Control Methods of Acoustic Parameters for Singing-Voice Synthesis

Takeshi Saitou, Masashi Unoki, and Masato Akagi

School of Information Science, Japan Advanced Institute of Science and Technology
{t-saitou; unoki; akagi;}@jaist.ac.jp

Abstract
To construct a natural singing voice synthesis system, it is important to adequately control acoustic parameters such as fundamental frequency (F0), phoneme duration, and spectrum information in the synthesis method, based on comparative analysis between spoken- and singing-voices. This paper proposes a transformation from read speech into singing-voice using STRAIGHT. This method is composed of an F0 control model, a duration control method for each type of phoneme, and a spectrum stretching method for controlling spectral shapes in read speech. Results showed that the proposed system can synthesize a natural singing-voice, whose sound quality is almost the same as that of sung voice.

1. Introduction
Singing a song is an important way in human communications to express linguistic and emotional information. It is an important issue to investigate how singing-voices are perceived and generated, as a part of studies of non-linguistic information in speech sounds. However, since most speech synthesis methods were not proposed for singing-voice synthesis but for spoken-voice synthesis, it is impossible to reveal how singing-voices are perceived and generated by using these methods. This paper aims to show the possibility of synthesizing a singing-voice by adding non-linguistic information to read speech.

In singing a song, a person makes an effort to express lyrics by changing notes corresponding to melody without changing one’s own tone color. Moreover, singing-voices have more dynamic and complicated characteristics than those of speaking-voices, and these characteristics are significant factors in the naturalness of singing-voices. Therefore, to transform from read speech into a singing-voice, the following points must be considered.
- how to control F0 based on melody
- how to control phoneme duration based on rhythm
- how to control spectrum based on F0 variation

It is well known that F0-Contours of singing-voices have the following three characteristics with regard to F0 fluctuations [2, 3]: (a) the dynamic range of the F0 contours is wider than that of spoken-voices; (b) a steady state F0 contour corresponds to a note, and the note changes of the F0 contours correspond to melody; and (c) there are many F0 fluctuations that are observed only in singing-voices. These characteristics are peculiar to singing-voices. Therefore, an F0 control model is required to be able to cope with these F0 characteristics.

With regard to phoneme duration, it is shown that, in a typical Japanese song, one or more notes are allocated in one mora. Note duration is determined by the duration peculiar to the note (e.g., crotchet or quaver) and the tempo of the song. However, the duration of one mora is almost the same in all Japanese spoken-voices. Therefore, to translate from read speech into a singing-voice, a duration control model is required to be able to stretch phoneme duration related to note duration.

With regard to spectrum information, there are two useful reports. Slawson [4] showed that the sound quality of synthesized speech improves by controlling formant frequency corresponding to variations of F0. Takano et al. [5] showed that glottal position varies with varying F0 in singing. This implies that the length of vocal tracts may be changed related to F0 changes in singing a song. Therefore, a spectrum control model is required to be able to cope with spectrum related to F0 change.

In this paper, we analyze significant acoustical parameters in singing-voices and construct a singing-voice synthesis method that can transform from read speech to singing-voices, based on the analyzed results.

2. Analysis of acoustic parameters

2.1. Singing-voice and Read speech data
The voice data used in this experiment were obtained from recordings of two adult females reading lyrics and singing the Japanese children’s song “Nanatsunoko.” The songs were recorded on a DAT with 48 kHz sampling and 16-bit accuracy, and then down-sampled to 20 kHz. Figures 1 and 2 show, for example, the sound spectrogram of the phrase “karasunazenakuno” in “Nanatsunoko”, analyzed by using STRAIGHT, for read speeches and singing-voices. Labels were plotted at the top of panels via manual.

2.2. Analysis of F0 in singing-voices
The F0s were estimated by using TEMPO in
STRAIGHT [1], from the recorded data that the singers were asked to sing “Nanatsunoko” with the Japanese vowel /a/ only, to investigate how the F0s vary in time. A melody component that represents note change in the extracted F0 is shown in Figure 3 (a). Figure 3 (b) shows four F0 fluctuations that are found in F0 contours. These fluctuations are defined as follows: (1) overshoot is deflection exceeding the target note after note changes; (2) vibrato is periodic frequency modulation (4 - 7 Hz); (3) fine-fluctuation is irregularly fine fluctuation higher than 10 Hz; and (4) preparation is deflection in the opposite direction of note change observed just before note changes.

In order to investigate how much these F0 fluctuations influence singing-voice perception, we removed each F0 fluctuation from the F0 contours and re-synthesized the singing-voices using the modified F0s. Moreover, we carried out psychoacoustical experiments using these synthesized singing-voices. The result shows that the effects of all F0 fluctuations on singing-voice perception are large, and the effect of overshoot is the largest. The details of this experiment are described in [3]. Therefore, we have to deal with all four F0 fluctuations to control the F0 contours of singing-voices.

2.3. Comparison of phoneme duration
Phoneme durations are measured from a labeled dataset by comparing the duration between consonants in singing-voices and in read speech. The ratios of the duration of each consonant in singing-voices to read speech were as follows: 2.61 (7) for /w/, 1.35 (4) for /m/, 1.37 (4) for /z/, 1.00 (2) for /d/, 2.12 (11) for /r/, 1.50 (26) for /n/, 1.18 (9) for /s/, 1.09 (5) for /t/, 1.14 (3) for /g/, 0.97 (35) for /k/, and 1.28 (6) for /h/. The numbers shown in parentheses indicate the total number of target consonants in all datasets. These results indicate that we can control phoneme duration by controlling articulation manner rather than articulation position. The stretching rate of phoneme durations for fricative, plosive, semivowel, nasal, and /y/ were 1.28, 1.00, 2.37, 1.43, and 1.22, respectively.

2.4. Spectrum variation related to F0 fluctuations
To cope with spectrum information in singing-voices, the relationship between formants in the spectrum of singing-voice and read speech was considered. Each formant frequency was calculated from its amplitude spectrum, represented using STRAIGHT, by the following steps: calculate the auto-correlation from the power spectrum, assume a p-order all-pole filter for the auto-correlation waveform, and then calculate the formant frequency from the root of the coefficients of the filter.

Formant frequencies calculated from vowels in singing-voices tend to increase as the F0 increases.
and these can be approximated as a linear function of F0. Here, let \( f_{\text{aug}} \), \( f_{\text{spk}} \), and \( F \), be the formant frequencies of singing-voices and read speeches, and the ratio of F0 in singing-voices to F0 in read speeches, respectively. It was shown that formant frequencies of singing-voices tend to be higher than those of read speeches as the ratio of F0 in singing-voices to F0 in read speeches increases, as shown in Figure 4. To cope with the spectrum information in singing-voices using \( f_{\text{aug}} \), by using least mean square fitting, this relationship can be represented as

\[
 f_{\text{aug}} = (0.13F + 0.84)f_{\text{spk}}. \tag{1}
\]

3. Schema of singing-voice synthesis

A block diagram of the proposed singing-voice synthesis system is shown in Figure 5. The inputs of the system are (1) a melody component which has a musical interval and duration corresponding to that of a melody of a song, and (2) read speech which reads the lyrics of a song. The read speech is decomposed into F0 and a spectral envelope by STRAIGHT (analysis part) as shown in Figure 3. F0 contours of singing-voices are generated by the F0 control model [3]. This model controls F0 contours of singing-voices by adding F0 fluctuations into the melody component. The spectral envelope is controlled by the following two methods: (1) the duration control method for each phoneme based on melody and (2) the spectrum stretching method in the frequency axis to modify formant frequencies as a function of the F0 ratio of the speaking-voices to read speeches. These controlled acoustic parameters are entered into STRAIGHT (synthesis part).

3.1. F0 control model

The F0 control model [3] generates F0 contours adding four fluctuations: overshoot, vibrato, preparation, and fine-fluctuation into the melody component. The inputs for the model are melody components described as a sum of a step function. Overshoot, vibrato, and preparation are controlled using the transfer function of a second-order system represented as

\[
 H(s) = \frac{K}{s^2 + 2\zeta\Omega s + \Omega^2}, \tag{2}
\]

where \( \Omega \) is natural frequency, \( \zeta \) is damping coefficient, and \( K \) is proportional gain. Here, the impulse response of \( H(s) \) can be obtained as

\[
 h(t) = \begin{cases} 
 \frac{K}{2\sqrt{\zeta^2 - 1}} & |\zeta| > 1 \\
 K & \sqrt{\zeta^2 - 1} \leq |\zeta| < 1 \\
 K\exp(-\Omega t) & |\zeta| = 1 \\
 K \sin(\Omega t) & |\zeta| = 0 
\end{cases} \tag{3}
\]

where \( \lambda_1, \lambda_2 = (-\zeta \pm \sqrt{\zeta^2 - 1})\Omega \), and Eqs. (3a) - (3d) represent solutions to 2nd order exponential damping, 2nd order damping, 2nd order critical oscillation, and 2nd order oscillation (no-loss) models, respectively.

In this paper, control parameters (\( \Omega \) [rad/ms], \( \zeta \), and \( K \)) are used for overshoot (0.0348, 0.5422, 0.0348; 2nd order damping), vibrato (0.0345, --, 0.0018; 2nd order oscillation), and preparation (0.0292, 0.6681, 0.0292; 2nd order damping). These parameter values were obtained by using a nonlinear least-squared-error method to minimize the error between the extracted and the controlled F0s. Controlling fine-fluctuation is done by lowpass filtering and normalizing white noise.
3.2. Spectrum control method

To transform the spectrum of read speech to the spectrum of singing-voice, the connection of consonant with vowel is assumed to be a segmentation of consonant part + coarticulation part + vowel part. The duration of coarticulation part is 40 ms, which is from -10 ms to 30 ms with respect to the boundary of consonant and vowel.

The spectrum control method with regard to duration is done by the following processing: (i) maintaining the same spectrum in coarticulation part; (ii) stretching the spectrum during consonant part (duration from boundary of consonant to vowel) using the stretching rate, described in Sec. 2.2, in which the start duration of 10 ms was maintained; and (iii) stretching the spectrum of speaking voices during vowel part by using the spline interpolation, in which stretched duration was maintained to be a duration subtracting the stretched duration in the consonant and coarticulation parts from the corresponding note duration.

The spectrum control method with regard to formants is done by using Eq. (1) at each 1 ms shift. This is stretching the spectrum information in the frequency domain toward a higher frequency.

3.3. Demonstration

The proposed transformation produces a singing-voice from read speech. A transformation result from Figure 1 is shown in Figure 6: the speech waveform, F0 generated by the F0 control method described in Sec. 3.1, and sound spectrogram generated by two types of stretching described in Sec. 3.2.

Scheffe’s method of paired comparison was used to evaluate the singing-ness of the synthesized singing-voices. The stimuli of this experiment were a real singing-voice, a read speech, and two types of synthesized singing-voice in which: (1) only duration control of the spectral envelope and (2) duration and frequency control of the spectral envelope were used. The F0 of these synthesized singing-voices was generated by the F0 control model. The result showed that the sound quality of the synthesized singing-voice improved by controlling the spectral envelope in not only the time domain (duration stretching) but also the frequency domain (formant stretching), and the sound quality of the synthesized singing-voice is almost the same as that of real singing-voice data.

4. Summary

This paper proposed a singing-voice transform from read speech using STRAIGHT, by considering (1) how to control the F0 contours based on the melody component, (2) how to control phoneme duration based on rhythm, and (3) how to control the spectral envelope related to F0 change. The results show that the proposed system can produce natural synthesized singing-voices, and sound quality of the synthesized singing-voice is almost the same as that of real singing-voice data.

In our future work, we will refine the proposed method in order to automatically control significant parameters based on rules.

5. Acknowledgements

This work was supported by a grant-in-aid for scientific research from the JSPS (No. 13610079) and conducted as a program for the “Fostering Talent in Emergent Research Fields” in Special Coordination Found for Promoting Science and Technology by Ministry of Education, Culture, Sports, Science and Technology.

6. References