ONLINE MELODY EXTRACTION: MIREX 2012

Vipul Arora
Department of Electrical Engineering
Indian Institute of Technology, Kanpur
vipular@iitk.ac.in

Laxmidhar Behera
Department of Electrical Engineering
Indian Institute of Technology, Kanpur
lbehera@iitk.ac.in

ABSTRACT

Melody is the pitch perceived from the most dominant source in the polyphonic music. This paper describes our algorithm for melody extraction submitted for the audio melody extraction task of the Music Information Retrieval Evaluation eXchange (MIREX) 2012. The task consists of two parts: melody estimation and voicing detection. The presented algorithm performs the melody estimation in an online fashion. It is based on tracking of the harmonic clusters followed by selection of the predominant source. The voicing detection is done in off-line fashion based on the harmonic energy in the clusters.

1. INTRODUCTION

Generally pitch has one to one correspondence with the fundamental frequency (F0) of the acoustic source. Melody estimation task refers to extracting the predominant F0 of the musical source with respect to time, while the voicing detection task refers to detecting whether the dominant melodic source is active/inactive in a particular time frame.

The following properties form the foundation of our method:

1. A melodic source has a harmonic structure, characterised by spectral peaks at the multiples of F0.
2. The melodic sources evolve slowly in time.
3. Predominant source is the one having maximum power in the harmonic structure. This power is quantified using various measures like Fourier transform, amplitudes of the peaks as well as deviation of the frequencies of the peaks from the harmonic constraints.

The proposed algorithm consists of two parts: melody estimation and voicing detection. The melody estimation part is based on an on-line framework, while the voicing detection part is implemented in off-line way. The melody estimation part can further be seen as two serial modules: 1) harmonic source tracking, which simultaneously tracks several harmonic sources, and 2) dominant source selection, which selects one of the tracked sources as the dominant melodic source.

The article is organized as follow: Section 2 presents the on-line melody estimation module. Section 3 describes the voicing detection module.

2. ON-LINE MELODY EXTRACTION

2.1 Feature Extraction

We divide the signal into short time overlapping frames and transform each frame to spectral domain. The log spectrum of the music signal is extracted using the magnitude of 1024-point short time Fourier transform (STFT) calculated using a 50ms long hanning widow, with a hop size of 10ms. Spectral peaks (called as partials) are extracted which forms the partial space representation of the signal. Also, the inverse Fourier transform (IFT) of the aforementioned log spectrum is calculated. The partial space and the magnitude of the inverse Fourier transform are used for tracking the harmonic sources and for selecting the dominant melodic source. They are also used for calculating the voicing detection score, as described in Section 3.

2.2 Harmonic Cluster Tracking

The top three peaks in the IFT of the log spectrum representation of the signal are used to initialize the new harmonic clusters. This is a quick and robust way to initialize the clusters.

These clusters are tracked in time in the partial space representation, by taking care of both the harmonic structure as well as the evolutionary constraints simultaneously. The positions of the partials of a harmonic cluster are predicted for the current time frame using their frequencies and phases in the previous frames [1] as well as their conformation to the harmonic structure [2, 3].

The cluster tracks are terminated when the amplitudes of their partials become very weak or if their F0 goes out of the desired range.

2.3 Dominant Source Selection

At a time, there are at most five clusters being tracked. The one having maximum recognition score is selected as the dominant source. This recognition score is calculated as a product of several measures like: 1) the IFT of the log spectrum, at the sinusoidal frequency equal to F0, 2) Amplitudes of the partials, and 3) the deviation of partial frequencies from the multiples of F0. To take the evolutionary constraints into consideration, the recognition score for each cluster is smoothed using a smoothing filter.

In this way, the on-line melody extraction module produces an estimate of the predominant pitch at every time frame.

3. VOICING DETECTION

The second problem of the task is to detect whether the dominant melodic source is active or not, for each time frame. Similar to the idea of normalized harmonic energy [4], we use the sum of the squares of the amplitudes of the partials of the dominant cluster at each time frame. This score is normalized by its maximum attained value in the entire input clip. The frames with the value of the voicing score more than a threshold are classified as voiced and the others are labeled as unvoiced.

4. REFERENCES

[1] T. Virtanen: *Audio Signal Modeling with Sinusoids Plus Noise*, M.S. Thesis, Tampere University of Technology, Department of Information Technology, 2000.

[2] V. Rao and P. Rao: "Vocal melody extraction in the presence of pitched accompaniment in polyphonic music," *IEEE Trans. Audio, Speech and Lang. Process.*, Vol. 18, pp. 2145–2154, 2010.

[3] J. Salamon, E. Gomez, and J. Bonada: "Sinusoid extraction and salience function design for predominant melody estimation," *Proc. Int. Conf. on Digital Audio Effects (DAFx)*, pp. 73–80, 2011.

[4] V. Rao and P. Rao: "Vocal melody detection in pitched accompaniment using harmonic matching methods," *Proc. of the 11th International Conference on Digital Audio Effects (DAFx-08)*, 2008.