PAPER

KARAOKE SYSTEM AUTOMATICALLY MANIPULATING A SINGING VOICE

Ikuyo Masuda-Katsuse[1)]

Abstract：A karaoke system that manipulates a singing voice has been developed. It has two functions. One is “pitch correction mode,” which converts an out-of-tune vocal sound into one that is in tune. The other is “sub-melody singing generation mode,” which converts solo singing into a duet. Target pitch height for pitch conversion was obtained by adding several F0 fluctuations to a vocal melody written in a MIDI file. Because a speech analysis and synthesis method named TANDEM-STRAIGHT is used for converting the pitch height in this system, it will be easy to introduce a “voice-quality conversion mode” into the system.

歌声を操作するカラオケシステムを開発した。このシステムは２つの機能を有する。一つは「ピッチ制御モード」であり、音高が外れた歌声を正しい音高の歌声に変換する。もう一つは「副旋律歌声生成モード」であり、単独歌唱を二重唱に変換するものである。ピッチ変換の際に目標となる音高は、MIDIファイルに記述されたボーカルメロディにF0摂動が付与されたものである。本システムではTANDEM-STRAIGHTと呼ばれる音声分析合成手法を使って音高を変換しているため、「声質変換モード」の導入も容易である。

Key words：Karaoke, pitch conversion, speech analysis and synthesis, TANDEM-STRAIGHT

キーワード：カラオケ、ピッチ変換、音声分析合成、TANDEM-STRAIGHT

1. Introduction

Karaoke was invented in Japan in the 1970s, and has become popular worldwide. Nowadays, karaoke is used not just for personal amusement but also as a communication tool. In Japan, karaoke parties are often held to promote intra-office relationships.

On the other hand, Independent newspaper dated from the United Kingdom, 8th January 2009 [1], said that karaoke was identified as the worst gadget. British government's survey asked more than 2,500 adults to name the gadgets they regarded as the most important as well as the most irritating. An audience doesn't want to listen to out-of-tune vocal sounds, and not all “tone deaf” singers take pleasure in singing songs in public.

Therefore, we incorporated new functions into a karaoke system [2] [3] [4]. One function is “pitch correction.” It enables out-of-tune vocal sounds to be transformed into singing that is in tune. The other function is “sub-melody singing generation.” It can convert the singing style from solo into duet by synthesizing a sub-melody singing voice using a main melody singing voice.

2. Method

The system's software is written in Java programming language and works on a personal computer. Figure 1 shows the processing flow. First, a MIDI file corresponding to a selected song is read. Next, a vocal melody is extracted from the file and modified to form “correct pitch height.” At the same time as the MIDI is played, pitch conversion of the singing voice begins. The pitch conversion is accomplished using a speech analysis and synthesis method. When “pitch correction mode” is selected, the re-synthesized singing voice is output instead of the input singing voice. When the “sub-melody singing generation mode” is selected, the input and the re-synthesized singing voices are output together.

Each processing is explained in the following subsections.

2.1. Selection of a song and several functions

A song is selected, along with the sex of the singer. Moreover, either “pitch correction mode” or “sub-melody singing generation mode” is selected.

As internal processing, information about a Java soft-

1）Faculty of Humanity-Oriented Science and Engineering, Kindai University　katsuse@fuk.kindai.ac.jp

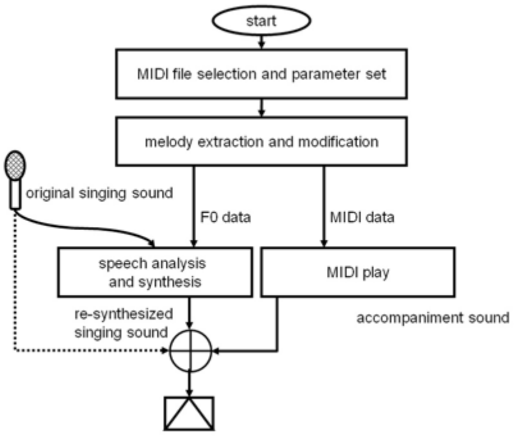
start
MIDI file selection and parameter set
melody extraction and modification
original singing sound
F0 data
MIDI data
speech analysis and synthesis
MIDI play
re-synthesized singing sound
accompaniment sound

Figure 1 Processing flow

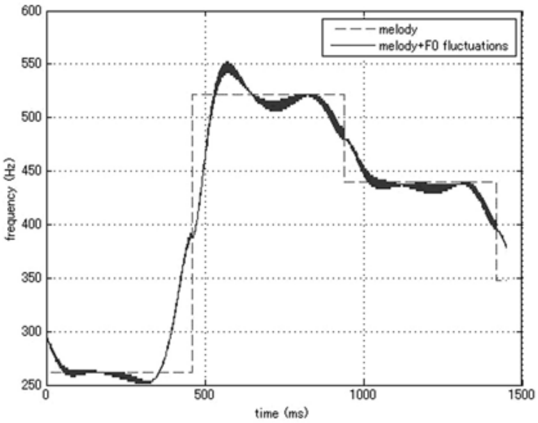

Figure 2 Pitch height on musical notation (broken line) and modified pitch height (solid line)

mixer and available input/output (I/O) device is obtained and the I/O lines are opened.

2.2. Vocal melody extraction

MIDI files under the standard MIDI format (SMF) [5] were prepared for testing. A main vocal melody was written on the first track and a sub-main vocal melody was written on the second track. According to SMF, data on tempo, delta-time, note-on, and note-off are obtained from the MIDI file. Using these data, the frequencies of the vocal melody are calculated as a function of time every 1 ms, from when the MIDI begins to play. In the rest period, the frequency is substituted for 0 Hz. When a male sings a female part of the song, the frequency is set at half pitch, and vice versa.

2.3. Frequency modification

The temporal change of frequency of a main or sub-main melody extracted from a MIDI file is discrete and unnatural as a human singing voice.

Saito et al. [6] showed that there are four kinds of pitch fluctuations that characterize a singing voice: overshoot, preparation, vibrato, and fine fluctuation. Overshoot is deflection exceeding the target note after note changes. Preparation is deflection in the opposite direction of a note change, observed just before the note changes. Vibrato is quasi-periodic frequency modulation. Fine fluctuation refers to irregular fine fluctuations. In this research, these fluctuations are introduced into the frequency of vocal melody extracted from a MIDI file, and the result is regarded as the target pitch height for pitch conversion. Figure 2 shows the frequency change on the musical notation and the results of introducing the fluctuations. These results verified that the target pitch height changes smoothly.

2.4. MIDI play

As soon as the calculation of the target pitch height is finished, the MIDI file is played. A singer starts to sing a song according to the accompaniment.

2.5. Speech analysis and synthesis

As soon as the MIDI file is played, pitch conversion is started. We use a speech analysis and synthesis system for converting the pitch because we intend to implement not only pitch conversion but also voice quality conversion in the future.

2.5.1. *Speech Analysis*

In the speech analysis process, the power, voiced/unvoiced decision, and the spectrum were calculated.

To estimate the spectrum of an input singing voice, the TANDEM-STRAIGHT method [7] is adopted. In the method, first, a temporally stable power spectrum, which is called the TANDEM spectrum, is calculated using two F0-adaptive time windows. Next, the TANDEM spectrum is converted into the TANDEM-STRAIGHT spectrum by convolving a particular smoothing function along the frequency axis.

Although the pitch of the input voice is needed for spectral estimation using the TANDEM-STRAIGHT method, precise pitch estimation in real-time seems to be difficult in such noisy environments as those where karaoke is performed.

Therefore, in this system, the pitch of input speech is not estimated; on the other hand, the frequency calculated from the main vocal melody in the MIDI file is used as a pitch height for TANDEM-STRAIGHT analysis instead of that of the original vocal sound.

When an input singing voice is out-of-tune, the pitch height of the singing differs from the pitch height on the notation. As a result, a time window and a smoothing filter for spectrum estimation are not optimum. This causes deteriorated voice quality; however, such deterioration is practically held low thanks to the robustness of the TANDEM-STRAIGHT method.

As a result of a pilot experiment using an extremely out-of-tune singing voice, the deterioration did not seem to be an obstacle in practical use even if it was detectable.

2.5.2. *Speech Synthesis*

In homomorphic signal processing, if it is assumed that the combined vocal tract and glottal pulse response is in a minimum phase, the vocal tract can be modeled as a linear time-invariant system whose output is the convolution of the impulse response of the vocal tract with the excitation waveform [8]. Singing sound is re-synthesized using a minimum-phase impulse response calculated from complex cepstra and an excitation signal.

The complex cepstra is calculated from the TANDEM-STRAIGHT spectrum.

When the input voice sound is determined to be “unvoiced” in speech analysis processing and the target pitch height is 0 Hz, noise is assigned to the excitation. When the input sound is “voiced,” the glottal pulse is assigned.

The glottal pulse is generated by an all-pass filter design [9], in which the characteristic of the temporal energy distribution of the excitation sound source is controlled by manipulating the group delay characteristic.

2.6. Output of singing voices and the end of process

Only the re-synthesized singing voice is output, when “pitch correction mode” is selected. Both the input singing voice and the re-synthesized singing voice are output when “sub-melody singing generation mode” is selected.

As soon as the MIDI file finishes playing, the speech analysis and synthesis processing ends and the I/O lines are closed.

3. Implementation examples

3.1. Pitch correction mode

To verify the effectiveness of the “pitch correction mode” of the system, an out-of-tune singing voice was input into the system. The song was “Happy Birthday to You.”

Figure 3 shows a part of the song, where change of the pitch height is large. The broken line shows the pitch height on the musical note. The dotted line shows the pitch height of the input singing-sound. The solid line shows the pitch height of the output singing-sound. We can confirm that the pitch height of the singing sound was modified to the target pitch.

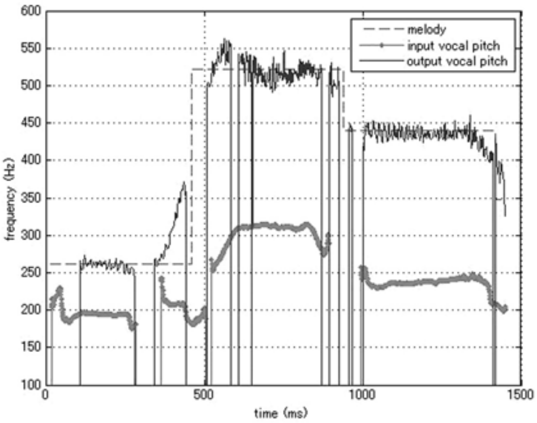

Figure 3 Pitch height of singing voice (dotted line), pitch height of output synthesized using singing voice (solid line), and pitch height on musical notation (broken line).

3.2. Sub-melody singing generation mode

To verify the effectiveness of the “sub-melody singing generation mode” of the system, a singing voice was input into the system. The song was “Summer Memory” composed by Yoshinao Nakada. Figure 4 shows the notation of main and sub melodies at the beginning of the song. The upper part shows a main melody, which a singer sings, written on the first track in a MIDI file. The lower part shows a sub melody, which is the target pitch height for speech synthesis, written on the second track. In short, a melody on the first track in the MIDI file is used for speech analysis and a melody on the second track in the MIDI file is used for speech synthesis.

Figure 5 shows the change of the pitch height at the beginning of the song. The dashed line shows the pitch height of the input singing voice. The dashed line shows the pitch height of the sub melody on the musical note.

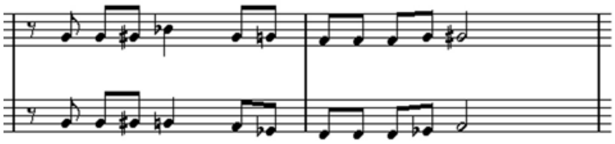

Figure 4 Musical notation of main melody and sub melody

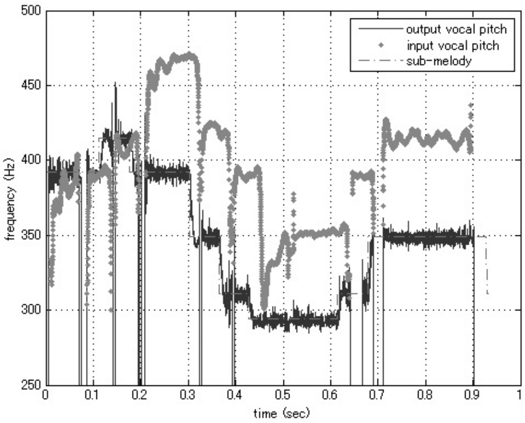

Figure 5 Pitch height of singing voice (dotted line), pitch height of synthesized by using singing voice (solid line), and pitch height of sub melody on the musical notation (chain line).

The solid line shows the pitch height of the output singing voice. We can confirm that the pitch height of the singing voice was modified to the target pitch height concerned with the sub-melody.

4. Future works

We will convert not only pitch height but also voice quality. Frequency axis conversion without deterioration of voice quality using the TANDEM-STRAIGHT has been reported. High-quality speech morphing processing has also been accomplished [10] [11]. In this system, in order to imitate the vocal quality of the original singer of the song, we will embed the singing voice of the original singer or its parameters into the MIDI file and covert the voice quality of the input singing voice to one resembling the original singing voice by introducing a part of the morphing function of the TANDEM-STRAIGHT system.

This system has a weak spot. If a singer sings *out-of-tempo*, timing of the turn of the syllable results in shifting timing of the turn of the pitch height. Saito et al. [12] developed 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 notation. In their system, the duration of each phoneme in the speaking voice is determined by considering the duration of its musical note. In the case of our system, timing of the turn of the pitch height must be decided by synchronizing the timing of the turn of the phonemes of the input singing voice. If several distinctive features instead of lyrics are written along with the musical notation in the MIDI file, synchronization between the turn of pitch height and the turn of the phoneme might be carried out.

Finally, in the present system, although a sub-melody is written on the second track, it is not difficult to introduce an automatic sub-melody generation function [13] [14].

5. References

[1] http://www.independent.co.uk/news/uk/home-news/stand-up-if-you-hate-karaoke-1232067.html

[2] I. Masuda-Katsuse and Y. Urakawa, "Development of KARAOKE system that automatically modifies out-of-tune vocals," Proc. Youngnam and Kyushu Joint Conference on Acoustics 2009, 151-154, 2009.

[3] I. Masuda-Katsuse, "Development of KARAOKE system that automatically modifies out-of-tune vocals," Proc. of annual meeting of the Acoustical Society of Korea, 1-4, 2009.

[4] I. Masuda-Katsuse, "Development of Karaoke system with pitch conversion of vocal sound," Proc. of autumn meeting of the Acoustical Society of Japan, 2-7-6, 2010.

[5] M. Kato, "The background of birth of MIDI standard, and the outline of MIDI standard: Change of the environment of electronic music," J. Acoustical Society of Japan, 64 (3), 158-163, 2008. (*in Japanese)*

[6] T. Saitou, M. Unoki, and M. Akagi, "Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis," Speech Communication 46, 405-417, 2005.

[7] H. Kawahara, M. Morise, T. Takahashi, H. Banno, R. Nisimura, and T. Irino, "Improving accuracy in spectral envelope estimation based on TANDEM-STRAIGHT - Recovery of higher frequency components exceeding Nyquist limit posed by the fundamental frequency," IEICE Technical Report, SP2008-23, 2008. (*in Japanese*)

[8] A. V. Oppenheim and R. W. Schafer, "Homomorphic Signal Processing in Digital Signal Processing," Prentice-Hall, New Jersey, 1975.

[9] H. 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, 27, 187-207, 1999.

[10] H. Kawahara and M. Morise, "TANDEM-STRAIGHT and Voice Morphing: Applications to Emotional Speech and Singing Research," J. of the Phonetic Society of Japan, 13 (1), 29-39, 2009. (*in Japanese*)

[11] H. Kawahara, R. Nisimura, T. Irino, M. Morise, T. Takahashi, and H. Banno, “Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown,” Proc. ICAASP 2009, 3905-3908, 2009.

[12] T. Saitou, M. Goto, M. Unoki, and M. Akagi, “Speech-To-Singing Synthesis: Converting Speaking Voices to Singing Voices by Controlling Acoustic Features Unique to Singing Voices,” Proc. 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA2007), 215-218, Oct. 2007.

[13] N. Yogev and A. Lerch, “ A System for Automatic Audio Harmonization,” 25. TONMEISTERTAGUNG – VDT INTERNATIONAL CONVENTION, 2008.

[14] R. Tamura, Y, Tajima, and Y. Kotani, “Automatic Sub-Melody Generation Using Hidden Markov Models of Pitch and Duration,” The Special Interest Group Notes of IPSJ, 2007-MUS-69, 2007. (*in Japanese*)