NOTE EVENT MODELING FOR AUDIO MELODY EXTRACTION

Matti Ryynänen
Institute of Signal Processing,
Tampere University of Technology
P.O.Box 553, FI-33101 Tampere, Finland
matti.ryynanen@tut.fi

Anssi Klapuri
Institute of Signal Processing,
Tampere University of Technology
P.O.Box 553, FI-33101 Tampere, Finland
anssi.klapuri@tut.fi

ABSTRACT

This paper outlines a method for the MIREX05 contest "Audio Melody Extraction". The method aims at extracting the fundamental frequency (F0) trajectory of the predominant melody, such as the lead singing, within polyphonic and multitimbral audio signals. The method is based on a multiple-F0 estimator front-end followed by the probabilistic modeling of note events and their relationships. The performance of the method was ranked second best in the MIREX05 evaluations.

Keywords: multiple-F0 estimation, music transcription, hidden Markov model

1 INTRODUCTION

Transcription of music refers to the process of generating parametric notations, i.e., musical transcriptions, for musical performances. The automatic transcription of realistic music signals, including polyphonic and multitimbral audio, is a very challenging task which has become an active research topic during the last ten years.

The method outlined in this paper is a slight modification of our method proposed for the automatic transcription of polyphonic music [4]. The main difference between these two is the format of output: the original method is intended for producing polyphonic MIDI-type transcriptions from music, whereas the outlined method is required to produce non-quantized F0 trajectory of the predominant melody in a music performance.

This paper briefly describes the proposed method. For detailed explanations of the multiple-F0 estimator and the note-event modeling method, please refer to [2], [4].

2 METHOD DESCRIPTION

Figure 1 shows the block diagram of the method. First, an audio recording is frame-wise processed with a multiple-F0 estimator to obtain two F0s and their saliences in each frame. The musicological model uses the F0 estimates to estimate musical key and to choose between-note transition probabilities accordingly. Note events are described with HMMs which allow the calculation of the likelihoods for different note events. A search algorithm then finds the best path through the models to produce a transcribed sequence of notes and rests. The output F0 trajectory is obtained by choosing the closest F0 estimates to the transcribed notes and by linearly interpolating output F0s at every 10 ms.

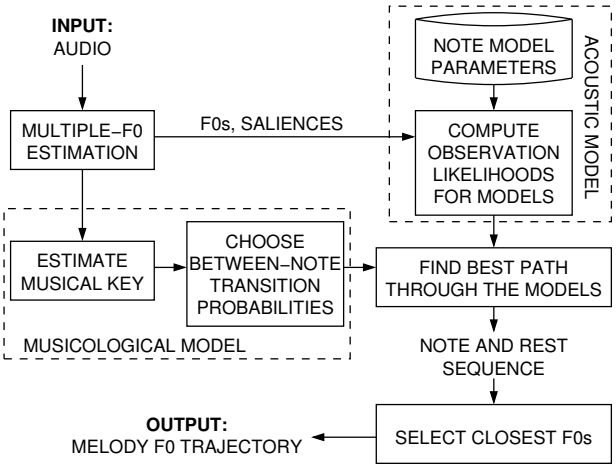

Figure 1: The block diagram of the method.

2.1 Multiple-F0 Estimation

The front-end of the method is a multiple-F0 estimator proposed in [2]. The estimator applies an auditory model where an input signal is passed through a 70-channel bandpass filterbank and the subband signals are compressed, half-wave rectified, and lowpass filtered with a frequency response close to $1/f$. Short-time Fourier transforms are then computed within the bands and the magnitude spectra are summed across channels to obtain a summary spectrum where all the subsequent processing takes place. Periodicity analysis is carried out by simulating a bank of comb filters in the frequency domain. F0s are estimated one at a time, the found sounds are canceled from the mixture, and the estimation is repeated for the residual.

We use the estimator to analyze audio signal in overlapping 92.9 ms frames with 23.2 ms interval between the beginnings of successive frames. Here, the estimator produces two F0s and their saliences in each frame.

2.2 Note Event Modeling

The transcription system applies 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 main melody notes are sounding. The F0 estimates are processed by a musicological model which estimates musical key and chooses bigram between-note transition probabilities in a manner similar to [3], [4].

Note events are described with a three-state left-to-right HMM. The note-HMM states represents the typical values of the F0s and their saliences during the main melody notes. We allocate one note HMM for each MIDI note number $n = 44, \ldots, 84$, covering F0 range from 100 Hz to 1 kHz. We used the annotated main melodies of the RWC popular music database to train the note event model [1]. The silence model is a 1-state HMM for which the the observation likelihood is the negation of the greatest observation likelihood in any state of any note model at a time.

The note models and the silence model constitute a network of models. The optimal path through the network is found using the Token-passing algorithm [5] after calculating the observation likelihoods for note model states and the silence model, and choosing the transition probabilities between different models by the musicological model. The found path is the transcribed sequence of notes and rests for the main melody.

3 METHOD EVALUATION RESULTS

The MIREX05 evaluation results of the method are given in the following. The reported results are the average values of the correctly transcribed frames within the test set including 50 music excerpts from different musical genres. When octave errors are ignored, the corresponding result is given within parentheses.

The method correctly transcribed 63.9% (67.3%) of voiced and unvoiced portions. For the voiced segments, the method estimated the correct F0 in 66.0% (69.9%) of the frames. The average F-measure was reported to be 70.2% (73.9%). When melody segmentation errors were ignored, the results were 68.2% (73.7%). The temporal boundaries of melodic segments were correctly decided in 81.9% of the cases. In the overall evaluation, the method performed second best among the nine participating methods.

For an Intel 1.7 GHz computer with Linux OS, the method takes approximately 20–30 times the duration of an audio recording with sampling rate 44.1 kHz. Most of the computational effort is required by the multiple-F0 estimator. The proposed method was among the four slowest systems and much slower than the winning method by K. Dressler. However, no optimization was carried out which would expectedly reduce the computational load.

ACKNOWLEDGEMENTS

We wish to thank the hardworking MIREX05 team for arranging and running the evaluation contest.

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*, October 2002.

[2] A. Klapuri. A perceptually motivated multiple-F0 estimation method. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2005.

[3] M. P. Ryynänen and A. Klapuri. Modelling of note events for singing transcription. In *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio*, October 2004.

[4] 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*, October 2005.

[5] 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.