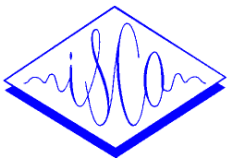

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

A STUDY ON THE PITCH PATTERN OF A SINGING VOICE SYNTHESIS SYSTEM BASED ON THE CEPSTRAL METHOD

Tomio Takara, Kazuto Izumi*, and Keiichi Funaki***

*Department of Information Engineering, **Computing and Networking Center
University of the Ryukyus 1 Senbaru, Nishihara, Okinawa 903-0213 JAPAN
takara@ie.u-ryukyu.ac.jp

ABSTRACT

We synthesize singing voice by rule based on cepstral method. Higher accuracy of analysis and synthesis is required to synthesize singing voice, comparing to rule-based speech synthesis. In this paper, we propose a method of analysis and synthesis with high accuracy. Also, we express pitch patterns minutely by curves that close to natural pitch by using this method. We apply Fujisaki model and add vibrato to the pitch pattern.

1. INTRODUCTION

A singing voice synthesis system by rule is a system that synthesizes a singing voice artificially from a music score and lyrics by controlling the parameters such as phoneme, interval, length and strength of the voice. We have developed rule-based singing voice synthesis system using a method based on cepstrum. It is possible to synthesize a singing voice tentatively by applying the pitch pattern of the singing voice to a common speech synthesis system. However, a higher level of accuracy is necessary in order to analyze and synthesize the singing voice compared to spoken voice. High accuracy is necessary especially in the analysis and generation of delicately changing patterns of pitch.

We have applied methods that were proposed for sound synthesis of musical instruments to the synthesis of the singing voice. Using Lagrange interpolation method, we were able to detect smaller pitch differences within the sampling period. And we have also adopted the Sinc function [1] with the same energy as an impulse in order to express these small pitch variations within the sampling period. Using these methods, we made it possible to analyze and synthesize the pitch pattern with high accuracy.

It was necessary for us to show the effectiveness of these analysis and synthesis methods through a listening test. In this paper, we propose to express the pitch pattern using Fujisaki model function, and the precise vibration of vibrato using the above method. We performed a listening test using the synthesized singing voice to evaluate the effectiveness of the proposed method.

2. CONFIGURATION OF THE SYSTEM

In order to synthesize singing voice, temporal change of the characteristics for singing voice i.e., spectrum, pitch, length and intensity (power) of singing voice need to be controlled. In this paper, we adopt cepstral method as a analysis and synthesis method of speech. Rule-based singing voice synthesis system is constructed by adopting Log Magnitude Approximation (LMA) filter [2] whose coefficients are cepstram parameters estimated by the modified cepstral method.

The parameters for each phoneme are stored in the system in advance to synthesize by rule. The parameters consist of the length of the frame for each phoneme, cepstram coefficients, and the parameter of voiced / unvoiced decision.

In this system, we input text data of the sound name (musical scale), mora, length, and tempo with the text form to generate singing voice. The system generates pitch period, voiced / unvoiced decision, cepstram coefficient with input text. The LMA filter consists of the cepstram parameters as spectral parameters, and is driven by Sinc function series for voiced sounds or by white noise for unvoiced sounds. This system synthesizes singing voice sound by controlling pitch and duration for each phoneme.

3. ACCURATE ANALYSIS AND SYNTHESIS

Needless to say, accurate analysis and synthesis method is required to synthesize high quality singing voice. In this paper, we extract the spectral parameters accurately by using the improved cepstral method [3]. We apply Lagrange interpolation as pitch estimation with sub-sample resolution. Sinc function was adopted to realize delicate pitch changes with sub-sample.

3.1 Analysis Method

The improved cepstral method is adopted to estimate precise spectral envelope in this study.

The improved cepstral method can estimate the log spectrum that ties the peak value of harmonics of FFT spectrum of speech signal. Therefore, more accurate and less biased spectrum can be extracted.

3.2 Pitch Estimation

The effective pitch estimation has been proposed in which fundamental frequency is extracted with higher resolution than sampling period by the Lagrange interpolation of cepstram parameters.

The method achieves better performance than conventional one. Accordingly, we adopt the method for a pitch extraction.

However, the accuracy of pitch extraction is not sufficient only with the Lagrange interpolation. In order to increase pitch extraction accuracy, smoothing of high quefrency between analysis frames is performed as the pre-processing of the Lagrange interpolation. As the post-processing, median smoothing with 5 pitch periods are performed.

3.3 Input Signal of Voiced Sound

In the conventional speech synthesis system, impulse train with discrete valued pitch period was applied as voice sources. When the sampling frequency is 10kHz, the sampling period is 0.1msec and the integer pitch period is discrete-value with 0.1msec interval. However the Lagrange interpolation method is enable to extract the pitch period with sub-sample floating point resolution.

In order to realize the sub-sample pitch period, Sinc function shown as in Eq. (1) is adopted. The pulse train with sub sample pitch period is generated by Eq. (1).

$$x_i(n) = \frac{\sin p\left(x - \frac{N_i - 1}{2} - E\right)}{p\left(x - \frac{N_i - 1}{2} - E\right)} \qquad 1$$

N_i Index of sinc function

E The shift width of peak of sinc function

4. GENERATION OF PITCH PATTERN

When the system synthesizes singing voice, it generates pitch pattern in accordance with a music score. However, in our early system, the pitch pattern is composed of only the straight line, as a result, the naturalness of singing voice was not sufficient. In order to generate pitch pattern more close to that of natural singing voice, we propose the method in which pitch pattern is expressed by curve in Fujisaki model [4] and with vibrato.

4.1 Fujisaki Model

It is well known that the Fujisaki model can generate the natural pitch pattern of speech. This model consists of two parts, i.e., voicing control mechanism as the system that generates the pattern associated with the outflow of expired air, and accent control mechanism to control accent. The command of voicing and accent is the unit step response function. The former is added continuously from immediately start to immediately stop of utterance. According to the accent type of each word, the latter is added synchronizing with the utterance of particular mora. The characteristics of each control mechanism are expressed by step function and represented as Gv (t) and Ga (t).

In this study, we omit stepping response Gv (t) of the voicing set up in this study. It is adequate that step response Ga (t) of the accent is regarded as a monotonous undecrease function, so it can be approximated by Eq. (2).

$$G_a(t) = A_a\{1-(1+ \quad t)e^{- \ t}\} \qquad (2)$$

Aa is the magnitude parameter that indicates the scale of the response, and is a time domain parameter that indicates response speed.

We supposed that voicing component is really associated with pitch pattern since in singing voice, height of the sound is tried consciously to keep it constant. Therefore, only accent function is applied in transient part of 12 frames (1 frame is 10[ms]). The parameters of the accent function are determined by using steepest descent method.

4.2 Estimation of Model Parameter

To generate the accent function of the Fujisaki model, two parameters Aa (size of reaction) and (reaction speed) are required.

Eight songs by two singers are used for the experiment. The model parameters of the accent function was approximated by steepest descent method at pitch pattern of first 10 frames of pitch transition part. We use eight songs that were also used in the listening test. In listening test, an average of the estimated parameters for each song is used. It was shown that the value of the parameters was not so different from each other. Therefore, in singing voice synthesis, we used the parameters averaged among the songs. The model of accent function was matched to the pitch pattern normalized by the difference of two pitches.

4.3 Vibrato

In order to improve the naturalness of synthesized singing voice, application of vibrato to long phonemes were examined. Vibrato can be expressed with two parameters of speed and magnitude since it is a cyclic fluctuation of

fundamental frequency.

Although various aspects of the vibratos have been studied, Seashore found out that the speed of vibrato depended on singer. According to his observation result for 29 singers, the average speed of the fluctuation was 6.6/sec time and the average magnitude was ±48 cents[5].

We adopt the result of Seashore's research and apply it the singing voice synthesis system. If the vibrato begins at 0-th frame, the fundamental frequency added with the vibrato at *i*-th frame, f_1[Hz] is calculated with Eq. (3) as follows.

$$f_1 = f_0 \times 2^{\frac{48}{1200}\sin(0.01i \times 6.6 \times 2\pi)} \quad (3)$$

Where f_0 is original fundamental frequency at 0-th frame.

Note that, vibrato does not added at 12 frames of the transition part, in which but the Fujisaki model is applied.

However, naturalness was lacked from just after transition of pitch since the vibrato is beginning. In order to cope with this, the amplitude of vibrato is gradually enlarged at the beginning of 400ms and is stabilized thereafter.

5. EXPERIMENT

Subjective listening test was conducted to examine the effectiveness of the proposed method.

5.1 Setting Condition of a Listening Test

In order to evaluate the difference in terms of tempo, pitch, transition of pitch, 4 songs shown in table 1 were tested.

The following 7 kinds of songs are compared.

A. Original sound

B. Analysis synthesis sound

C. Synthetic sound by rule that does not adopt interpolation of pitch on transition parts of pitch.

D. Synthetic sound by rule that adopts linear interpolation on transition parts.

E. Synthetic sound by rule that adopts Fujisaki model on transition parts.

F. Synthetic sound by rule that adopts linear interpolation and vibrato at long phonemes.

G. Synthetic sound by rule that adopts Fujisaki model and vibrato at long phonemes.

Table 1 : Synthetic singing voice

Song	Tempo	Average of pitch	Register	Transition of pitch
Doremi no uta	120	3D	4cent	3.3cent
Tsubasa wo kudasai	100	4C	4cent	1.3cent
Wakamono tachi	100	3F	9cent	2.3cent
Kumo yo wake senbaru no sora	120	3F	12cent	4.7cent

The power of the sounds are normalized.

The spectrum parameter of the original sound was used in all conditions in order to compare with only pitch component rigidly. We made synthesized song replacing the spectrum parameters by the original ones through DP matching technique [6].

The singing persons were two adult males, professional singer (tenor) and man of experience of chorus (baritone).

Evaluation is done in comparison of one pair with judging "which is more natural singing voice?". Listens evaluate the sounds with the headphones in a simple sound-proof room by using the comparative audition system with a personal computer.

5.2 Experimental Result

The typical experimental result by the synthetic sound of ( ) is shown in Figure 1. The results of two singers for the song ( ) were almost similar. The graph is the average of 2 experimental results for the songs ( ), and color of the graph shows each listeners' results.

We got similar results of two singers. The result was that a number of choice of (E), synthetic sound by applying Fujisaki model, and (G), synthetic sound by applying Fujisaki model and adding vibrato, were high. In conclusion, Application of the Fujisaki model is effective in the case of songs with large transition of pitch. Combination of the Fujisaki model and vibrato is effective.

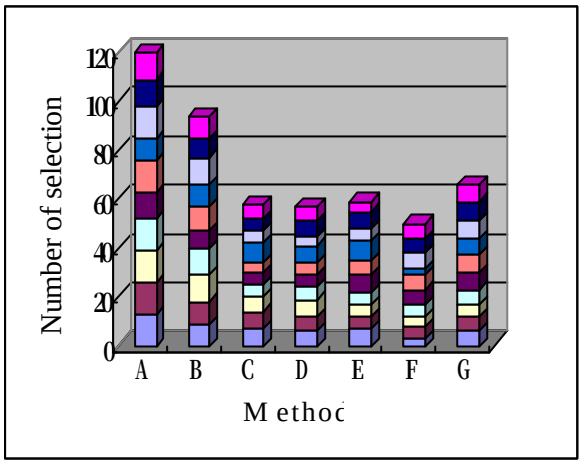

Figure 1 Result of experiment
Song Kumo yo wake senbaru no sora

Among songs ( ) ( ), there was not the difference in numbers of choices, and difference between singers existed. Therefore, it was indicated that it was decided by preference of the listeners. The standard deviation of the select number of synthetic sound C, D, E, F, G is shown in table 2. From the table, we can find out that difference of the result among synthetic sounds are small except in case of song ( ).

6. CONCLUSION

In this paper, we proposed the method of high accuracy analysis and synthesis of singing voice, and difference of the sound quality of the synthetic sound was examined using the different pitch patterns in listening test. Then we showed that the Fujisaki model and vibrato are useful for high quality of singing voice.

There are many factors in a singing voice such as timbre, clearness and so on besides the pitch pattern. In order to improve the quality of the singing voice more, further study on these factors are needed in a future.

Table 2 : Standard deviation among sounds

Song	Register	Transition of pitch	Standard deviation	
( )	small	small	small	3.32
( )				4.69
( )	big			7.75
( )		big	big	11.40

ACKNOWLEDGMENT

This study is supported by Grant-in-Aid for Scientific Research (A) of Japan Society for the Promotion of Science.

REFERENCES

1. Tomio Takara, Tetsuya Kuniyoshi, Takashi Ooishi, and Itaru Nagayama, "Analysis - Synthesis System of Singing Voice with High-Fidelity of Pitch, " ICSP'97, Y3, Seoul, Korea (1997-8).

2. Satoshi Imai, "Log Magnitude Approximation (LMA) Filter," (in Japanese) Trans. IECE Japan J63-A, 12, pp.886-893 (1980-12)

3. Satoshi Imai and Yoshiharu Abe, "Spectral Envelope Extraction by Improved Cepstral Method," (in Japanese) Trans. IECE Japan, J62-A, 4, pp.217-223 (1979-4).

4. Hiroya Fujisaki and Keikichi Hirose, "Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese," J. Acoust. Soc, Jpn. (E) 5, 4, pp.233-242 (1984).

5. Diana Deutsch, "The Psychology of Music," (translated to Japanese) Nishimura Co., Lyd. Printed in Japan, pp.99-100 (1987).

6. Hiroaki Sakoe and Seibi Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition, " IEEE Trans., Speech & Signal Process., vol.ASSP-26, no.1, pp.43-49, (1978-2).