MULTIPLE FUNDAMENTAL FREQUENCY ESTIMATION & TRACKING IN POLYPHONIC MUSIC FOR MIREX 2010

F. Canadas, F. Rodriguez, P. Vera, N. Ruiz and J. Carabias
Telecommunication Engineering Department, University of Jaen
Polytechnic School, Linares, Jaen, Spain
fcanadas@ujaen.es

ABSTRACT

The goal of the MIREX (Music Information Retrieval Evaluation eXchange) contest is to compare state-of-the-art algorithms and systems relevant for Music Information Retrieval. This paper briefly describes our algorithm for the MIREX 2010 Multiple Fundamental Frequency (multiple-F0) Estimation & Tracking task. Specifically, it is submitted for the first two subtasks: i) estimation of several fundamental frequencies (F0s) at frame-level and ii) Tracking note contours on a continuous time basis. Our algorithm, designed as a trade-off between accuracy rate and computational cost, is based on the minimization of a spectral distance which is calculated using the convolution between candidates and harmonic patterns belonging to the same combination at frame-level and maximization of temporal continuity at temporal interval-level.

1. INTRODUCTION

The typical process of writing a musical score has been made by handwriting. However, this process is high time-consuming and requires musical education. The goal of an automatic music transcription system is to efficiently extract a musical score from an audio signal. In this context, a note is defined by a pitch and an onset-offset time. For this purpose, polyphonic musical sounds are suitable signals for the problem of multiple-F0 estimation as well as vocal sounds to speech recognition.

A multiple-F0 estimator is the main block of a automatic music transcription system. Tempo detection and key estimation complement multiple-F0 estimation to correctly extract the music score. Detecting all active pitches in a polyphonic signal is still a challenging problem despite the huge research effort that has been made in the last years. For an illustrative overview of different methods for multiple-F0 estimation, we refer to [1] [2].

2. ALGORITHM DESCRIPTION

Our submitted algorithm is composed of two stages: onset detection and a jointly multiple-F0 estimator. Next, each stage is explained in details.

The first stage is composed by an onset estimator which defines each temporal interval between two consecutive onsets. We assume that a temporal interval locates an excerpt of signal in which the spectral content varies slowly within it (non-transients). We propose to analyze the time envelope of a collection of sinusoids, extracted using a sinusoidal model, which clearly allows to discriminate transients sinusoids (onset times) from stationary ones. Onset information, from frequency domain, is provided by time envelope of each isolated sinusoid. To compute the time envelope, an 1-order linear prediction is applied to each spectral peak neighborhood. Thus, the calculated pole will be located at the middle of the energy burst and it is moved (phase information) along temporal frames. Linear prediction coefficient is calculated with the autocorrelation function of a few samples around each previous spectral peaks. Next, most significant perceptual sinusoids at each MIDI-note range are selected. In order to avoid spurious onsets, a clustering process is performed to select only sequences of poles which mark the same onset time. To conclude, the sum of their perceptual importance is computed followed by a Hidden Markov Model to determine onset activation times.

The basic idea of the second stage is to perform a jointly multiple-F0 estimation, within each temporal interval provided by the onset detector, using a criterion based on distances and level of active presence of each possible pitch. Each temporal interval is composed of a variable number of fixed length frames.

From now on, all actions are referred to the analysis of an arbitrary frame. Using the STFT from the considered frame, we compute a new spectrum composed of only significant spectral peaks using a frequency-dependent threshold [3]. Next, we construct all harmonic patterns whose fundamental frequency is located between MIDI number 36 to MIDI number 95, a typical interval of multiple-F0 analysis [4] [5]. Considering a joint multiple-F0 estimation, an exhaustive search for those combinations which are composed of the most predominant harmonic patterns is performed taking into account a suitable trade-off between accuracy rate and computational cost. Supposing

that an overlapped partial amplitude is equal to the sum of all overlapped partial amplitudes shared out among all partials involved in the spectral collision, each overlapped partial is re-estimated by linear interpolation of the nearest non-overlapped partials [6]. Under assumption that the amplitude spectrum of a polyphonic music signal is additive, we maximizes a spectral distance which is calculated using the convolution between candidates (fundamental frequency) and re-estimated patterns belonging to the same combination at frame-level and maximizes temporal continuity at temporal interval-level. In this manner, the most likely combination of actives pitches at the considered frame is selected using a criterion which penalizes abrupt changes among input-modeled partials and rewards the spectral similarity for overall spectral patterns between the input spectrum and the modeled spectrum for each combination.

From now on, all actions are referred to the analysis of an arbitrary temporal interval. Once a collection of active pitches has been estimated at each frame of the same temporal interval, we set as active pitch those ones which are active in most of the frames belonging to the temporal interval. An example of multiple-F0 estimation performed by our algorithm is shown in Figure 1 in which most of reference notes (black rectangles) are correctly estimated (white rectangles).

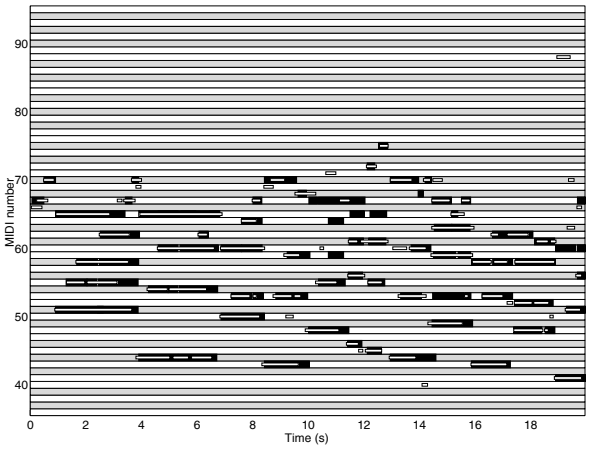

Figure 1. Transcription of a polyphonic musical excerpt. *x-axis* indicates time in seconds. *y-axis* indicates MIDI notes from MIDI number 36 to MIDI number 95.

3. ACKNOWLEDGEMENTS

The authors would like to thank the MIREX 2010 organization committee and the IMIRSEL team for running the MIREX evaluations.

This work was supported by the Spanish Ministry of Education and Science under Project TEC2009-14414-C03-02.

4. REFERENCES

[1] Klapuri, A. and Davy, M. "Signal Processing Methods for Music Transcription", *Springer Science+Business Media LLC*, 2006.

[2] Wang, D. and Brown, G. "Computational Auditory Scene Analysis : Principles, Algorithms and Applications", *IEEE Press / Wiley*, 2006.

[3] Every, M. and Szymanski, J. "Separation of synchronous pitched notes by spectral filtering of harmonics", *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no 5, pp. 1845-1856, 2006.

[4] Duan, Z. and Zhang, C. "A probabilistic approach to multiple fundamental frequency estimation from the amplitude spectrum peaks", *in Proc. Music, Brain and Cognition workshop in the Twenty-first Annual Conference on Neural Information Processing Systems (NIPS)*, 2007.

[5] Klapuri, A. "Multipitch analysis of polyphonic music and speech signals using an auditory model", *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, no 2, pp. 255-266, 2008.

[6] Pertusa A., Inesta J.M., "Multiple Fundamental Frequency estimation using Gaussian smoothness," Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP 2008, pp.105-108, Las Vegas, USA, 2008.