MIREX 2020 SUBMISSION: AUDIO KEY DETECTION BASED ON AUTOENCODER AND ANALYSIS OF MUSICAL STRUCTURE

Yang Deng
NetEase, Inc.
dengyang02@corp.netease.com

Ziyao Xu
Malong Technologies
ziyxu@malong.com

ABSTRACT

Audio key estimation is the basis on which various musical tasks can be performed. We introduce a system that divides it into two parts: convert audio to MIDI, and key estimation based on MIDI files. For the former, a deep learning model, the autoencoder, is used to detect melody and obtain MIDI data. For the latter, we adopt the method based on the analysis of musical structure.

1. INTRODUCTION

The main idea of our method is to use the pattern of music structure as far as possible to identify the key, so we need to extract melody from the audio. We adopt the network proposed in [1], namely MSnet. This is a lightweight network for melody detection, which consists of encoder, decoder and classifier. The input of MSnet is the CFP representation of audio, and the high-level representation is obtained by the encoder, and then the existence of melody and saliency frequency map are obtained by classifier and decoder respectively. The melody line can be obtained by combining the existence of melody and saliency frequency map. We submitted two systems listed in Table 1, the difference is that the training data of autoencoder is different.

Submission ID	Training Dataset
DX1	MIR-1K[2], MedlyDB[3]
DX2	MedlyDB

Table 1. An overview of the submitted systems.

After extracting the MIDI data, we analyze the music elements such as the main melody, intervals, notes, bars, and chords to estimate the key according to the composition rules.

2. PROPOSED METHODS

Extracting melody from audio is to learn the mapping between a real-valued, dense matrix that represents the input audio and another relatively sparser matrix that represents the melody line. It is similar to semantic pixel-wise segmentation of images, so MSnet adopts a structure similar with SegNet[4] in image segmentation. As shown in Figure 1, the input of the model is the CFP representation of the audio. The high-level feature of the audio is obtained by the encoder, and the saliency frequency map is obtained by the decoder. Since the melody does not exist in all frames, the network predicts the existence of melody by the classifier. By concatenating the saliency frequency map with the existence of melody, the melody line can be detected.

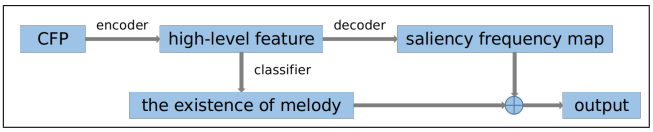

Figure 1. Model structure

After obtaining MIDI data from the melody line, the analysis of musical structure is performed, which includes the analysis of the musical elements such as main melody, intervals, notes, bars, and chords.

After analyzing above elements, the system obtains the tonic and mode of the input audio file.

3. REFERENCES

[1] Hsieh, Tsung Han, L. Su, and Y. H. Yang. "A Streamlined Encoder/Decoder Architecture for Melody Extraction." (2018).

[2] https://sites.google.com/site/unvoicedsoundseparation/mir-1k

[3] R. Bittner, et al. "MedleyDB: A multitrack dataset for annotation-intensive MIR research." Proc. ISMIR, 2014, [Online] http://medleydb.weebly.com/

[4] V. Badrinarayanan, A. Kendall, and R. Cipolla. "Seg-Net: A deep convolutional encoder-decoder architecture for image segmentation", IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 39, No. 12, pp. 2481–2495, 2017.