MIREX2014: AUDIO MELODY EXTRACTION

Yukara Ikemiya **Kazuyoshi Yoshii** **Katsutoshi Itoyama**
Department of Intelligence Science and Technology
Graduate School of Informatics, Kyoto University
{ikemiya, yoshii, itoyama}@kuis.kyoto-u.ac.jp

ABSTRACT

This paper describes our submission for the audio melody extraction task of the Music Information Retrieval Evaluation eXchange (MIREX 2014). Our algorithm first separates the vocal spectra from polyphonic sound spectra. Melody extraction and vocal activity detection are applied to the separated spectra.

1. INTRODUCTION

Automatic melody extraction is an important task for music information retrieval (MIR), and many methods for this task are proposed. Most methods directly calculate candidates of fundamental frequency (F0) from polyphonic spectra and select most-likely peaks in them. Since they premise that vocal part has the most predominant harmonic structure in polyphonic signal, the F0 estimation accuracy decreases rapidly when accompanying sound is larger than vocal sound.

2. MELODY EXTRACTION

2.1 Vocal Separation

We first extract the power spectra from an input signal using short time Fourier transform (STFT). We use a 2048-point hamming window and a hopsize of 10 [*msec*] for the STFT.

The vocal spectra are separated by applying Robust PCA (RPCA) and binary time-frequency masking to a STFT spectrogram [1]. A trade-off parameter k of RPCA is set to 1.0. The separated vocal spectra are used for calculation of a salience function and vocal activity detection.

2.2 Salience Function

The frequency resolution of an STFT spectrum is increased by spline interpolation. Frequencies of 200 bins per octave on a cent scale (6 cents per bin) are calculated.

We use subharmonic summation method with amplitude weighting [2] for a salience function. This function is calculated as:

$$\mathrm{SF}(t,f) = \mathrm{A}(t)\sum_{n=1}^{N}(0.84)^{n-1}\mathrm{S}(t, f + 1200\log_2 n) \quad (1)$$

where $\mathrm{S}(t,f)$ denotes an interpolated spectrum for which f and t are a cent-scale frequency and a time index, respectively. The N is the number of harmonics being considered and the $\mathrm{A}(t)$ is a normalization factor for each time index. The frequency range from 30 to 4000 [*Hz*] is considered.

2.3 F0 Estimation using Viterbi Search

Using the salience function, vocal F0 contour is calculated as:

$$\hat{F} = \underset{f_1,\ldots,f_T}{\arg\max}\sum_{t=1}^{T-1}\{\log \mathrm{SF}(t,f_t) + \log \mathrm{T}(f_t - f_{t+1})\} \quad (2)$$

where $\mathrm{T}(f)$ denotes an F0 transition probability of f cents transition. We use the Laplace distribution as a function $\mathrm{T}(\cdot)$ described in [3]. This can be effectively computed using the Viterbi search. We assume that the vocal F0s exist in the frequency range from 100 to 700 [*Hz*].

3. VOCAL ACTIVITY DETECTION

3.1 VAD based on HMM

We apply vocal activity detection based on a hidden Markov model (HMM) that transitions between vocal state s_v, and non-vocal state s_n. The sequence of vocal and non-vocal states is estimated as:

$$\hat{S_H} = \underset{s_1,\ldots,s_T}{\arg\max}\sum_{t=1}^{T-1}\{\log p(\mathbf{x}|s_t) + \log p(s_t|s_{t+1})\} \quad (3)$$

where $p(\mathbf{x}|s)$ denotes an output probability of state s with an acoustic feature vector $\mathbf{x}$, and $p(s_i|s_j)$ denotes a state transition probability from state s_j to state s_i.

We use 13-th order LPC-derived mel cepstral coefficients (LPMCCs) as an acoustic feature vector. LPMCCs are mel-cepstral coefficients of an LPC spectrum. The $p(\mathbf{x}|s)$ is a 64-components Gaussian mixture model (GMM) and previously trained for each state. Training data for the GMMs are 100 popular songs from the "RWC Music Database: Popular Music" (RWC-MDB-P-2001) [4] that are applied vocal separation using RPCA [1].

3.2 VAD based on Thresholding

It is difficult to detect short vocal sections between two notes by HMM-based VAD. Therefore, we use thereshold-ing method for VAD. First we design a cost function for thresholding as follows.

$$\mathrm{CF}(t) = \sum_{f} \left\{ \frac{1}{H_f} \sum_{h=1}^{H_f} \mathrm{S}(t, hf) \right\}^{1.7} \tag{4}$$

where H_f is the number of all harmonics within 4000 [Hz] for each frequency f. Using this function, vocal and non-vocal state are estimated by thresholding.

$$s_t = \begin{cases} s_v & \mathrm{CF}(t) > k \\ s_n & otherwise \end{cases} \tag{5}$$

where k is a threshold.

4. REFERENCES

[1] P. S. Huang, S. D. Chen, P. Smaragdis and M. H. Johnson: "Singing-Voice Separation from Monaural Recordings using Robust Principal Component Analysis," *Proc. ICASSP*, pp. 57-60, 2012.

[2] D. J. Hermes: "Measurement of Pitch by Subharmonic Summation.," *J Acoust Soc Am.*, pp. 257-264, 1988.

[3] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata and H. G. Okuno: "F0 Estimation Method for Singing Voice in Polyphonic Audio Signal Based on Statistical Vocal Model and Viterbi Search," *Proc. ICASSP*, vol. 5, pp. 253-256, 2006.

[4] M. Goto, H. Hashiguchi, T. Nishimura and R. Oka: "RWC Music Database: Popular, Classical, and Jazz Music Databases.," *Proc. ISMIR*, pp. 287-288, 2002.