MIREX2015: 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 2015). This submission is an extended work from [1].

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). The vocal spectra are separated by applying Robust PCA (RPCA) and binary time-frequency masking to a STFT spectrogram [2]. The separated vocal spectra are used for calculation of a salience function and vocal activity detection.

2.2 F0 Estimation using Viterbi Search

Given a time-frequency salience spectrogram H base on subharmonic summation (SHS) [3], vocal F0 contour is calculated as:

$$\hat{F} = \mathop{\arg\max}_{f_1,\ldots,f_T} \sum_{t=1}^{T-1} \{\log \mathrm{H}(t, f_t) + \log \mathrm{T}(f_t - f_{t+1})\} \qquad (1)$$

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 [4]. This can be effectively computed using the Viterbi search. We assume that the vocal F0s exist in the frequency range from 80 to 720 [*Hz*].

2.3 Vocal Activity Detection

We use theresholding method for vocal activity detection (VAD) based on volume dynamics of a separated singing voice signal.

3. REFERENCES

[1] Y. Ikemiya, K. Yoshii and K. Itoyama: "Singing Voice Analysis and Editing based on Mutually Dependent F0 Estimation and Source Separation," *Proc. ICASSP*, pp. 574-578, 2015.

[2] 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.

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

[4] 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.

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