Proceedings of 20th International Congress on Acoustics, ICA 2010

23–27 August 2010, Sydney, Australia

Four-tone modeling for natural singing synthesis in Chinese and comparing synthesized singings with speaking voices

Kenko OTA (1) and Terumasa EHARA (2)

(1) Faculty of Systems Engineering, Tokyo University of Science Suwa, Nagano, Japan
(2) Faculty of Human Cultures, Yamanashi Eiwa College, Yamanashi, Japan

PACS: 43.72.Ja, 43.75.Rs

ABSTRACT

Currently, many researchers work on singing synthesis in Japanese or English etc. However, there are few researches on singing synthesis in Chinese. Thus, this paper studies four-tone modeling for natural singing synthesis in Chinese. Four-tone is one of the characteristics of the Chinese syllable, which is modeled as follows: 1st tone is a horizontal linear function, 2nd tone is a linearly increasing function, 3rd tone is a quadratic function and 4th tone is a linearly decreasing function. Four types of four-tone models have been defined in order to clarify an optimal four-tone model. Proposed four-tone models are controlled by a parameter which determines the changing rate of fundamental frequency. As the results of subjective evaluations, the following things have been clarified about the fundamental frequency control for natural singing synthesis: 1st tone is no need to change the fundamental frequency from that of a score, the fundamental frequency of 2nd tone is controlled to change at the last half of the duration of a note and the fundamental frequency of both 3rd and 4th tones are controlled to change at the first half of the duration of a note, and the optimal changing rate for 2nd, 3rd and 4th tones are 1.5%, 1.0% and 1.5% respectively. In this paper, the changing rate of fundamental frequency for singing voices synthesized by the above-mentioned system is compared with that for speaking and singing voices. Firstly, the changing rate of speaking voices in Chinese is calculated. It can be seen that the changing rate for each tone varies widely in individuals. However, the trend of changing rate among 2nd∼4th tones is similar to each speaker. Secondly, the changing rate of real singing voices in Chinese is calculated. It seems that the changing rate of a singing voice is similar to optimal parameter values for singing synthesis except 3rd tone. Moreover, it has been clarified that the changing rate of a singing voice depends on the level of singing. It seems that the changing rate of good singers has a tendency to be smaller than that of poor singers. Thirdly, the similarity between synthesized singings and real singings by Chinese is investigated by comparing the fundamental frequency contour of synthesized singings with that of real singings. It seems that the synthesized singing voice is closed to the real singing voice of good singers.

INTRODUCTION

Currently, many researchers work on singing synthesis techniques and there are fundamental researches on singing voices or applied researches for software products[(1)]. Singing synthesis techniques can be classified into two types. One is corpus-based techniques[(2)][(3)], and the other one is techniques which synthesize a singing voice from a speaking voice[(4)][(5)].

Although corpus-based techniques are highly practicable, there are some defects denoted as follows. It is necessary to record enormous amount of singing voices in order to develop a corpus. Moreover, individuality of synthesized singing voices is lost. These techniques have been applied to singing synthesis of Japanese or English songs. However, there are few techniques for singing synthesis of Chinese songs.

On the other hand, techniques which synthesize a singing voice from a speaking voice can keep the individuality of a speaker. Saito et al. have been proposed one of these techniques. Kubo et al. have been proposed a technique for synthesizing Chinese singing voices. However, the study by Kubo et al. has not considered four-tone which is one of the characteristics of Chinese syllable. Hence, synthesized singing voices could not be heard as natural Chinese singings.

Authors have been proposed a singing synthesis technique in Chinese[(6)][(7)]. In this paper, these results are briefly introduced. Moreover, synthesized singing voices are compared with both speaking and singing voices.

The rest of this paper consists of the following four sections. In the section of “Related researches”, related researches are introduced and the position of our research is clarified. In the section of “Singing synthesis system in Chinese”, the overview of our singing synthesis system and four-tone models are denoted. In the section of “Comparison of synthesized singings with speaking and singing voices”, synthesized singings are compared with speaking and singing voices. Finally, in the section of “Conclusion”, this paper is concluded.

RELATED RESEARCHES

Vocaloid

Vocaloid is one of the corpus-based singing synthesizers. Vocaloid can synthesize arbitrary singing voices by inputting notes and lyrics. The corpus named “singer library” contains samples extracted from enormous singing voices by voice actors. Vocaloid employs a technique which can smoothly concatenate samples, so it can realize natural singing synthesis. Currently, however Vocaloid can treat Japanese and English songs, it cannot treat

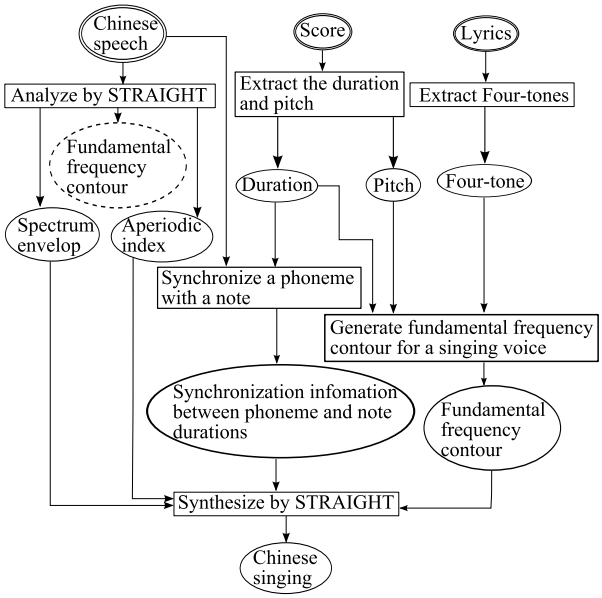

Figure 1: Flowchart of singing synthesis system in Chinese

Chinese songs.

Vocal conversion from speaking voice to singing voice by Saito et al.

In this research, acoustic characteristics affecting the perception of singing voices are clarified by investigating various singing voices. Concretely, these characteristics are overshoot, vibrato, preparation and fine fluctuation. Vocal conversion from a speaking voice to a singing voice can be realized by adding these acoustic characteristics to a speaking voice. Moreover, Saito et al. have also reported that synthesized singings can be perceived as more natural singing voices by adding the singer's formant which is one of the spectral characteristics of singing voices.

Vocal conversion from speaking voice to singing voice in Chinese

As with the synthesizer proposed by Saito et al., this research has also proposed a synthesizer which synthesizes singing voices from speaking voices. This research focuses on synthesizing Chinese singing voices. However, in synthesizing a singing voice, this synthesizer has not considered four-tone in Chinese. Hence, synthesized singing voices could not be heard as natural Chinese singings.

SINGING SYNTHESIS SYSTEM IN CHINESE

In this research, STRAIGHT[(8)], a speech analysis and synthesis software, is used for extracting fundamental frequency contour, spectrum envelope and aperiodic index. Singing synthesis is realized by controlling these speech characteristics to fit the duration and the pitch of a score.

System

Figure 1 shows a flowchart of the developed system. Input data for this system are a Chinese speech, a score and a lyric in Chinese. In this system, spectrum envelop, fundamental frequency contour and aperiodic index are extracted from a speech by STRAIGHT. The fundamental frequency contour of a singing voice is generated from the pitch of a score. Moreover, control of the duration is realized using a piecewise-linear function. Finally, a synthesized singing is generated by STRAIGHT.

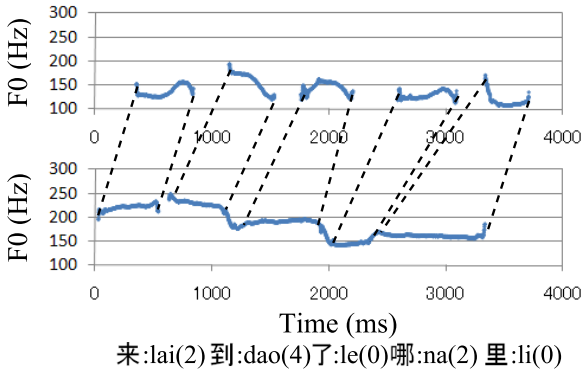

Figure 2: Comparison of fundamental frequency contours.(Top: speaking voice, Bottom: singing voice, The number in brackets represents the tone number of each syllable.)

The research target of this paper is a part for generating a fundamental frequency contour for a singing voice. Normally, the fundamental frequency contour of a singing voice is controlled according to the pitch of a score. However, from Fig. 2, it can be seen that the fundamental frequency contour of a real singing voice by human is not flat. Hence, the naturalness of synthesized singings can be improved by adding the fundamental frequency fluctuation to synthesized singings with a flat fundamental frequency contour. The fundamental frequency fluctuation is caused by four-tone.

Four-tone modeling

Four-tone is one of the characteristics of Chinese language. The pitch fluctuation pattern is classified into four main types in Chinese. The pitch fluctuation pattern of 1st tone is flat high-pitch. That of 2nd tone is a tone starting with mid pitch and rising to a high pitch. That of 3rd tone is a low tone which dips briefly before rising a high pitch. That of 4th tone is a sharply falling tone starting high and falling to a low pitch. Proposed four-tone models are given by the following equations:

model1:

$$F0_{11}(t) = (1+\alpha)F_n \quad (1)$$

$$F0_{12}(t) = \frac{2\alpha F_n}{e_n - s_n}t + F_n\left(1 - \alpha\frac{e_n+s_n}{e_n-s_n}\right) \quad (2)$$

$$F0_{13}(t) = \frac{8\alpha F_n}{(e_n-s_n)^2}\left(t - \frac{s_n+e_n}{2}\right)^2 + (1-\alpha)F_n \quad (3)$$

$$F0_{14}(t) = -\frac{2\alpha F_n}{e_n - s_n}t + F_n\left(1 + \alpha\frac{e_n+s_n}{e_n-s_n}\right) \quad (4)$$

model2:

$$F0_{21}(t) = F_n \quad (5)$$

$$F0_{22}(t) = \frac{\alpha F_n}{e_n - s_n}t + F_n\left(1 - \alpha\frac{s_n}{e_n-s_n}\right) \quad (6)$$

$$F0_{23}(t) = \frac{4\alpha F_n}{(e_n-s_n)^2}\left(t - \frac{s_n+e_n}{2}\right)^2 + (1-\alpha)F_n \quad (7)$$

$$F0_{24}(t) = -\frac{\alpha F_n}{e_n - s_n}t + F_n\left(1 + \alpha\frac{s_n}{e_n-s_n}\right) \quad (8)$$

model3:

$$F0_{31}(t) = F_n \quad (9)$$

$$F0_{32}(t) = \begin{cases} F_n, & s_n \le t < \frac{s_n+e_n}{2} \\ F0_{12}(t), & \frac{s_n+e_n}{2} \le t \le e_n \end{cases} \quad (10)$$

$$F0_{33}(t) = \begin{cases} F_n, & s_n \le t < \frac{s_n+e_n}{2} \\ \frac{25\alpha F_n}{(e_n-s_n)^2}\left(t-\frac{3s_n+7e_n}{10}\right)^2 & \\ +(1-\alpha)F_n, & \frac{s_n+e_n}{2} \le t \le e_n \end{cases} \quad (11)$$

$$F0_{34}(t) = \begin{cases} F_n, & s_n \le t < \frac{s_n+e_n}{2} \\ F0_{14}(t), & \frac{s_n+e_n}{2} \le t \le e_n \end{cases} \quad (12)$$

model4:

$$F0_{41}(t) = F_n \quad (13)$$

$$F0_{42}(t) = \begin{cases} F0_{12}(t), & s_n \le t \le \frac{s_n+e_n}{2} \\ F_n, & \frac{s_n+e_n}{2} < t \le e_n \end{cases} \quad (14)$$

$$F0_{43}(t) = \begin{cases} \frac{25\alpha F_n}{(e_n-s_n)^2}\left(t-\frac{7s_n+3e_n}{10}\right)^2 & \\ +(1-\alpha)F_n, & s_n \le t \le \frac{s_n+e_n}{2} \\ F_n, & \frac{s_n+e_n}{2} < t \le e_n \end{cases} \quad (15)$$

$$F0_{44}(t) = \begin{cases} F0_{14}(t), & s_n \le t \le \frac{s_n+e_n}{2} \\ F_n, & \frac{s_n+e_n}{2} < t \le e_n \end{cases} \quad (16)$$

where, α denotes a changing rate of four-tone, F_n denotes the pitch of a n-th note, s_n denotes the start time of a n-th note and e_n denotes the end time of a n-th note.

Optimal model and parameter

In this paper, the proposed system is optimized as follows: optimal models for 1st∼4th tones are model 2, model 3, model4 and model 4, respectively, and the optimal changing rates α are 0.015 (1.5%) for 2nd tone, 0.010 (1.0%) for 3rd tone and 0.015 (1.5%) for 4th tone, respectively[(6)][(7)].

COMPARISON OF SYNTHESIZED SINGINGS WITH SPEAKING AND SINGING VOICES

From the subjective evaluations, it has been clarified that the naturalness of synthesized singings was improved by adding four-tone to synthesized singings with flat fundamental frequency contour. In this section, the similarity between synthesized singings and real singings by Chinese is investigated by comparing the fundamental frequency contour of synthesized singings with that of real singings. Moreover, this section describes the comparison of the changing rate among synthesized singings, speaking and real singings by Chinese.

Synthesized singings are generated using the above-mentioned system. The lyric input into the system is shown as follows.

春:chun(1) 天:tian(1) 已:yi (3) 来:lai (2) 了:le (0)
春:chun(1) 天:tian(1) 已:yi (3) 来:lai (2) 了:le (0)
来:lai (2) 到:dao (4) 了:le (0) 哪:na (2) 里:li (0)
来:lai (2) 到:dao (4) 了:le (0) 山:shan(1) 里:li (0)
来:lai (2) 到:dao (4) 了:le (0) 村:cun (1) 落:luo(4)
也:ye (3) 来:lai (2) 到:dao(4) 野:ye (3) 地:di (4)

This is a lyric translated into Chinese from a Japanese Children's song named "haruga kita (Spring has come)". The number in brackets represents the tone number where "0" denotes a light tone which is pronounced lightly. Figure 3 shows the score of "haruga kita".

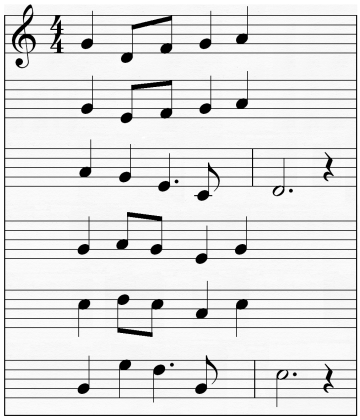

Figure 3: Score of "haruga kita"

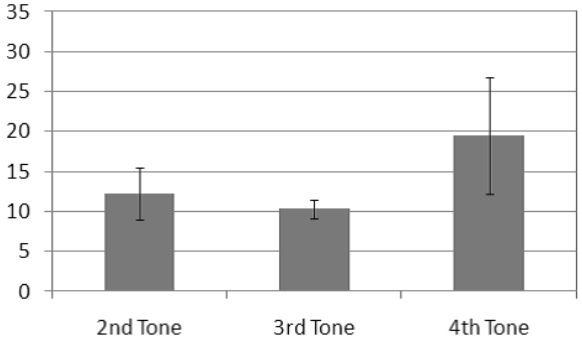

Figure 4: Changing rate of a speaking voice of Speaker #1.(Error bar: Standard deviation of changing rate)

Comparison of speaking voices

Firstly, the changing rate of speaking voices in Chinese is calculated. Here, the changing rate of each syllable is calculated as the ratio of the difference between the maximum and the median of fundamental frequency to the median of fundamental frequency within the each syllable interval. Figures 4 and 5 show the changing rate of speaking voices of Speaker #1 and Speaker #2, respectively who are students from China. From this figure, it can be seen that the changing rate for each tone varies widely in individuals. However, the trend of changing rate among 2nd∼4th tones is similar to each speaker.

Comparison of real singing voices

Secondly, the changing rate of real singing voices in Chinese is calculated. Figure 6 shows the changing rate of a singing voice of Singer #1. From this figure, it has been clarified that the

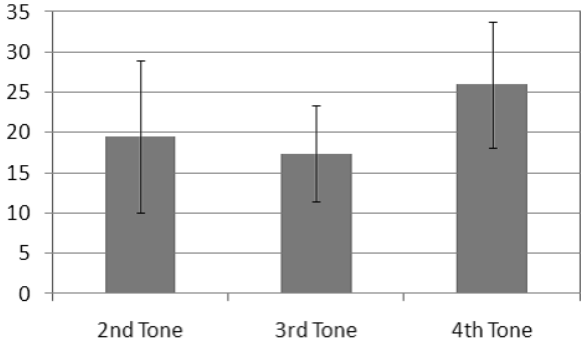

Figure 5: Changing rate of a speaking voice of Speaker #2.(Error bar: Standard deviation of changing rate)

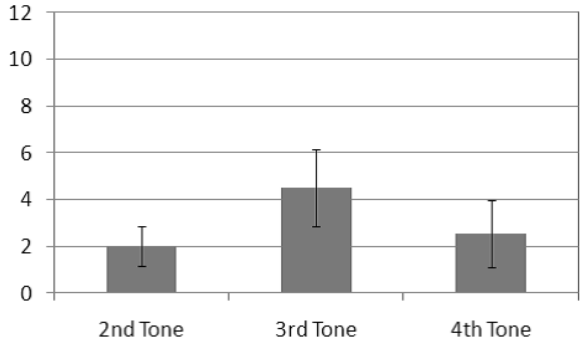

Figure 6: Changing rate of a singing voice of Singer #1.(Error bar: Standard deviation of changing rate)

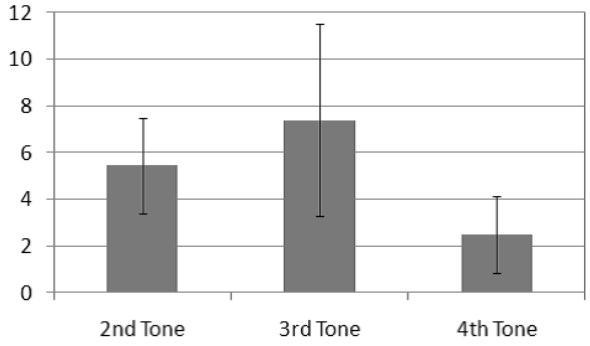

Figure 7: Changing rate of a singing voice of Singer #2.(Error bar: Standard deviation of changing rate)

changing rate of a singing voice is similar to optimal parameter values for singing synthesis except 3rd tone. However, there is a possibility that the changing rate of a singing voice depends on the level of singing. Here, in Fig. 7, the changing rate of a singing voice of Singer #2 is shown. Singer #1 sings better than Singer #2. From Fig. 7, it can be seen that the changing rates for 2nd∼4th tones of Singer #2 are larger than those of Singer #1.

Comparison of synthesized singings with real singings

Thirdly, the similarity between synthesized singings and real singings by Chinese is investigated by comparing the fundamental frequency contour of synthesized singings with that of real singings. Figure 8 shows the fundamental frequency contours of a singing voice of Singer #2, a singing voice of Singer #1 and a synthesized singing. From this figure, it has been clarified that the synthesized singing voice is closed to the real singing voice of a better singer.

CONCLUSION

In this paper, discussed is four-tone modeling on Chinese singing synthesis. As a result of comparison of a singing voice synthesized by the proposed method with speaking voices, it has been clarified that the changing rate of speaking voices varied widely in individuals. However, the trend of changing rate among 2nd∼4th tones is similar to each speaker. Moreover, from the comparison of singing voices recorded by students from China, it has been clarified that the changing rate of a singing voice depends on the level of singing. It seems that the changing rate of good singers has a tendency to be smaller than that of poor singers. Hence, it seems that the synthesized singing voice is closed to the real singing voice of good singers.

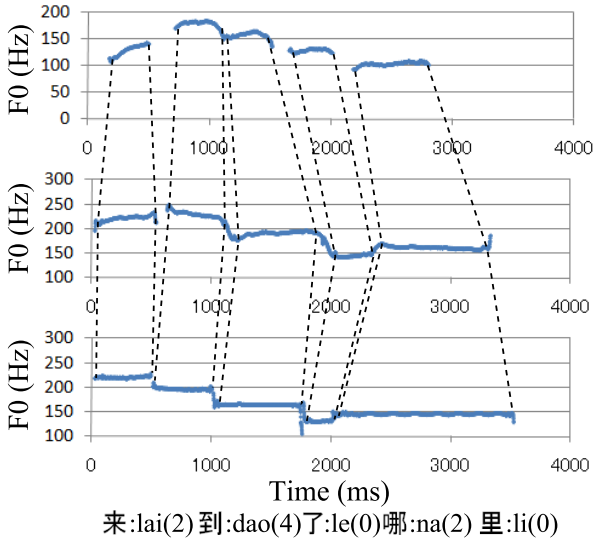

Figure 8: Comparison of fundamental frequency contours.(Top: singing voice of Singer #2, Middle: singing voice of Singer #1, Bottom: synthesized singing)

A future work is to record real singing voices of various singers and to confirm the findings in this paper. The other one is to remove ambiguity of syllable boundaries of singing voices.

ACKNOWLEDGEMENT

We appreciate Prof. Hideki Kawahara, Wakayama University, for accepting us to use STRAIGHT. Moreover, we appreciate students from China, Yamanashi Eiwa College and Tokyo University of Science Suwa, for cooperating with the recording and the evaluation.

REFERENCES

[1] M. Goto, T. Saitou, T. Nakano and M. Fujiwara, “Recent studies on singing information processing”,J. Acoust. Soc., 64,10, pp. 616–623, 2008. (in Japanese)

[2] H. Kenmochi and H. Ohshita, “VOCALOID – Commercial singing synthesizer based on sample concatenation”,Proc. on Interspeech 2007, Antowerp, Aug. 2007.

[3] Shu-Sen Zhou, Qing-Cai Chen, Dan-Dan Wang and Xiao-Hong Yang, “A corpus-based concatenative Mandarin singing voice synthesis system,” Proc. on ICMLC 2008, pp. 2695–2699, Jul. 2008.

[4] T. Saitou, M. Unoki and M. Akagi, “Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis,” Speech Communication, 5, pp. 267–277, 2005.

[5] K. Kubo and T. Ehara, “Conversion from a speaking voice to a singing voice in Chinese”, Proc. on NLP annual meeting, pp. 985–988, 2008. (in Japanese)

[6] K. Ota, H. Matsue and T. Ehara, “A Study on Four-tone Modeling for Automatic Singing in Chinese”, Proc. on ASJ autumn meeting, pp. 407–410, 2009 (in Japanese).

[7] K. Ota, H. Matsue and T. Ehara, “Four-tone Modeling for Natural Singing Synthesis in Chinese”, Proc. on ASJ autumn meeting, pp. 457–458, 2010. (in Japanese)

[8] H. Kawahara, “Restructuring speech representation using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds,” Speech Communication, pp. 187–207, 1999.