TRANSFERRING VOCAL EXPRESSIONS OF A PROFESSIONAL SINGER TO UNACCOMPANIED SINGING SIGNALS

Yukara Ikemiya, Kazuyoshi Yoshii, Katsutoshi Itoyama
Graduate School of Informatics, Kyoto University, Japan
{ikemiya,yoshii,itoyama}@kuis.kyoto-u.ac.jp

ABSTRACT

This paper presents a singing-style transfer system that enables users to attach their favorite vocal expressions (fluctuation patterns of vocal F0 contours) extracted from polyphonic music signals to existing unaccompanied singing signals. Since our system can separately manipulate three major types of vocal expressions (vibrato, glissando, and kobushi), users can easily understand the effect of singing-style transfer. This is the important first step towards our final goal that aims to replace vocal expressions included in a musical piece with those of an arbitrary singer. Such music touch-up forms the basis of *active music listening* that gives a new dimension to how users enjoy musical pieces found by music information retrieval (MIR) techniques.

1. INTRODUCTION

Active music listening [1] has recently been considered to be one of the most attractive directions of music information retrieval (MIR) research. While listening to music, we often wish that a song be sung by another singer or a particular part be performed in a different way. Recent advances of music signal processing enable us to *actively* customize the content of existing CD recordings, *e.g.* edit drum parts as in MIDI sequencers [2], control the volume balance between harmonic instruments [3, 4] or between vocal parts and accompanying parts [5].

So far, many methods have been proposed for manipulating the pitches, timbres, and volumes of singing voices (main melodies of popular music). A speech analysis-and-synthesis system called STRAIGHT [6], for example, can decompose speech or singing voices into fundamental frequencies (F0s), spectral envelopes, and non-periodic components. High-quality pitched-changed or timbre-changed singing voices can be resynthesized by manipulating F0s and spectral envelopes. Ohishi *et al.* [7] proposed a method that represents the temporal dynamics of vocal F0 contours by using a probabilistic model, which can be used for generating expressive F0 contours from music scores. To overcome the conventional limitation that input data are unaccompanied singing signals, Fujihara and Goto [8] tried to directly modify vocal timbres by modeling spectral envelopes in polyphonic music signals.

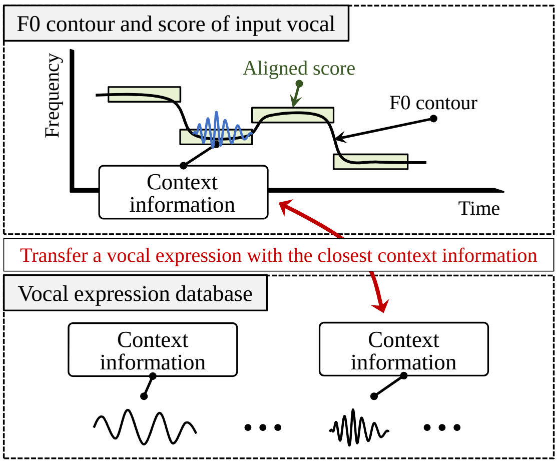
F0 contour and score of input vocal
Frequency
Aligned score
F0 contour
Context information
Time
Transfer a vocal expression with the closest context information
Vocal expression database
Context information
Context information

Figure 1. Overview of vocal expression transfer.

In this paper we propose a singing-style transfer system that explicitly represents the characteristic patterns of vocal F0 contours (*e.g.*, vibrato, glissando, and kobushi) for separately editing each type of vocal expressions in an intuitive way. The vocal expressions can be extracted from audio signals of existing musical pieces [9]. A main contribution of this study is to enable users to transfer collected vocal expressions to existing singing signals, not to synthesize completely new singing signals by using singing synthesizers such as Vocaloid and Sinsy.

2. PROPOSED SYSTEM

This section describes our singing-style transfer system based on automatic extraction of vocal expressions.

2.1 Extraction of Vocal Expressions

To make a database of vocal expressions from polyphonic music, we parameterize vocal expressions such as vibrato, glissando, and kobushi included in vocal F0 contours [9]. Note that vibrato is a periodically-fluctuated F0 contour, kobushi is a short tremolo that often appears in Japanese folk songs (called *enka*), and glissando is a gliding-down F0 contour at a note offset (called glissdown) or a gliding-up contour at a note onset (called glissup). The F0 contours can be estimated from polyphonic music signals by using a sub-harmonic summation method with appropriate constraints on temporal continuity of F0s. The three types of

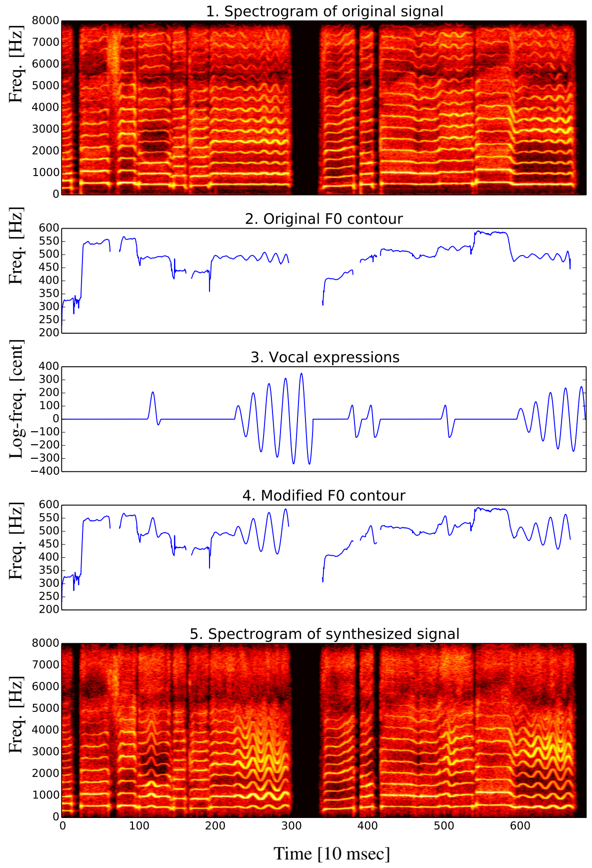

Figure 2. An example of vocal expression transfer.

vocal expressions (short F0 contours) are then extracted by using a pattern matching method [9]. To make a distinction between vibrato and kobushi having similar F0 contours, vibrato, glissando and kobushi are detected in this order. More than one vocal expression is not allowed to be detected in the same region.

We compactly represent the F0 shapes of extracted vocal expressions as parametric functions having several parameters that enable us to resynthesize F0 contours. More specifically, the vibrato is represented as a sine wave with periodicity and amplitude (deepness) parameters, the glissando as a half quadratic function, and the kobushi as a 6th polynomial function (see detailes in [9]). Each vocal expression is indexed by the information of the original context (a note number of the target note, note numbers of the previous and next notes, note length, and a relative position in a phrase).

2.2 Vocal Expression Transfer

Our system takes as input an unaccompanied singing signal with a temporally-aligned music score (Fig. 1). First, context information is extracted for each musical note. A vocal expression that has the closest context information is selected from the database and then added to the horizontal F0 contour obtained by removing original vocal expressions (simply averaging F0s in the musical note). We use STRAIGHT [6] for resynthesizing a singing signal from the modified F0 contour with preserving the timbral characteristics of the original singing signal.

3. DEMONSTRATION

We made a vocal-expression database of a well-known professional Japanese folk singer named *Misora Hibari*, who is characterized by deep vibrato and kobushi, and transferred her vocal expressions to an unaccompanied singing signal of another singer in the chorus part of No.7 from the RWC Music Database: Popular Music [10]. Fig. 2 shows from top to bottom the original spectrogram, the original F0 contour, vocal expressions to be transferred, a modified F0 contour, and a spectrogram synthesized by STRAIGHT. Sample audio files are available on our website.[1]

4. CONCLUSION

In this paper we presented a vocal-expression transfer system for active music listening. A database of vocal expressions is obtained by analyzing F0 contours for commercial CD recordings. The vocal expressions are then transferred to unaccompanied singing signals sung by another singer.

We plan to deal with more types of vocal expressions that characterize singing styles and subjectively evaluate the usefulness of our system. Our system could be used for investigating how singing styles (F0 fluctuations) affect music perception. Towards the ultimate goal that aims to edit singing styles included in polyphonic music signals, we are developing a method that can directly modify only vocal components in the polyphonic mixture.

Acknowledgment: This study was supported in part by JSPS KAKENHI 26700020 and JST CREST OngaCREST project.

5. REFERENCES

[1] M. Goto: "Active Music Listening Interfaces Based on Signal Processing," *Proc. ICASSP*, 2007.

[2] K. Yoshii et al.: "Drumix: An Audio Player with Real-time Drum-part Rearrangement Functions for Active Music Listening," *IPSJ Journal*, Vol. 48(3), 2007.

[3] J. Fritsch et al.: "Score Informed Audio Source Separation using Constrained Nonnegative Matrix Factorization and Score Synthesis," *Proc. ICASSP*, 2013.

[4] N. J. Bryan et al.: "Source Separation of Polyphonic Music with Interactive User-Feedback on a Piano Roll Display," *Proc. ISMIR*, 2013

[5] Z. Rafii et al.: "Combining Modeling of Singing Voice and Background Music for Automatic Separation of Musical Mixtures," *Proc. ISMIR*, 2013

[6] H. Kawahara et al.: "Tandem-STRAIGHT: A Temporally Stable Power Spectral Representation for Periodic Signals and Applications to Interference-free Spectrum, F0, and Aperiodicity Estimation," *Proc. ICASSP*, 2008.

[7] Y. Ohishi et al.: "Mixture of Gaussian Process Experts for Predicting Sung Melodic Contour with Expressive Dynamic Fluctuations," *Proc. ICASSP*, 2014.

[8] H. Fujihara et al.: "Concurrent Estimation of Singing Voice F0 and Phonemes by Using Spectral Envelopes Estimated from Polyphonic Music," *Proc. ICASSP*, 2011.

[9] Y. Ikemiya et al.: "Transcribing Vocal Expression from Polyphonic Music," *Proc. ICASSP*, 2014.

[10] M. Goto et al.: "RWC Music Database: Popular, Classical, and Jazz Music Databases," *Proc. ISMIR*, 2002.

[1] winnie.kuis.kyoto-u.ac.jp/members/ikemiya/demo/ismir2014.html