MIREX 2019 SUBMISSION: CHORD ESTIMATION

Song-Rong Lee
National Taiwan University
hubert.lee@mirlab.org

I Chien
National Taiwan University
eden.chien@mirlab.org

Tzu-Chun Yeh
National Tsing Hua University
kenshin.yeh@mirlab.org

Jyh-Shing Roger Jang
National Taiwan University
jang@mirlab.org

ABSTRACT

In this submission, we provide a chord estimation system based on the Convolution Neural Network (CNN) for the chord estimation task of MIREX 2019. In the training process, we added several audio files which mixed with environmental noise as noise training set for the robustness of the real-world recordings.

1. INTRODUCTION

Chord estimation is a fundamental task for multiple applications. For example, in music performance assessment, the system needs to recognize the chords correctly for a reasonable assessment result. Music key detection also relies on accurate chord recognition in order to give the right key. Although the chords can be identified by people who are well trained, it is a still a difficult task for machines.

Most of the submissions which are based on neural network have better performance in these years. In our submission, we also do data augmentation including pitch-shift, down sampling and adding noises to expand the training data size for better result. The details of the method is described in section 2 and the training procedure is described in section 3.

2. METHOD

In this section, we will introduce the system overview and the methods applied in our system, including the chord representation, feature extraction and the model architecture that this submission used. The system diagram shows in Figure 1.

2.1 Chord Label Representation

There are five evaluation classes in the chord estimation task. Two of them have to predict both quality and inversion and the other three only include quality. Our submission tried to predict more chord quality types than the evaluation classes define by MIREX because our system aim for the practical usage. By the quality frequency statistics provided by MIREX, our system cover almost 90% type of chords. In addition, this system ignore the inversion part of the chord. The qualities that our system predict and the full label definitions are described in Table 1.

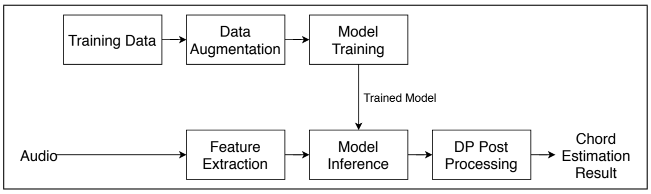

Figure 1. System Overview.

Table 1. Label representation of the chords.

Dimension	**Meaning**
0-11	Root
12	The chord not in our prediction lists
13	Major
14	Minor
15	Seventh
16	Minor seventh
17	Major seventh
18	Augmented
19	Diminished
20	Fifth
21	Major sixth

2.2 Feature Extraction

The system used Constant-Q transform spectrogram as features and is calculated by librosa [2]. The extraction procedure use 192 bins for Constant-Q transform starting from C1 and use 24 bins in one octave. The hop length is set to 2048 and the sample rate of input audio is 16000.

2.3 Model Architecture

The neural network architecture that system used is proposed by Jiang et al. [1] in MIREX 2018. It is based on multiple layer of Convolution Neural Network and a bi-directional LSTM at the end. The model structure is shown in Table 2.

Table 2. Model structure.

Layer type	Parameters
Convolution	16*3*3
Convolution	16*3*3
Convolution	16*3*3
Max Pool	3*3
Convolution	32*3*3
Convolution	32*3*3
Convolution	32*3*3
Max Pool	3*3
Convolution	64*3*3
Convolution	64*3*3
Max Pool	3*4
Bi-directional LSTM	128*2
Fully connected	145

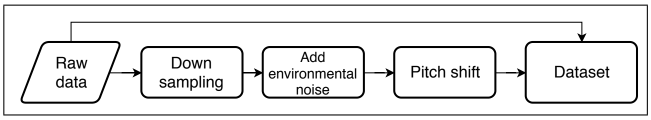

Figure 2. Flowchart of data augmentation.

2.4 Post-processing

In order to smooth the prediction result, the system also applied a dynamic programming based post processing proposed by Jiang et al. [1] in MIREX 2018. The target is to find a sequence of c to minimize J, where c denote the chord prediction. The α that our system use is 0.25.

$$J(c_1, ..., c_n|\texttt{Feature}) \\ = \sum_{i=1}^{n} \log p(c_i|\texttt{Feature}) - \sum_{i=1}^{n-1} \alpha \cdot [c_i \neq c_{i+1}] \quad (1)$$

3. MODEL TRAINING

In this section, we will describe the data we used for training and related details of this submission.

3.1 Datasets

The dataset we used for training the neural network includes McGill Billboard[1] and Isophonics[2]. There are total 819 songs we used in this submission.

3.2 Data Augmentation

The data is augmented by the following steps. First, we down sample the audio. Second, we add the environmental noise to the audio. At the end, we shifted the pitch of the audio from -6 semitone to +5 semitone. The flowchart of the complete procedure is in Figure 2. After the data augmentation, there have a total of 9872 audios.

3.3 Training Process

The system used Adam optimizer with learning rate 1e-4. The model was trained with 200 epochs and set the mini batch as 15. The feature of each audio was padded to the sample shape. When iterating the training data, each sample is a hole audio file.

The total loss is calculated by two subtasks, a root part and a quality part. These subtasks are treated as classification problems and used binary cross entropy as its loss function.

4. REFERENCES

[1] Junyan Jiang, Ke Chen, Wei Li, and Guangyu Xia. Mirex 2018 submission: A structural chord representation for automatic large-vocabulary chord transcription. 2010.

[2] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, 2015.

[1] `https://ddmal.music.mcgill.ca/research/The_McGill_Billboard_Project_(Chord_Analysis_Dataset)/`

[2] `http://www.isophonics.net/`