POLYPHONIC MUSIC TRANSCRIPTION USING NOTE EVENT MODELING 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 "Multiple Fundamental Frequency Estimation & Tracking". The task includes two subtasks: i) estimation of several fundamental frequencies (F0s) at frame-level and ii) detection of note events performed with pitched instruments. Our submission consists of a previously published method for the automatic transcription of pitched instruments in polyphonic music. The method was originally designed to perform the latter subtask but can handle the multiple-F0 estimation task in a straightforward manner. The method performed well for the tasks in MIREX 2007, and therefore, the submission is a replica of our last year's submission.

1 INTRODUCTION

Transcription of music refers to the analysis of an acoustic music signal for producing a parametric representation of the signal. In this case, the parametric representation is a set of notes. A note is here defined by a discrete note pitch with an onset and an offset time. In automatic music transcription, notes are extracted from a music signal by a machine.

The submitted method for polyphonic music transcription has been published in [5]. The method and its parameters in this submission are exactly the same as in the above publication. Figure 1 shows a block diagram of the method. First, both the left and the right channel of an audio recording are processed frame-by-frame with a multiple-F0 estimator to obtain several F0s and their related features. The F0 estimates are processed by a musicological model which estimates musical key and chooses between-note transition probabilities. Note events are described with HMMs which allow the calculation of the likelihoods for different note events. Finally, a search algorithm finds multiple paths through the models to produce transcribed note sequences. The method is directly applicable to monaural audio input as well. Later, we have applied similar framework in melody transcription [3] and in bass line transcription [4].

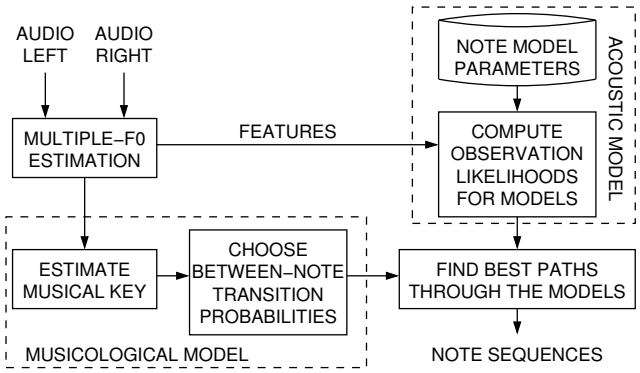

Figure 1. A block diagram of the transcription method.

2 METHOD DESCRIPTION

We briefly introduce the method in the following. For more details, please see [5].

The front-end of the transcription method is a multiple-F0 estimator proposed in [1]. F0s are estimated one at a time, the found sounds are canceled from the mixture, and the estimation is repeated for the residual. In addition, the method performs detection of the onsets of pitched sounds by observing positive changes in the estimated strengths of different F0 values. Here we used the estimator to analyze audio signal in 92.9 ms frames overlapped with 11.6 ms interval between the beginnings of successive frames. In each frame, the estimator produces five distinct fundamental frequency values per one audio channel.

The transcription method uses three probabilistic models: a note event HMM, a silence model, and a musicological model. The note HMM uses the output of the multiple-F0 estimator to calculate likelihoods for different notes, and the silence model corresponds to time regions where no notes are sounding. The musicological model controls transitions between note HMMs and the silence model, analogous to a "language model" in automatic speech recognition. Transcription is done by searching for disjoint paths through the note models and the silence model.

Note events are described with a three-state HMM where the states represent the typical values of the features in con-

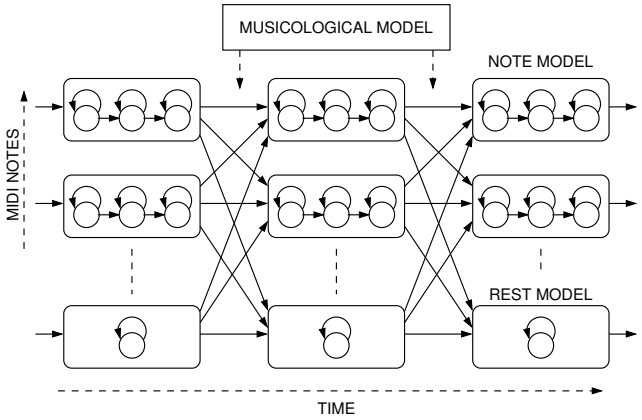

Figure 2. The network of note models and the rest model.

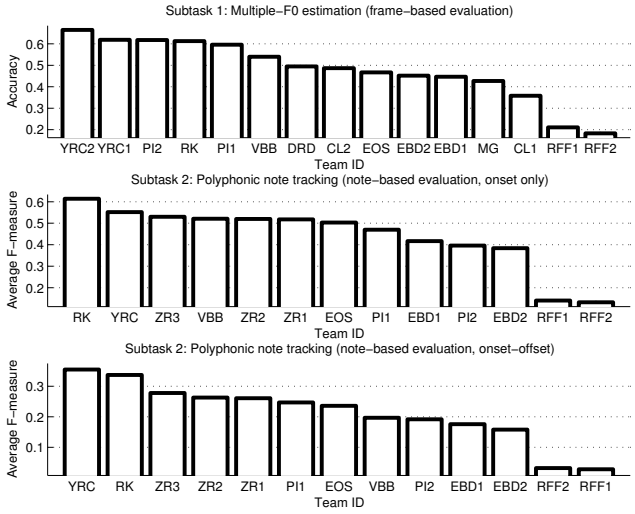

Figure 3. Overall results for subtasks 1 and 2.

secutive temporal segment of note events. The method allocates one note HMM for each MIDI note number in desired pitch range. A silence model, which is a simple one-state HMM, models the time regions where no notes are sounding.

The musicological model controls transitions between note HMMs and the silence model. The musicological model is based on an idea that some note sequences are more common than others in a certain musical key. The musicological model first finds the most probable relative-key pair (e.g., C major and A minor) using a musical key estimation method. This key pair is then used to choose transition probabilities between note models and the silence model. The transition probabilities between models are given by a bigram trained on a large database of monophonic melodies.

The note models and the silence model constitute a network of models shown in Figure 2. A path through the network can be found using the Token-passing algorithm [6] after calculating the observation likelihoods for note model states and the silence model, and choosing the transition probabilities between different models. This simultaneously produces the discrete MIDI note numbers and note segmentation, i.e., the note onsets and offsets. A note starts when the found path enters into a note model and ends when the path exits the model. Rests are produced when the path goes through the silence model. In order to transcribe polyphonic music, we need to find several paths through the network. For this, we apply the Token-passing algorithm iteratively by prohibiting the use of note models (except the silence model) on the found path during the following iterations. As a result, the method finds several non-overlapping note sequences.

Finally, the transcribed notes are converted into the required output format for the MIREX task. For the note detection task, the method output is similar to the required format. For the multiple-F0 estimation task, the transcribed notes directly determine the sounding pitches or silence on a 10 ms time grid.

3 ABOUT THE IMPLEMENTATION

The method has been implemented as Matlab M-files and MEX-files, and it runs in Linux Matlab versions 6.5, 7.3, and 7.4. The execution time on a 3.2 GHz Pentium 4 processor is about four times the length of a monaural audio recording. Most of the time is spent in the multiple-F0 estimation module which could be replaced with a faster implementation (e.g., [2]).

4 RESULTS

Figure 3 shows the summary results for both tasks. Since our submission is the same as last year, our results are identical to MIREX 2007 evaluation (team "RK"). Pleasingly the results have been improved by other teams and for the first subtask (frame-based multiple-F0 estimation), the results are clearly improved by the submissions of C. Yeh, A. Roebel, and W.-C. Chang (team "YRC") and A. Pertusa and J. M. Iñesta (team "PI"), for example. Our submission was ranked as third in this subtask.

Our method was originally designed to produce MIDI notes as an output, and in the note tracking subtask, our method was top-ranked also this year according to the "onset only" criterion. When also note offsets are taken into account, the submission by the team "YRC" performs better.

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] A. Klapuri. A perceptually motivated multiple-F0 estimation method. In *Proc. 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 291–294, New Paltz, NY, USA, Oct. 2005.

[2] A. Klapuri. Multiple fundamental frequency estimation by summing harmonic amplitudes. In *Proc. 7th International Conference on Music Information Retrieval*, pages 216–221, 2006.

[3] M. Ryynänen and A. Klapuri. Transcription of the singing melody in polyphonic music. In *Proc. 7th International Conference on Music Information Retrieval*, 2006.

[4] M. Ryynänen and A. Klapuri. Automatic bass line transcription from streaming polyphonic audio. In *Proc. 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1437–1440, Honolulu, Hawaii, USA, Apr. 2007.

[5] M. P. Ryynänen and A. Klapuri. Polyphonic music transcription using note event modeling. In *Proc. 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 319–322, New Paltz, NY, USA, Oct. 2005.

[6] S. J. Young, N. H. Russell, and J. H. S. Thornton. Token passing: a simple conceptual model for connected speech recognition systems. Technical report, Cambridge University Engineering Department, July 1989.