Analysis of Acoustic Features Affecting "Singing-ness" and Its Application to Singing-Voice Synthesis from Speaking-Voice

Takeshi Saitou, Naoya Tsuji, Masashi Unoki, and Masato Akagi

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

Abstract

To construct a natural singing-voice synthesis system, it is important to adequately control acoustic features such as fundamental frequency (F0), spectrum shapes, and phoneme duration in the synthesis method. This paper reveals acoustic features affecting singing-voice perception by comparative analyzing singing- and speaking-voices, and then proposes a transforming method from speaking-voice into singing-voice using STRAIGHT [1]. This method is composed of an F0 control model for generating F0 contours of singing-voices, a spectral sequence control model for modifying spectral shapes in speaking-voice, and a duration control model based on rhythm. Results showed that the proposed system could synthesize a natural singing-voice, whose sound quality is almost the same as that of real one.

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, not only acoustic features affecting singing-voice perception has not been investigated deeply, but also methods for controlling acoustic features of singing-voice are not proposed. Therefore, most speech synthesis methods were not proposed for singing-voice synthesis but for speaking-voice synthesis. This paper aims to reveal important non-linguistic information for singing-voice perception, and to show the possibility of synthesizing a singing-voice by adding non-linguistic information to speaking-voice.

In singing a song, one 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 quality of singing-voices. Therefore, to transform speaking-voice into a singing-voice, the following points must be considered:

- how to control F0 based on melody;
- how to control spectrum based on F0 variation; and
- how to control phoneme duration based on rhythm.

It is well known that F0 contours of singing-voices have the following characteristics with regard to F0 fluctuations [2]: (a) the dynamic range of the F0 contours is wider than that of speaking-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.

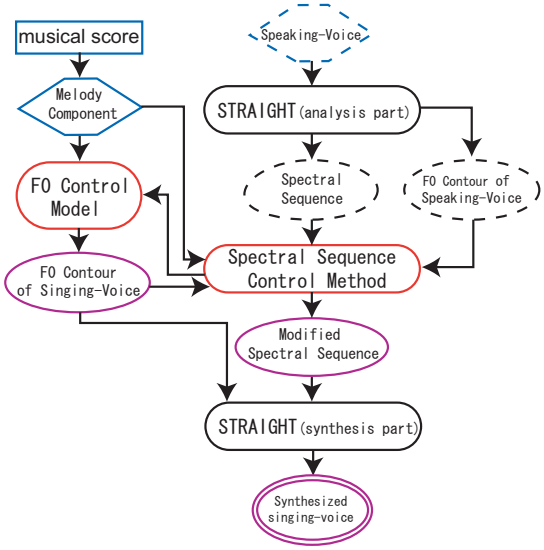

Figure 1: *Singing-voice synthesis system.*

With regard to spectrum information, there are three useful reports. Sundberg [3] showed that there is "Singing-Formant" in singing-voices that is a remarkable peak in the spectrum at around 3 kHz. 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 spectral sequence control model is required to be able to cope with spectrum related to F0 change.

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 speaking-voices. Therefore, to transform speaking-voice into a singing-voice, a duration control model is required to be able to stretch phoneme duration related to note duration.

This paper reveals acoustic features affecting singing-voice perception by comparative analyzing singing- and speaking-voices. Moreover, singing-voice synthesis method that can transform speaking-voice to singing-voice is constructed by developing control model for some acoustic features affecting singing-voice perception.

2. Schema of singing-voice synthesis

A block diagram of the proposed singing-voice synthesis system is shown in Fig 1. This system consist with two procedures; (1) generating F0 contour of singing-voice by controlling melody component produced by a musical score, and (2) modifying spectral sequence of speaking-voice read the lyrics of a song. The melody component has a musical duration corresponding to that of a melody of a song, and the speaking-voice is decomposed into F0 and a spectral envelope by STRAIGHT (analysis part). To fix this system, F0 control model and spectral sequence control model are proposed. The F0 control model can generate F0 contours of singing-voices by adding F0 fluctuations into the melody component. The spectral sequence is controlled by the following two methods: (1) the duration controlling method for each phoneme based on melody and (2) the spectrum modifying method according to F0 variation. These controlled acoustic features are entered into STRAIGHT (synthesis part), and a synthesized singing-voice is produced.

3. F0 control model for singing-voice

3.1. F0 fluctuations in singing-voice

The F0s were estimated using TEMPO in STRAIGHT [1], from the recorded data that the singers were asked to sing Japanese children's song "Nanatsunoko" with a 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, psychoacoustical experiments were carried out using these synthesized singing-voices. The results show that the effects of all F0 fluctuations on singing-voice perception are large, and effect of overshoot is the largest. The details of this experiment were described in [2]. Therefore, it is required to deal with all four F0 fluctuations to control the F0 contours of singing-voices.

3.2. Development of F0 control model for singing-voice

The F0 control model [2] 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

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

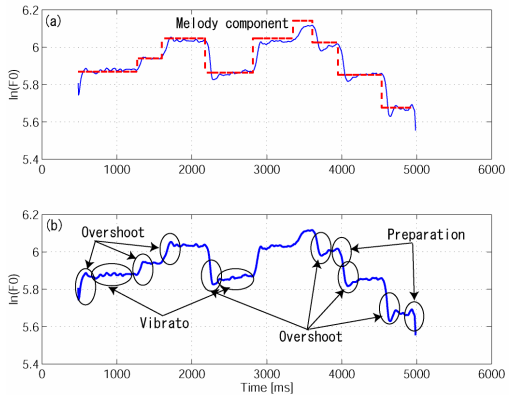

Figure 2: *Extracted F0 using TEMPO. (a) Melody component, (b) F0 fluctuations: overshoot, vibrato, and preparations. Fine-fluctuation in whole contour.*

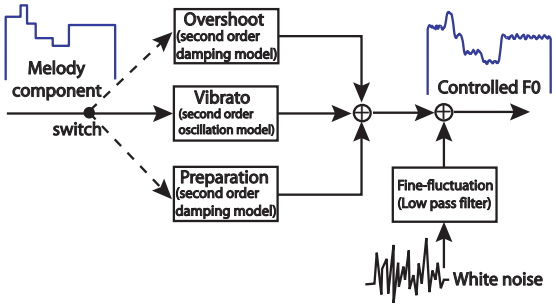

Figure 3: *F0 control model for singing-voice.*

where Ω is natural frequency, ζ is damping coefficient, and K is proportional gain. Overshoot and Preparation are represented with 2nd order damping model, and Vibrato is represented with 2nd order oscillation (no-loss) model.

In this paper, the control parameters (Ω [rad/ms], ζ, 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 using a nonlinear least-squared-error method to minimize the error between the extracted and the controlled F0s. Adding fine-fluctuation is done by lowpass filtering and normalizing of white noise.

4. Spectral sequence control method

In order to consider how to control spectral sequence for transforming speaking-voice into singing-voice, a three-layer model was proposed as shown in Fig 4. This model describes the relationship between "singing-ness" (1st layer) and psychoacoustical primitive feature (2nd layer) and acoustic features (3rd layer). It is able to not only analyze acoustic features affecting "singing-ness" but also reveal psychoacoustical primitive features which constitute "singing-ness" by investigating relation between the layers.

The singing- and speaking-voices data used for experiment are shown Table 1. They are selected by following procedures; (1) selecting 80 data of vowel /a/ from several singing- and spoken-voices data [6], (2) performing listening test to order these data according to "singing-ness", and (3) choosing 11 voices from 80 voices based on the result of listening test.

4.1. Psychoacoustical features affecting "singing-ness"

In this paper, it is considered that "singing-ness" consists of several psychoacoustical primitive features. Thus, the relationship between the first layer and the second layer is investigated using Multi-Dimensional Scaling (MDS). Since stress in 3-D analysis was first less than 10 %, 3-D analysis can be adopted for MDS. Figure 5-(a) shows scatter plots of the voices in 3-D psychoacoustical space. The numbers in the figure correspond to the data number in Table 1. According to the order of "singing-ness" shown in Table 1, it is clear that the 11 voices are divided into 3 groups; group 1 contains "more singing-ness" voices, group 3 contains "less singing-ness" data, and group 2 is located in the space between group 1 and 3. Moreover, the same grouping is shown in 2-D psychoacoustical plane as shown in Fig 5-(b).

To reveal psychoacoustical primitive features affecting "singing-ness", some experiments were done by the following steps; (1) selecting some psychoacoustical primitive features that constitute "singing-ness", (2) calculating the psychological distance in each selected psychoacoustical primitive features by Scheffe's paired comparison (5 grade evaluation), and (3) calculating the direction of each psychoacoustical features using multiple regression analysis in the space of "singing-ness". In step (1), "vibration", "ringing", and "clearness" were selected for the candidates of psychoacoustical primitive feature of "singing-ness". Then, in step (3), multiple correlation coefficients of adjectives in the 3-D were 0.99 for "vibration", 0.99 for "ringing", and 0.84 for "clearness". Since these values are higher and vector of each psychoacoustical feature indicates different direction each other, "singing-ness" is strongly associated with "vibration", "ringing" and "clearness".

4.2. Acoustic features affecting "singing-ness"

To investigate acoustic features affecting "vibration", this paper focused on periodic fluctuation. As mentioned in Sec. 3.1, there is periodic fluctuation that is called Vibrato in F0 contour of singing-voice, and this characteristic affects singing-voice perception. Therefore, fluctuation in amplitude envelope and formants were analyzed using STRAIGHT. From the result, the following characteristics affected "vibration".

- 4 - 6 Hz modulation in the F0 and amplitude envelope
- Formants are fluctuated in frequency and amplitude with the same modulation frequency.
- Intervals and phases of formant deviation are corresponding to those of F0 deviation.

To investigate how much these characteristics affect "vibration", some synthesized voices were generated by adding these features into speaking-voice data and psychoacoustical experiments were carried out using synthesized voice. The result shows that "vibration" increase by adding each characteristic.

Sundberg reported that there was a remarkable peak in the spectrum at around 3 kHz, and this characteristic is peculiar to singing-voice [3]. Thus, spectral envelope and aperiodicity index were analyzed using STRAIGHT. The results showed that there are 2 types of variation in spectral sequence impressing higher "ringing". Type 1 is remarkable peak at around 3 kHz that reported as "singing-formant" [3] as shown in Figure 6-(a). Type 2 is dip of aperiodicity index at around 3 kHz as shown in Fig 6-(b). Remarkable peak indicates a strong harmonic component.

Table 1: *Experimental data for three layer model.*

Data No	Singing / Speaking	Vocalism	Sex	Order of "singing-ness"
1	Singing	Tenor	male	1
2	Singing	Baritone	male	4
3	Singing	Mezzo-soprano	female	11
4	Singing	Minyou *1	male	25
5	Singing	Warabe-uta *2	female	30
6	Singing	Soprano	female	33
7	Singing	Nagauta *3	male	40
8	Singing	Shōmyō *4	male	51
9	Speaking	Tenor	male	66
10	Speaking	Soprano	female	68
11	Speaking	Nagauta *3	male	72

(*1: Japanese folk song, *2: Japanese children's song, *3: Long epic song, *4: Chanting the holy invocation)

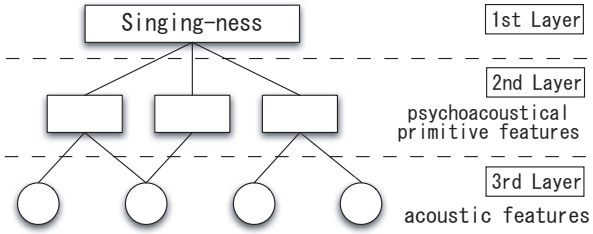

Figure 4: Three-layer model for "singing-ness."

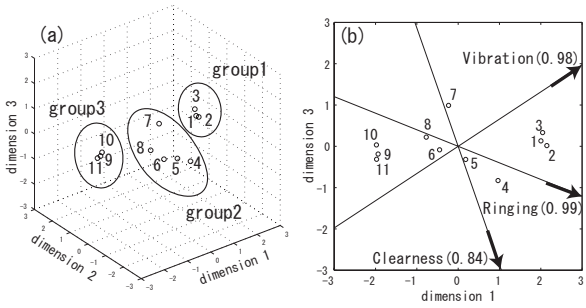

Figure 5: Psychoacoustical space of "singing-ness". (a) 3-dimension, (b) 2-dimension.

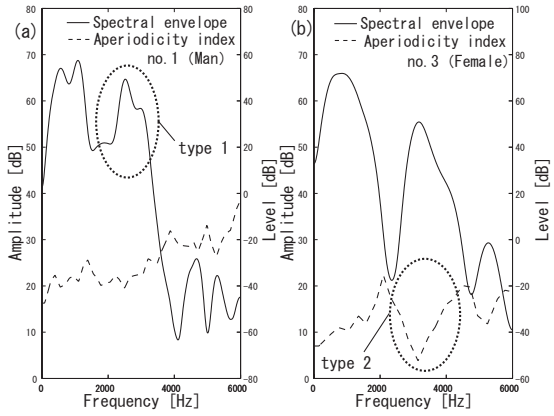

Figure 6: Two types of characteristics for "ringing."

4.3. Spectral sequence control

Based on these results, relationship between the first layer (singing-ness) and the third layer (acoustic features) was investigated. The result showed that "singing-ness" of the voice increases by adding each acoustic feature into speaking-voice. Especially, the effect of acoustic features affecting "vibration" on "singing-ness" is larger than that of acoustic features affecting "ringing". Therefore, in order to control spectral sequence for transforming speaking-voice to singing-voice, the following procedures are required; (1) adding AM and FM corresponding to vibrato in F0, (2) emphasizing peak of spectral envelope or dip of aperiodicity index at around 3 kHz. In this paper, modulation frequency of AM and FM is 5 Hz as the same as that of Vibrato in F0, and peak amplitude is increased 12 dB.

5. Duration control method

To stretch the duration of speaking-voice against the duration 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 duration control is done by the following processing: (1) maintaining the same spectrum in coarticulation part; (2) stretching the spectrum during consonant part (duration from boundary of consonant to vowel) using the stretching rate in which the start duration of 10 ms was maintained; and (3) stretching vowel portion fitting syllable duration into note duration by spline interpolation. In processing (2), the stretching ratio of phoneme durations for fricative, plosive, semivowel, nasal, and /y/ were 1.28, 1.00, 2.37, 1.43, and 1.22, respectively. These values were determined by measuring phoneme duration and comparing the duration between consonants in singing-voices and in read speech.

6. Singing-voice synthesis from read speech

The proposed transformation method produces a singing-voice, Japanese children's song "Nanatsunoko", from speaking-voices. A transformed result from speaking-voice is shown in Figure 6: the speech waveforms, F0 contour, and sound spectrogram of the phrase /karasunazenakuno/ for speaking-voice (Figure 6-(a)) and singing-voice (Figure 6-(b)). F0 contour of synthesized singing-voice is generated the proposed F0 control model, and spectral sequence is modified by the method described in Sec.4 and 5.

In order to evaluate "singing-ness" of synthesized singing-voice, Scheffe's paired comparison was carried out. The result shows that the "singing-ness" of the synthesized singing-voice is almost the same as that of real singing-voice.

7. Conclusions

This paper proposed a singing-voice synthesis method that can transform from speaking-voice to singing-voice using STRAIGHT, by considering (1) how to control the F0 contours based on the melody component, (2) how to control spectral sequence for making voice more "singing-ness", and (3) how to control phoneme duration based on rhythm. The results show that the proposed system can produce natural

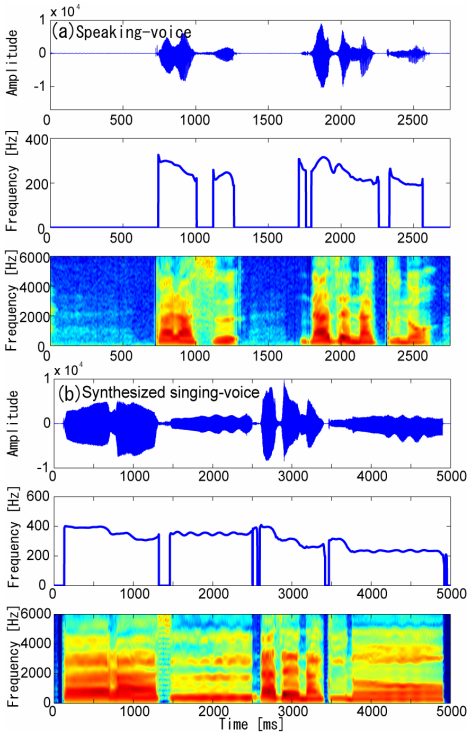

Figure 7: *(top) speech waveform, (middle) F0, and (bottom) sound spectrogram for (a) speaking-voice and (b) synthesized singing-voice.*

synthesized singing-voices, and sound quality of the synthesized singing-voice is almost the same as that of real singing-voice.

8. Acknowledgements

This work was supported by a grant-in-aid for scientific research from the JSPS (No. 13610079).

9. References

[1] 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 Communication*, Vol. 27, pp. 187-207, 1999.

[2] Saitou, T., Unoki, M., and Akagi, M., "Extraction of F0 dynamic characteristics and development of F0 control model in singing voice," *Proc. ICAD2002*, Kyoto, 275-278, 2002.

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

[4] Slawson, A. W. "Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency," *J. Acoust. Soc. Am*., 43, 1, 87-101, 1968.

[5] Takano, S., Honda, K., Masaki, S., Shimada, Y., and Fujimoto, I., "High-resolution imaging vocal gesture using a laryngeal MRI coil and a synchronized imaging method with external triggering," *Proc. Spring Meeting of the Acoustical Society of Japan*, 2-3-8, 2003.

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