Vocal Conversion from Speaking Voice to Singing Voice Using STRAIGHT

Takeshi Saitou [1], *Masataka Goto* [1], *Masashi Unoki* [2], and *Masato Akagi* [2]

[1] National Institute of Advanced Industrial Science and Technology (AIST)
[2] School of Information Science, Japan Advanced Institute of Science and Technology
{saitou-t,m.goto} [at] aist.go.jp, {unoki,akagi} [at] jaist.ac.jp

Abstract

A vocal conversion system that can synthesize a singing voice given a speaking voice and a musical score is proposed. It is based on the speech manipulation system *STRAIGHT* [1], and comprises three models controlling three acoustic features unique to singing voices: the F0, duration, and spectral envelope. Given the musical score and its tempo, the F0 control model generates the F0 contour of the singing voice by controlling four 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 showed that the proposed system could convert speaking voices into singing voices whose quality resembles that of actual singing voices.

1. Vocal conversion system

A block diagram of the proposed vocal conversion system is shown in Fig. 1. The system takes as the input a speaking voice of 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 synthesizes the singing voice in five steps: (1) decompose the speaking voice into three acoustic parameters − F0 contour, spectral envelope, and aperiodicity index (AP) − estimated by STRAIGHT (analysis part), (2) generate the continuous F0 contour of the singing voice from discrete musical notes by using the F0 control model, (3) lengthen the duration of each phoneme by using the duration control model, (4) modify the spectral envelope and AP by using the spectral control model 1, (5) synthesize the singing voice by using STRAIGHT (synthesis part), and (6) modify the amplitude of the synthesized voice by using the spectral control model 2.

2. F0 control model

Figure 2 shows a block diagram of the proposed F0 control model [2] that generates the F0 contour of the singing voice by adding F0 fluctuations to musical notes. Our model can deal with four types of dynamic F0 fluctuations: (1) overshoot, which is a deflection exceeding the target note after a note change [3]; (2) vibrato, which is a quasi-periodic frequency modulation (4–7 Hz) [4]; (3) preparation, which is a deflection in the direction opposite to a note change observed just before the note change; and (4) fine fluctuation, which is an irregular frequency fluctuation higher than 10 Hz [5]. Figure 3 shows examples of F0 fluctuations. Our
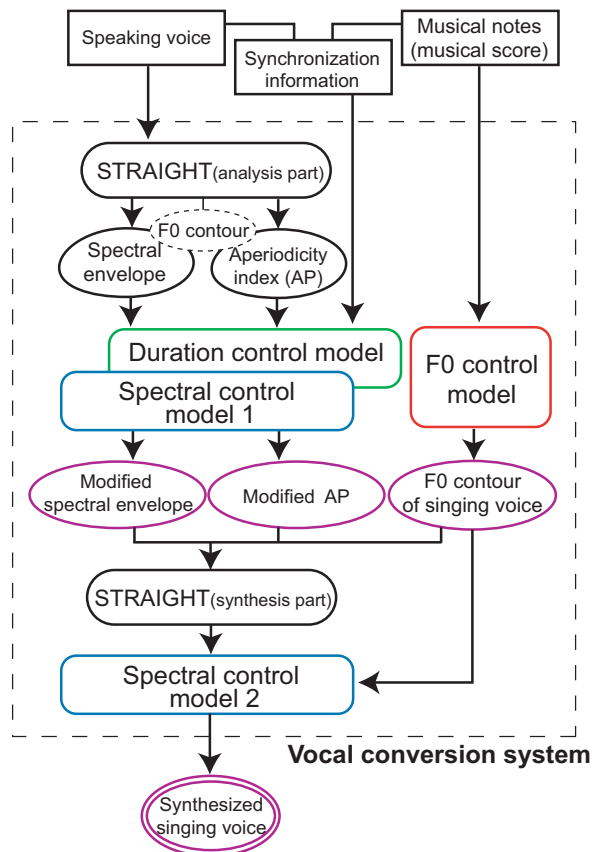

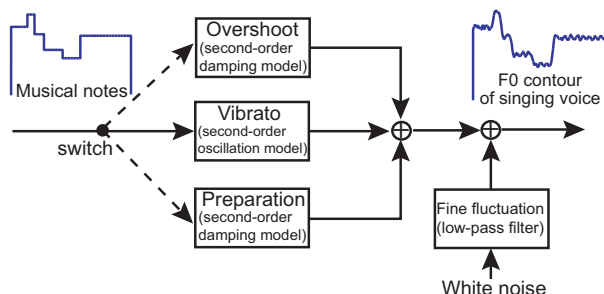Figure 1: *Block diagram of the vocal conversion system.*

Figure 2: *Block diagram of the F0 control model for singing voices.*

psychoacoustic experiments confirmed that each F0 fluctuation affects the perceptual quality of singing voices [2].

3. Duration control model

The duration control model (Fig. 4) 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. The boundary part is from 10 ms before the boundary to 30 ms after the boundary, so its duration is 40 ms. The parts are controlled as follows: (1) the consonant part is lengthened according to fixed rates that were determined experimentally by comparing speaking and singing voices (1.28 for a fricative, 1.00 for a plosive, 2.37 for a semivowel, 1.43 for a nasal, and 1.22 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.

4. Spectral control model

The spectral envelope of the speaking voice is modified by two models as shown in Fig. 1. These models can respectively control two different acoustic features that are frequently contained in vowels sung by professional singers: a singing formant [6], which is a remarkable peak of the spectral envelope near 3 kHz; and a formant amplitude modulation (AM) synchronized with the frequency modulation of each vibrato in the generated F0 contour. The spectral control model 1 controls the singing formant by emphasizing a peak of the spectral envelope or a dip of the AP at about 3 kHz during vowel parts of the speaking voice. During each vibrato, the spectral control model 2 adds the corresponding AM to the amplitude envelope of the synthesized singing voice. Our psychoacoustic experiments also confirmed that each acoustic feature affects the perceptual quality of singing voices [7].

Acknowledgements

References

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

[2] Saitou, T. *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.

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

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

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

[6] Sundberg, J., "Articulatory interpretation of the 'singing formant'," J. Acoust. Soc. Am., Vol. 55, pp. 838–844, 1974.

[7] Saitou, T. *et al.*, "Analysis of acoustic features affecting "singing-ness" and its application to singing voice synthesis from speaking voice," Proc. ICSLP04, Vol. III, pp. 1929–1932, 2004.
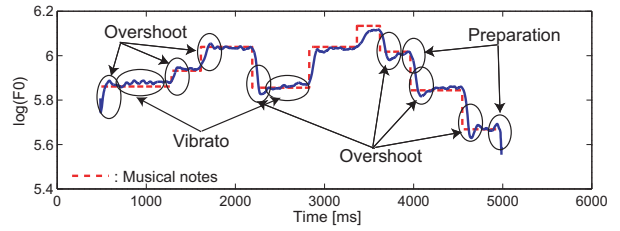

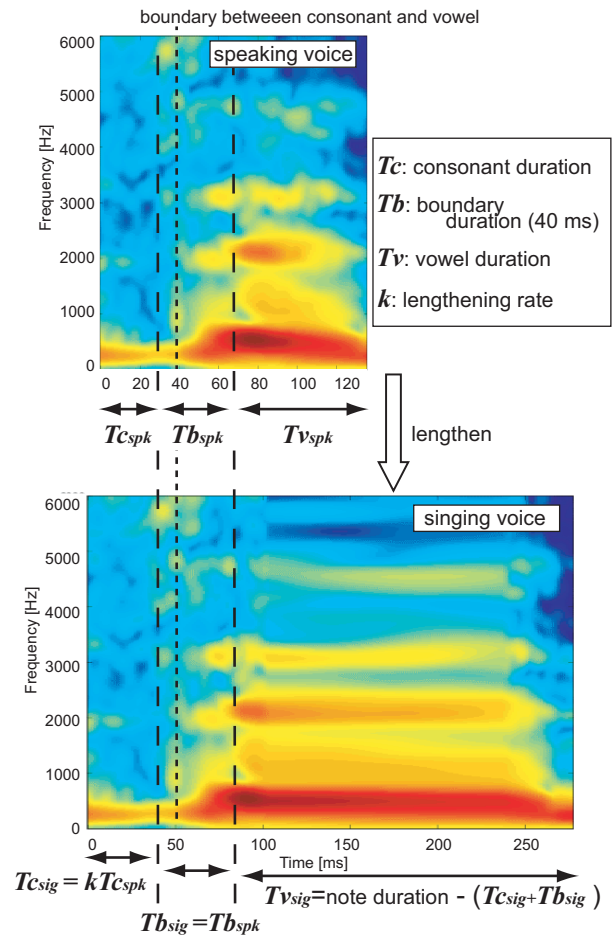Figure 3: *Examples of F0 fluctuations in the singing voice of an amateur singer.*

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