MIREX2016: AUDIO MELODY EXTRACTION USING NEURAL NETWORK

Chung-Che Wang[1], Zhe-Cheng Fan[2], Jyh-Shing Roger Jang[2], and Kazuyoshi Yoshii[3]
[1]Dept. of CS, National Tsing Hua Univ., Taiwan
[2]Dept. of CSIE, National Taiwan Univ., Taiwan
[3]Dept. of IST, Kyoto Univ., Japan
{geniusturtle, lambert.fan, jang}@mirlab.org, yoshii@kuis.kyoto-u.ac.jp

ABSTRACT

This extended abstract presents our submission to the MIREX 2016 audio melody extraction task, which uses neural network for model training and melody recognition.

1. INTRODUCTION

Melody extraction is an important topic of music information retrieval. Currently most of the methods are signal-processing based [1]. This motivates us to investigate how neural network performs in this task.

2. MELODY EXTRATION

2.1 Network Settings

A network with three hidden layers (1,024 nodes for each layer) is used for our submissions. Since the frame size is 1,024, the number of input nodes is 513. The number of output nodes depends on the data, which will be described in the next sub-section.

We use Chainer [2] as the implementation of our network. Sigmoid function is used as the activation function and Adam is used as the optimization method. Detailed parameters are the default settings of Chainer.

2.2 Training Data and Output Range

Some or all of the following datasets are used for our different submissions:

- 1KP: partial of the MIR-1K dataset [3]. More specifically, the first male (abjones) and first female (amy) were selected.
- 1KPG: generated data based on 1KP.
- M05T: *mirex05TrainFiles*, containing 13 clips, with average length of 30 seconds [4].
- GM: generated MIDI files using [5].

Our different versions of submissions use different training data and output range:

- WFJY1: 1KP, 1KPG, M05T, and GM (semitone range is 80 to 97, with step value of 0.25 semitones) are used for training. The output semitone range is 29 to 97, with 4 bins per semitone.
- WFJY2: separated vocal (use [6] as implementation) from 1KP and 1KPG (which were mixed at 0dB) are used for training. The output semitone range is 29 to 88, with 4 bins per semitone.
- WFJY3: 1KP, 1KPG, M05T, and GM (semitone ranges are 30 to 43 and 80 to 97, with step value of 0.25 semitones) are used for training. The output semitone range is 29 to 97, with 4 bins per semitone.

For vocal/non-vocal detection, we used 1KP, 1KPG, and M05T for training.

3. REFERENCES

[1] MIREX Wiki. Available: http://www.music-ir.org/mirex/wiki/MIREX_HOME

[2] S. Tokui, K. Oono, S. Hido and J. Clayton, Chainer: a Next-Generation Open Source Framework for Deep Learning, *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS)*, 2015

[3] C.-L. Hsu and J.-S. R. Jang, “On the Improvement of Singing Voice Separation for Monaural Recordings Using the MIR-1K Dataset,” *IEEE Trans. Audio, Speech, and Language Processing*, volume 18, issue 2, p.p 310-319, 2010

[4] G. Poliner and D. Ellis, “A Classification Approach to Melody Transcription,” *Proc. Int. Conf. on Music Info. Retrieval (ISMIR)*, London, September 2005.

[5] K. Schutte, “MIDI file tools for MATLAB, ” Available: https://github.com/kts/matlab-midi/blob/master/src/synth.m

[6] Y. Ikemiya, K. Yoshii, K. Itoyama: “MIREX2015: AUDIO MELODY EXTRACTION,” *extend abstract of the International Symposium on Music Information Retrieval*, 2015.