MIREX 2019 SUBMISSION: CROWD ANNOTATION FOR AUDIO KEY ESTIMATION

Junyan Jiang
Carnegie Mellon University
junyanj@cs.cmu.edu

Gus G. Xia
New York University Shanghai
gxia@nyu.edu

David B. Carlton
Hooktheory, Inc.
dave@hooktheory.com

ABSTRACT

We here propose an early version of our key estimation model trained on a huge crowd-annotated dataset, the Hooktheory dataset. We use a deep learning model, the Convolutional Recurrent Neural Network (CRNN), for key classification. Another submission further adopts a multi-task training strategy by learning a model that jointly predicts several inter-related high-level features (keys, chords, beats & melody notes) with the intuition that different tasks might help correct each other.

1. INTRODUCTION

Submission ID	Algorithm	Training Dataset
JXC1	CRNN (Single-Task)	Hooktheory (may overlap with MIREX test datasets)
JXC2	CRNN (Multi-Task)	

Table 1. An overview of the submitted systems

We submitted two models listed in Table 1. The main idea of our method is to utilize the crowd-annotated labels as a good complementary to current audio key estimation datasets for data-hungry deep learning methods. If the annotation contains other high-level music features (e.g., chords and main melody), they may also help under a multi-task setting, as these inter-related music features might help correct each other.

It is important to notice that both submissions should only be treated as a reference instead of a comparative target because of the possible overlap between the Hooktheory dataset and the MIREX test datasets. The problem might be severe if the overlap rate is large. Fair comparisons should be conducted in the future.

2. PROPOSED METHODS

We adopt a Convolutional Recurrent Neural Network (CRNN) for audio key estimation. The model accepts the spectrogram as input and processes it with a Convolutional Neural Network (CNN) feature extractor first. The feature extractor is identical to [1] with one small modification in JXC2 by replacing batch normalization layers with instance normalization (since we trained the model under a small batch size). Then, a Bi-directional Long Short-Term Memory (Bi-LSTM) layer (with 128 hidden units for JXC1 and 256 hidden units for JXC2 in each direction) is applied to introduce temporal information. The output then goes through a linear unit to get the activation for frame-wise classification.

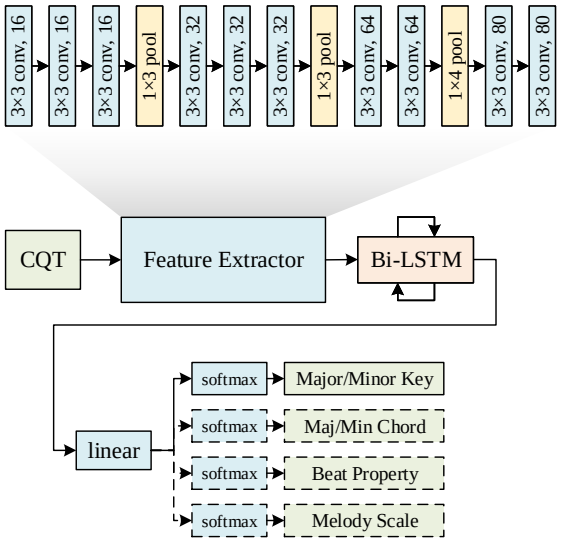

Figure 1. Model structure.

For the multi-task submission, the model predicts the following frame-wise classification task altogether:

1. Key (24 classes): 12 classes for major keys and 12 classes for minor keys;
2. Major/minor chords (25 classes): 1 class for non-chord class, 12 classes for major chords and 12 classes for minor chords;
3. Beat property (3 classes): 1 class for downbeat, 1 class for upbeat, and 1 class for non-beat.
4. Dominant melody notes (13 classes): 12 classes for 12 pitch classes, and 1 class for the case of no dominant melody.

To acquire a global key estimation during model inference, we simply select the most likely key class that maximizes the average frame-wise activation across the whole piece.

3. DATASET

Hooktheory [1] is a crowd-annotated data source with manually labeled keys, chord sequences, tempos, meters, dominant melody notes, and aligned audio pieces.

Mode	Number of Pieces	Percentage (%)
major	7642	49.5
minor	6231	40.4
dorian	645	4.2
mixolydian	518	3.4
lydian	193	1.3
phrygian	164	1.1
locrian	36	0.2

Table 2. Mode distribution in the collected dataset.

In the submission, we collected 15,429 well-aligned pieces with a mode distribution shown in Table 2. We only use the major/minor keys (13,873 pieces) for model training (13,147 pieces) and validation (726 pieces). The train-validation split is performed on song-level (i.e., no training piece and validation piece come from the same song). The training set is augmented by pitch shifting within a range of -5 semitones to 6 semitones.

Audio features are extracted by the hybrid Constant-Q Transform (CQT) using the librosa package [4] with a hop length of 512 and a sample rate of 22050. We use 324 frequency bins starting from C0 with 36 bins per octave.

The model is trained using the Adam optimizer [2] with a scheduled learning rate (1e-4 for 8 epochs, 1e-5 for 4 epochs, 1e-6 for 1 epoch and 1e-7 for 1 epoch). We optimize the sum of the cross-entropy loss for each classification task. If some annotation (e.g., the main melody) is missing for a piece, we simply ignore the loss term for that task.

4. EXPERIMENT

We perform an evaluation on the validation set of Hooktheory using metrics provided by the mir_eval package [5]. We compare our system against the pre-trained model from [3]. The results are shown in Table 3.

Systems	CNN [3]	JXC1	JXC2
Correct	0.6488	0.7507	0.7700
Fifth	0.0386	0.0289	0.0193
Relative	0.1584	0.1033	0.1047
Parallel	0.0124	0.0083	0.0138
Other	0.1419	0.1088	0.0923

Table 3. Evaluation results on the validation set.

[1] https://www.hooktheory.com/

The multi-task model shows better performance in key classification compared to the single-task model on the validation set. Further experiments are required to fully evaluate the proposed systems.

5. REFERENCES

[1] Junyan Jiang, Ke Chen, Wei Li, and Gus Xia. Large-vocabulary chord transcription via chord structure decomposition. In *the 20th International Society for Music Information Retrieval Conference, ISMIR*, 2019.

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[3] Filip Korzeniowski and Gerhard Widmer. Genre-agnostic key classification with convolutional neural networks. *arXiv preprint arXiv:1808.05340*, 2018.

[4] 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.

[5] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel. mir_eval: A transparent implementation of common mir metrics. In *In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*. Citeseer, 2014.