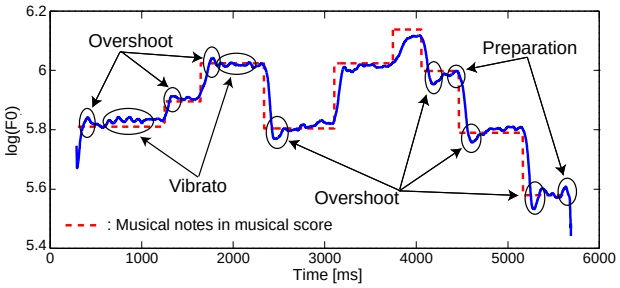

Figure 2: Examples of F0 fluctuations in singing voice of amateur singer.

4. Duration characteristics and its control model

Because the duration of all phonemes in the speaking voice differs from that in the singing voice, it should be lengthened or shortened according to the duration of corresponding musical notes. The duration of each phoneme is determined by the kind of musical note (e.g., crotchet or quaver) and the given local tempo.

Figure 3 shows a schema of the duration control model, which assumes that each boundary between a consonant and a succeeding vowel consists of a consecutive combination of a consonant part, a boundary part, and a vowel part. Note that the boundary is automatically segmented by using Viterbi alignment method. As the boundary part occupies a region ranging from 10 ms before the boundary to 30 ms after the boundary, its duration is 40 ms. The three parts are controlled in three ways:

- 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/).
- The boundary part is not lengthened.
- The vowel part is lengthened so that the duration of the whole combination corresponds to the note duration.

5. Spectral characteristics and its control model

5.1. Spectral characteristics

Two typical spectral characteristics unique to singing voices have been reported in previous studies.

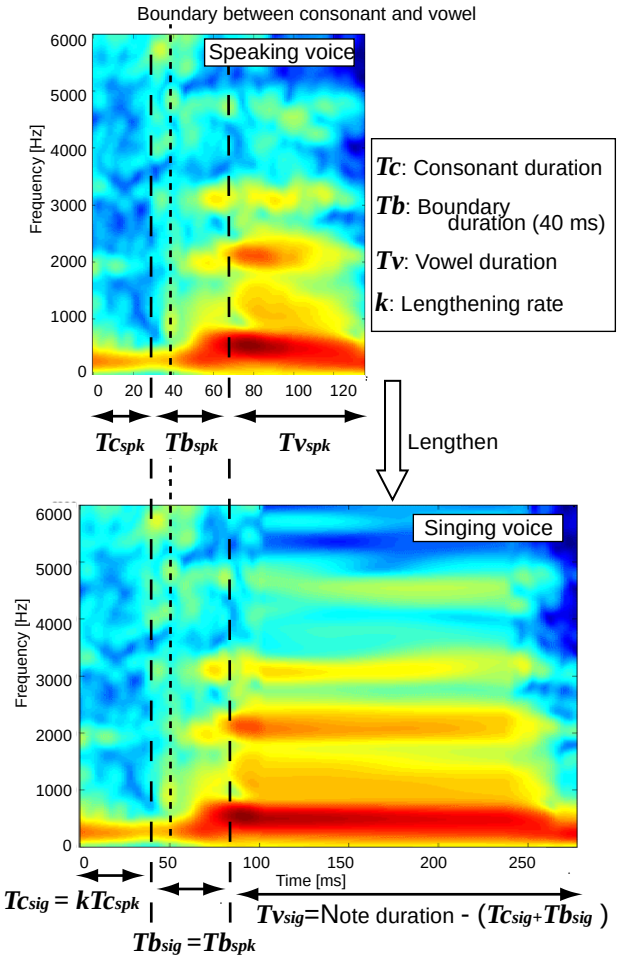

Figure 3: ***Schema for duration control model***.

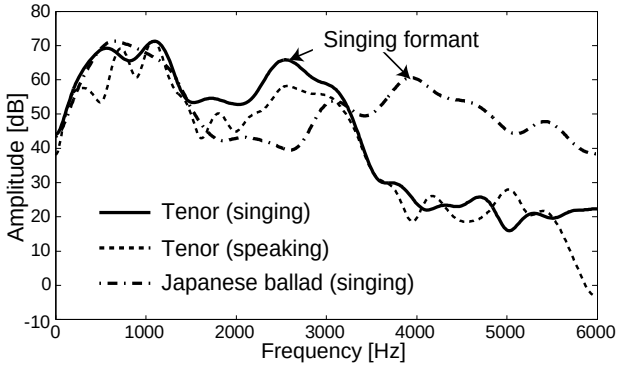

Figure 4: ***Examples of singing formant near 3 kHz in operatic singing and traditional Japanese singing.***

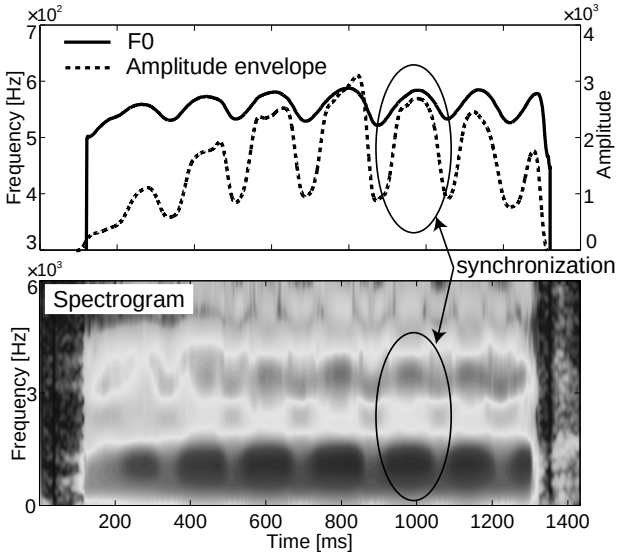

Figure 5: ***Example of formant amplitude modulation (AM) in synchronization with vibrato of F0.***

Sundberg [15] found that the spectral envelope of a singing voice has a remarkable peak called the ``singing formant" near 3 kHz. Nakayama[16] also discovered singing formant in traditional Japanese singing. Oncley [17] reported that the formant amplitude of a singing voice was modulated in synchronization with the frequency modulation of each vibrato in the F0 contour. Figure 4 shows examples of the singing formant, and Fig. 5 shows an example where the formant amplitude in the lower panel as well as the amplitude envelope in the upper panel are modulated in synchronization with the frequency modulation of the F0 contour. Our previous study [8, 9] also confirmed that these two types of acoustic features were contained in various kinds of singing voices and that they affected singing voice perception.

5.2. Spectral control models

As seen in Fig. 1, the spectral envelope of the speaking voice is modified by two spectral control models (1 and 2) corresponding to the two spectral characteristics.

Spectral control model 1 adds singing formant tothe speaking voice by emphasizing the peak of the spectral envelope at about 3 kHz during vowel parts in the speaking voice. The bandwidth of the spectral envelope for emphasis is 2000 Hz, and the gain used for adjusting the degree of emphasis is 12 dB. These values were determined by analyzing the characteristics of singing formants in several singing voices [8, 9]. The dip in AP at about 3 kHz during the vowel part, at the same time, is emphasized in the same way.

After synthesizing the singing voice, spectral control model 2 adds the corresponding amplitude modulation (AM) to the amplitude envelope of the synthesized singing voice. As shown in Fig. 1, the AM is added to the amplitude envelope during each vibrato in the generated F0 contour. The rate (modulation frequency) of AM is set to 5.5 Hz as the same as that of the vibrato in the generated F0 contour.

6. Performance of SingBySpeaking

We assessed the performance of SingBySpeaking by evaluating the quality of synthesized singing voices in a psychoacoustics experiment, where perceptual contributions of F0 control and spectral control models were also investigated.

6.1. Singing voice synthesis

Speaking voices taken as the input for SingBySpeaking 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.

Figure 6 shows the waveform, F0 contour, and spectrogram of the male speaking voice and SING-ALL.

6.2. Psychoacoustic experiment

Scheffe's method of paired comparison (Ura's modified method) [18] was used to evaluate the naturalness of the synthesized singing voices. Ten subjects, all graduate students with normal hearing ability, listened to paired stimuli through a binaural headphone at a comfortable sound pressure level and rated the naturalness of the synthesized singing voices on a seven-step scale from “-3” (The former stimulus was very natural in comparison with the latter) to “+3” (The latter stimulus was very natural in comparison with the former). Paired stimuli having either female or male voices were randomly presented to each subject.

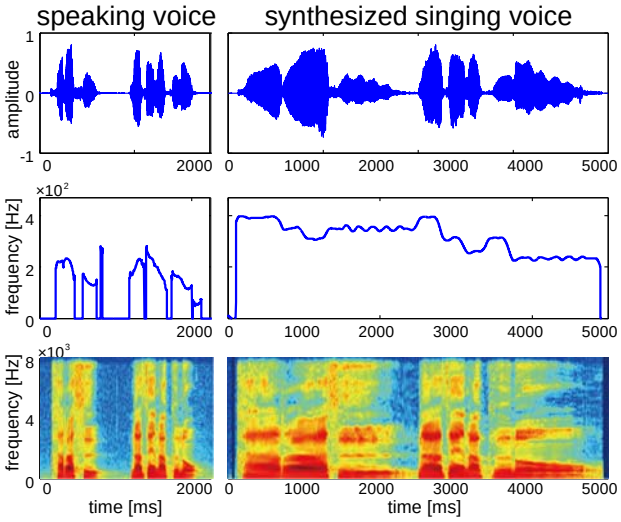

Figure 6: ***Acoustic parameters of male speaking voice and synthesized singing voice.***

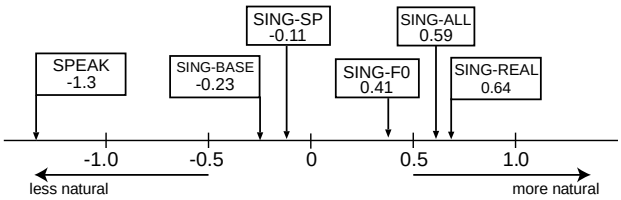

Figure 7: ***Results from psychoacoustic experiment: Degree of naturalness of speaking voices (SPEAK), actual singing voices (SING-REAL), singing voices synthesized by our system (SING-ALL), and singing voices synthesized by disabling control models (SING-BASE, F0, and SP).***

Figure 7 shows the experimental results. The numbers under the horizontal axis indicate the degree of naturalness of the synthesized singing voices. The results of the F-test confirmed that there were significant differences amongst all stimuli at the 5 % critical rate. This means that the naturalness of the synthesized singing voices could 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 this was almost the same as that of actual singing voices (SING-REAL) when using all the control models (SING-ALL). The results demonstrate that SingBySpeaking can synthesize natural and human-like singing voices. Moreover, the SING-F0 result was better than the SING-SP result, indicating that the perceptual effects of F0 fluctuations were greater than those of the spectral characteristics. These results indicate that acoustic features unique to singing voices are important acoustic cues not only for perceiving singing voices but also for discriminating singing and speaking voices.

7. Conclusion

This paper introduced a speech-to-singing synthesis system, called SingBySpeaking, that can convert speaking voices to singing voices by adding acoustic features unique to singing voices to the F0 contour and spectral envelope and lengthening the duration of each phoneme. The evaluation results revealed that SingBySpeaking made it possible to synthesize singing voices whose naturalness was close to that of actual singing voices and that the F0 fluctuations were more dominant acoustic cues than the spectral characteristics in the perception of singing voices.

The contributions made by SingBySpeaking demonstrate the potential of this system, which can be applied not only to constructing novel application of singing voice synthesis but also to investigating the mechanism underlying the perception and production of singing voices. We intend to investigate acoustic features that affect perceptions of a singer's individuality and singing style in the future and extending SingBySpeaking to express these.

Acknowledgements

We thank Ken-Ichi Sakakibara for many useful comments and invaluable advice.

References

[1] P. R. Cook, ``Identification of Control Parameters in an Articulatory Vocal Tract Model, With Applications to the Synthesis of Singing," Ph.D. Thesis, Stanford Univ.1991.

[2] J. Sundberg, ``The KTH synthesis of singing," Adv. Cognit. Psychol. (*Special issue on music performance*), 2(2–3), pp. 131–143, 2006.

[3] J. Bonada and X. Serra, ``Synthesis of the Singing Voice by Performance Sampling and Spectral Models," IEEE Signal Processing Magazine, Vol. 24, Iss. 2, pp. 67–79, 2007.

[4] H. Kenmochi and H. Ohshita, ``VOCALOID–Commercial Singing Synthesizer Based on Sample Concatenation," Proc. INTERSPEECH 2007, pp. 4011–4010, 2007.

[5] K. Saino, H. Zen, Y. Nankaku, A. Lee and K. Tokuda, ``HMM-based singing voice synthesis system," Proc. ICSLP06, pp. 1141–1144, 2006.

[6] T. Saitou, M. Unoki and M. Akagi, ``Development of the F0 Control Model for Singing-Voice Synthesis," Proc. Speech Prosody 2004, pp. 491–494, 2004.

[7] T. Saitou, M. Unoki and M. Akagi, ``Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis," Speech Commun., Vol. 46, pp. 405–417, 2005.

[8] T. Saitou, M. Unoki and M. Akagi, ``Analysis of acoustic features affecting ``singing-ness" and its application to singing voice synthesis from speaking voice," Proc. ICSLP2004, Vol. III, pp. 1929–1932, 2004.

[9] T. Saitou, N. Tsuji, M. Unoki and M. Akagi, ``Analysis of proper acoustic features to singing voice based on a perceptual model of “singing-ness”," J. Acoust. Soc. Jpn., Vol. 64, No. 5, pp. 267–277, 2008 (in Japanese).

[10] T. Saitou, M. Goto, M. Unoki and M. Akagi, ``Speech-to-Singing Synthesis: Vocal conversion from speaking voices to singing voices by controlling acoustic features unique to singing voices," Proc. Proc. 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA2007), pp. 215–218, 2007.

[11] H. Kawahara, I. Masuda-Katsuse, A. de Cheveigne, ``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.

[12] H. Mori, W. Odagiri and H. Kasuya, `` F0 Dynamics in Singing: Evidence from the Data of a Baritone Singer," IEICE Trans. Inf. & Syst., Vol. E87-D, No. 5, pp. 1086–1092, 2004.

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

[14] M. Akagi, H. Kitakaze, ``Perception of synthesized singing-voices with fine-fluctuations in their fundamental frequency fluctuations," Proc. ICSLP2000, Vol. 3, pp. 458–461, 2000.

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

[16] I. Nakayama, ``Comparative studies on vocal expression in Japanese traditional and western classical-style singing, using a common verse," Proc. ICA2004, pp. 1295–1296, 2004.

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

[18] S. Ura, ``Sensory Evaluation Handbook," JUSE Press Ltd., pp. 366-384, 1973 (in Japanese).