Autocomplete Vocal-f_0 Annotation of Songs Using Musical Repetitions

Tomoyasu Nakano
t.nakano@aist.go.jp
National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan

Yuki Koyama
koyama.y@aist.go.jp
National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan

Masahiro Hamasaki
masahiro.hamasaki@aist.go.jp
National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan

Masataka Goto
m.goto@aist.go.jp
National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan

ABSTRACT

Audio annotation for music clips is an important task for machine-learning-based music analysis and applications. However, it is a time-consuming task because it often requires repetitive manipulations even though typical audio files often contain repetitive structures (*e.g.*, a song often has similar phrases used multiple times). In this paper we present a new interaction technique, to intelligently automate repetitive manipulations for audio annotation. It mimics the "autocompletion" functions used in source code editors and spreadsheet software and is called *Autocomplete Audio Annotation*. We developed a proof-of-concept system for annotating the continuous *fundamental frequency* (f_0) of a vocal part of a song.

CCS CONCEPTS

• **Applied computing** → **Sound and music computing**.

KEYWORDS

Autocompletion; audio annotation; fundamental frequency (f_0); song structure; musical repetitions.

ACM Reference Format:
Tomoyasu Nakano, Yuki Koyama, Masahiro Hamasaki, and Masataka Goto. 2019. Autocomplete Vocal-f_0 Annotation of Songs Using Musical Repetitions. In *24th International Conference on Intelligent User Interfaces (IUI '19 Companion), March 17–20, 2019, Marina del Rey, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308557.3308700

1 INTRODUCTION

Annotation of musical pieces extends music information retrieval, and the annotated pieces are important as training data in machine learning. Fully automating the annotation of existing songs is difficult and the annotation task requires human labor, but human annotation can be supported by a semi-automatic and interactive method. To improve the speed and accuracy of interactive annotation, we propose *Autocomplete Audio Annotation* of songs using structured repetitions.

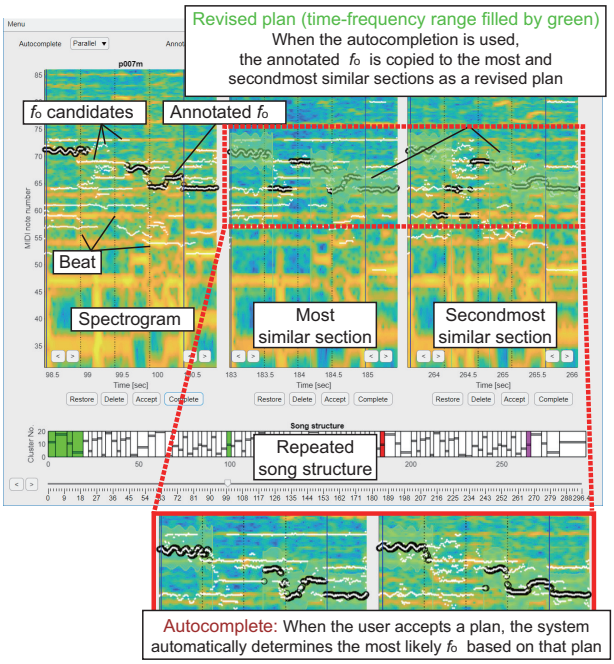

Figure 1: Screenshot of Autocomplete Audio Annotator, which provides autocompleted annotations (Bottom). Once the user provides an annotation (Top left), these autocompletions are generated by using an estimated repetitive musical structure shown at the "Song structure" part.

Music usually has sections like choruses and similar phrases that are repeated. This characteristic repetition has been used for music analysis, such as sound source separation [12], analyzing music structure [3, 8], and lyric transcription [9]. Interaction that speeds up the next input based on past input by the user is called *autocompletion* and is utilized in various situations, such as speech input [5] and 3D sculpting [11]. Autocompletion and the repetitions in music, however, have not been applied to annotation.

The annotation we target is the *fundamental frequency* (f_0) of the vocal in the popular music. The existence of instrument sounds,

IUI '19 Companion, March 17–20, 2019, Marina del Rey, CA, USA

ACM ISBN 978-1-4503-6673-1/19/03…$15.00
https://doi.org/10.1145/3308557.3308700

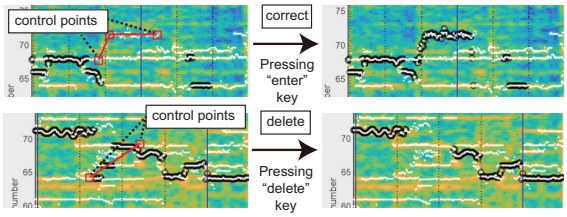

Figure 2: Manual edit of annotation by mouse interaction.

however, makes its automatic estimation with high accuracy difficult. Manual annotations are also difficult and time-consuming. Although Songle [6] and Ensemble [13] provide a semi-automated and crowdsourced architecture used to generate vocal f_0 annotations from the audio signal of a music piece, there has been no work using the repeated structure of music and the autocompletion approach.

2 IMPLEMENTATION

As shown in Figure 1, our proposed system first automatically estimates the vocal f_0 values (black circles), which may contain errors and thus need manual correction, and visualizes them on the spectrogram with vocal f_0 candidates (white dots). They are estimated by using PreFEst [2]. We then estimate a repetitive musical structure by using a method based on the spectral clustering [8]. To estimate the repetitive structure, we perform beat tracking based on the recurrent neural network [1] through the *madmom* library.

On the interface screen the section currently being annotated and the two sections that are the most and secondmost similar are visualized in parallel. The f_0 similarity between repetition sections is calculated by using a dynamic time warp method based on the Euclidean distance of an f_0 candidate.

A mouse is used for manual editing of annotation (Figure 2). If a user selects multiple points by left-clicking them, they are connected by interpolation. The points can be deleted by right-clicking them. f_0 annotation is corrected by pressing the *enter key* and a section is annotated as having no vocal by pressing the *delete key*. To confirm the annotated f_0, the vocal can be resynthesized by using the sinusoidal model and played back.

When the user edits f_0, the system proposes a revised plan, f_0 time-frequency range (green polygons), for other sections as the autocompletion. When the user accepts that plan, the system automatically determines the most likely and smooth f_0 based on the plan by using the Viterbi algorithm to adapt different f_0 values even though two sections are musically similar.

3 USER FEEDBACK

In order to verify the concept of using autocompletion for annotation, an initial pilot study was conducted under the condition that a user selects from the given f_0 candidates. A male musician who had experience with melody transcription and correcting melody (f_0) using Songle [6] and *Melodyne* was asked to annotate vocal f_0 of six songs from RWC-MDB-P-2001 [4]. The six songs were four Japanese songs sung by two males and two females and two English songs sung by one male and one female. The user performed two annotation tasks for five minutes on each song: (T1) annotating f_0 of Nos. 1, 80, and 98 without using the autocompletion, and (T2) annotating f_0 of Nos. 2, 78, and 96 with the autocompletion.

The user was presented the six songs in the order of Nos. 1, 80, 2, 78, 98, and 96. For each song, the progression rate of annotation completed in 5 minutes of task was 9.3%, 25.0%, 19.1% for Nos. 1, 80, and 98 (T1) and 19.1%, 9.0%, 45.0% for Nos. 2, 78, and 96 (T2), respectively. In T2 the autocompletion was used at 0%, 48.8%, and 57.0% for Nos. 2, 78, and 96, respectively.

The user commented on the effectiveness of the autocomplete function and the similar part display function. He also suggested the following improvements for the system: rejection when the similar part is erroneous, increase of the number of similar sections, and visualization of f_0 likeliness.

4 CONCLUSION

In this paper we introduced an autocomplete concept that speeds vocal f_0 annotation for music. The concept can be applied to various annotations other than those of music and can also be applied for purposes other than annotation, such as pitch editing (re-synthesis) of singing voice.

For annotating music data other than f_0, some interaction techniques have been investigated so far; for example, interactions for tagging sound events [7] and for correcting speech recognition results [10] have been proposed. Our system could be further extended by combining these techniques.

ACKNOWLEDGMENTS

This work was supported in part by JSPS KAKENHI Grant Number JP17K12721 and JST ACCEL Grant Number JPMJAC1602.

REFERENCES

[1] Sebastian Böck, Florian Krebs, and Gerhard Widmer. 2016. Joint Beat and Downbeat Tracking with Recurrent Neural Networks. In *Proc. ISMIR 2016.* 603–608.
[2] Masatala Goto. 2004. A Real-time Music Scene Description System: Predominant-F0 Estimation for Detecting Melody and Bass Lines in Real-world Audio Signals. *Speech Communication* 43, 4 (2004), 311–329.
[3] Masataka Goto. 2006. A Chorus-Section Detection Method for Musical Audio Signals and Its Application to a Music Listening Station. *IEEE Trans. on Audio, Speech and Language Processing* 14, 5 (2006), 1783–1794.
[4] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. 2002. RWC Music Database: Popular, Classical, and Jazz Music Databases. In *Proc. ISMIR 2002.* 287–288.
[5] Masataka Goto, Katunobu Itou, and Satoru Hayamizu. 2002. Speech Completion: On-demand Completion Assistance Using Filled Pauses for Speech Input Interfaces. In *Proc. ICSLP-2002.* 1489–1492.
[6] Masataka Goto, Kazuyoshi Yoshii, Hiromasa Fujihara, Matthias Mauch, and Tomoyasu Nakano. 2011. Songle: A Web Service for Active Music Listening Improved by User Contributions. In *Proc. ISMIR 2011.* 311–316.
[7] Bongjun Kim and Bryan Pardo. 2017. I-SED: an Interactive Sound Event Detector. In *Proc. IUI 2017.* 553–557.
[8] Brian McFee and Daniel P.W. Ellis. 2014. Analyzing Song Structure with Spectral Clustering. In *Proc. ISMIR 2014.* 405–410.
[9] Matt McVicar, Daniel PW Ellis, and Masataka Goto. 2014. Leveraging Repetition for Improved Automatic Lyric Transcription in Popular Music. In *Proc. ICASSP 2014.* 3117–3121.
[10] Jun Ogata and Masataka Goto. 2012. PodCastle: Collaborative Training of Language Models on the Basis of Wisdom of Crowds. In *Proc. Interspeech 2012.* 2370–2373.
[11] Mengqi Peng, Jun Xing, and Li-Yi Wei. 2018. Autocomplete 3D Sculpting. *ACM Trans. Graph.* 37, 4 (2018), 132:1–132:15.
[12] Zafar Rafii and Bryan Pardo. 2013. REpeating Pattern Extraction Technique (REPET): A Simple Method for Music/Voice Separation. *IEEE Trans. on Audio, Speech and Language Processing* 21, 1 (2013), 73–84.
[13] Tim Tse, Justin Salamon, Alex Williams, Helga Jiang, and Edith Law. 2016. Ensemble: A Hybrid Human-Machine System for Generating Melody Scores From Audio. In *Proc. ISMIR 2016.* 143–149.