AUDIO MELODY EXTRACTION FOR MIREX 2008

Matti Ryynänen and Anssi Klapuri
Department of Signal Processing, Tampere University of Technology, Finland
matti.ryynanen@tut.fi, anssi.klapuri@tut.fi

ABSTRACT

We submit a method for the MIREX 2008 task "Audio Melody Extraction". The method first transcribes input audio file into MIDI notes and then estimates fundamental frequency (F0) trajectory for each note. The melody transcription method uses a frame-wise pitch salience estimator to measure the strength of different fundamental frequencies. HMMs are used to model melody notes, other instrument notes, and silence or noise segments. Key estimation and between-note transition are applied as a musicological model. After transcribing the melody notes, a detailed F0 trajectory is estimated for each note by using the most salient F0 estimates and interpolation with cubic splines.

1 INTRODUCTION

Melody transcription refers to the automatic extraction of a parametric representation (e.g., a MIDI file) of the melody notes within a polyphonic music piece. Melody of a piece is an organized sequence of consecutive notes and rests, usually performed by a lead singer or by a solo instrument. More informally, the melody is the part one often hums along when listening to a music piece.

Several melody transcription methods have been proposed which either produce F0 trajectory for the melody (as in this MIREX task) or produce discrete note events as an output (see [3] for a comparison of different approaches). Automatic melody transcription facilitates applications in music-information retrieval and content-based audio modification, for example.

Our submission uses a melody transcription method [5] which produces discrete note events as an output. This is extended with a method for estimating detailed F0 trajectory for each note [4]. These are briefly introduced in the following. For more details, please see [5, 4].

2 METHOD DESCRIPTION

The melody transcription method [5] uses a frame-wise pitch salience estimator to measure the strength of different fundamental frequencies in 92.9 ms analysis frames with 23.2 ms interval between successive frames (i.e., 4096 and 1024 samples at 44.1 kHz sampling frequency f_s). In the following, $s_t(f)$ denotes the salience of fundamental frequency f in frame t. Fundamental frequency f ranges from MIDI note 44 to 84 with approximately 800 values distributed between these limits. The estimated saliences, their time differentials, and a measure of incoming spectral energy are used as features for computing observation likelihoods for melody notes, other instrument notes, and silence or noise segments, each of which is modeled using a hidden Markov model. The parameters of the models have been estimated from several hours of music using RWC Pop and Genre databases [1, 2]. A musicological model uses the saliences to estimate the musical key of the piece and employs the corresponding between-note transition probabilities. These probabilities are modeled with a note bigram trained with thousands of MIDI melodies. The Viterbi algorithm is used to find the optimal path through the melody note models in order to produce a transcribed melody note sequence, where the i:th note in the sequence is defined by its fundamental frequency n_i, onset time, and offset time. Here, fundamental frequency (F0) values are expressed on a scale of unrounded MIDI note numbers $f_{\text{MIDI}} = 69+12\log_2(f_{\text{Hz}}/440)$.

After transcribing the melody into notes, a detailed F0 trajectory is estimated for each note n_i as described in [4]. For each frame t in the time region between the note onset and offset, the maximum-salience F0 is obtained by

$$\widehat{f}_t = \underset{f}{\operatorname{argmax}}\, s_t(f), \text{ where } |f - n_i| < 3. \qquad (1)$$

The condition $|f - n_i| < 3$ in Eq. (1) limits the possible search range in frequency to ± 3 semitones from the transcribed note F0 n_i. The mean $\bar{s}_i$ of the detected salience maxima for the note is then calculated. Only the F0 values $\widehat{f}_t$, for which $s_t(\widehat{f}_t) > \alpha\bar{s}_i$, are preserved for further processing, where α was empirically set to 0.8. This is done to avoid less reliable frames. The preserved F0 values are finally used to interpolate a detailed F0 trajectory at 10 ms time grid for the note time regions, i.e., the trajectory is not defined during rests. The interpolation is performed using piecewise cubic splines.

Figure 1 shows an example transcription with melody notes n_i (the black boxes) and the detailed F0 trajectory (the white line). The gray-level intensity on the background indicates the salience values $s_t(f)$, where darker color shows

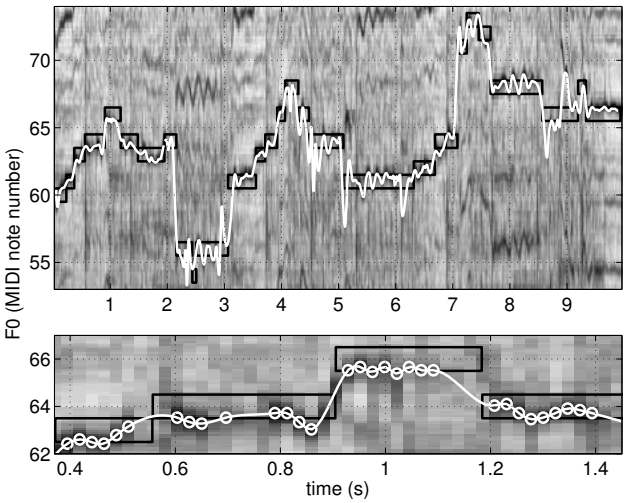

Figure 1. Automatically transcribed melody and its detailed F0 trajectory for the beginning of a verse in "Sick Sad Little World" by Incubus. See text for details.

greater salience. The lower panel shows a close-up of a few notes, where the white circles indicate the preserved F0 values. Piecewise cubic splines are fitted to these points and then used to interpolate the detailed F0 trajectory.

3 ABOUT THE IMPLEMENTATION

The melody transcription method has been implemented with C++ and it takes about five seconds to process one minute of audio on a 3.2 GHz Pentium 4 processor. The entire method with the F0 trajectory estimation implemented as Matlab m-file and mex-file takes about six seconds.

4 RESULTS

The submitted method was ranked third in the overall summary results (overall accuracy 71.1%). The methods by Cancela (1st) and Durrieu *et al.* (2nd) performed clearly better with overall accuracies 76.1% and 75.0%, respectively. The submitted method runs clearly faster than the more accurate methods (over 400 and 20 times faster than the first-ranked and the second-ranked methods, respectively).

5 ACKNOWLEDGMENTS

The authors would like to thank the IMIRSEL group for running the evaluations and organizing MIREX in general, the task organizers, people involved in the discussions, and all the participants.

6 REFERENCES

[1] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: Popular, classical, and jazz music databases. In *Proc. 3rd International Conference on Music Information Retrieval*, pages 287–288, 2002.

[2] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: Music genre database and musical instrument sound database. In *Proc. 4th International Conference on Music Information Retrieval*, 2003.

[3] G. Poliner, D. Ellis, A. Ehmann, E. Gómez, S. Streich, and B. Ong. Melody transcription from music audio: Approaches and evaluation. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1247–1256, May 2007.

[4] M. Ryynänen, T. Virtanen, J. Paulus, and A. Klapuri. Accompaniment separation and karaoke application based on automatic melody transcription. In *Proc. 2008 IEEE International Conference on Multimedia and Expo*, pages 1417–1420, Hannover, Germany, June 2008.

[5] M. P. Ryynänen and A. P. Klapuri. Automatic transcription of melody, bass line, and chords in polyphonic music. *Computer Music Journal*, 32(3):72–86, 2008.