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.

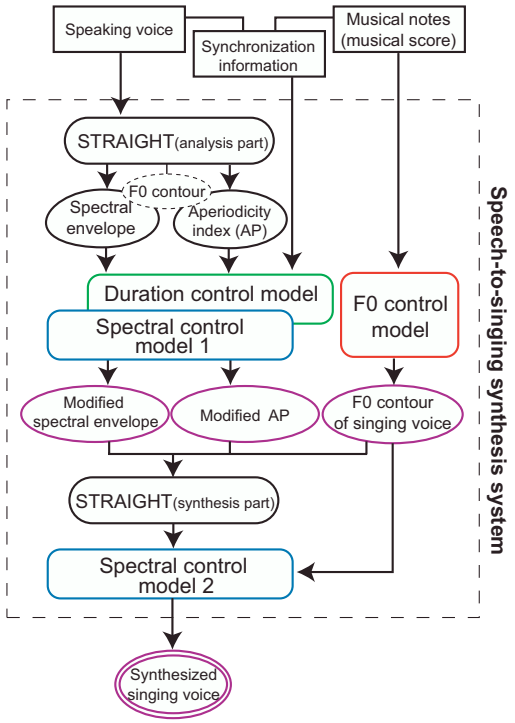

Figure 1: *Block diagram of the speech-to-singing synthesis system.*

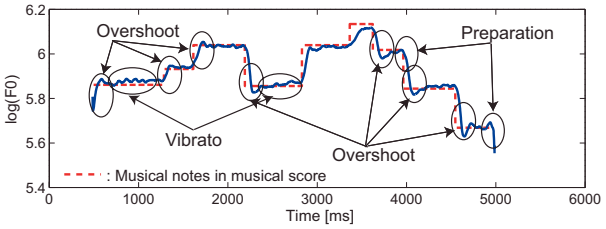

Figure 2: *Examples of F0 fluctuations in the singing voice of an amateur singer.*

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.

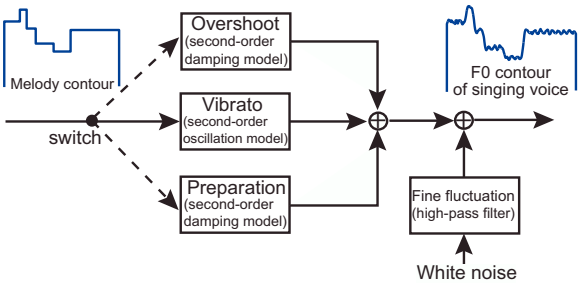

Figure 3: *Block diagram of the F0 control model for singing voices.*

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: