MUSIC CHORD RECOGNITION BASED ON MIDI-TRAINED Deep Feature and BLSTM-CRF HYBRID DECODING

Yiming Wu, Wei Li
School of Computer Science and Technology,
Fudan University, Shanghai, China

QUICK ABSTRACT

Motivation	Proposed solution	Datasets used	Results
✓ Annotating chord sequence requires skilled annotators and considerable time, thus the growth of annotated data amount for supervised learning is relatively slow. In addition, collecting raw audio data for those annotations is often not so easy. ✓ **We want to obtain big training dataset in an alternative way.**	✓ **Automatically "generate" supervising data from MIDI musics** (pairs of **synthesized audio** and **synchronized pitch class activation states**) for feature extractor training. ✓ This method leads to much easier and more robust acoustic feature learning/ extraction. It **does not try to completely replace those annotated data**.	✓ 1100 SMFs from RWC and Lakh MIDI datasets (for training feature extractor). ✓ Isophonics (217 tracks) and RWC-Popular (100 tracks) annotation datasets (for training the BLSTM-CRF decoder).	✓ The feature sequence extracted by MIDI-trained extractor shows better performance for chord classification than raw-CQT and other Chroma feature. ✓ The combined chord recognition system achieves competitive performance in cross-validation experiments and MIREX evaluation.

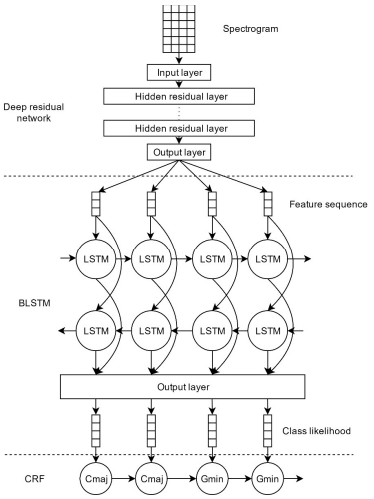

Figure 1. Overview of the system

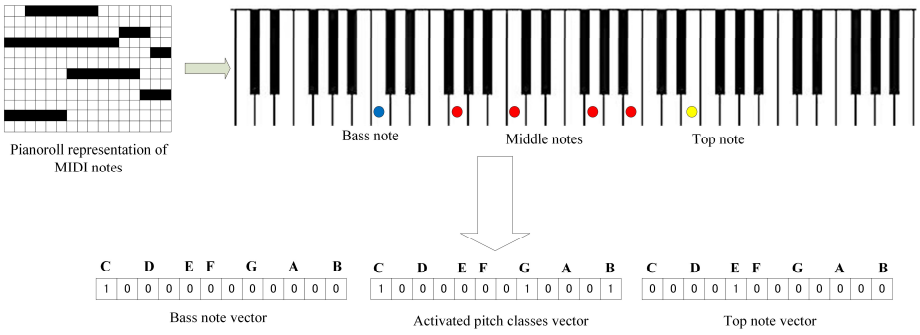

Figure 2. Obtaining target vectors from pianoroll representation

Models implementation/Training details

Feature Extractor Training

- The pianoroll representation of MIDI note data is transformed into a sequence of target vectors.
  - 36-dimention vectors, each 12-dimension indicate the pitch class(es) of bass note, middle notes, and top note of the corresponding frame (Fig.2).
- A deep neural network is trained with audio (synthesized from MIDI data) and corresponding target vectors. It is trained to estimate the bass note, active pitch classes and top notes simultaneously from each spectrum, and output feature vector in a Chroma-like form.
- Network architecture:
  - 5 fully-connected layers with 512 units per layer, activated with tanh function.
  - 36 unit output layer, activated with sigmoid function.
  - The network is trained to minimize mean-squared error between the network output and target vectors.

BLSTM-CRF Sequence Decoder

- A BLSTM-CRF model is trained to transform the extracted feature sequences into 25-class label sequences (major/minor triads +"N").
- The BLSTM is first trained to estimate the class probability of each frame. After BLSTM is trained, a linear-chain CRF is then optimized with the same dataset.
- Given the sequence of class possibility, the CRF infers the optimal label sequence via Viterbi algorithm.

Towards Larger Vocabulary Chord Recognition

- Following the hierarchy in chord vocabulary, chord recognition can be divided into two steps: roughly estimate chords in triad chord level, then determine whether the seventh or inversion is present in each chord, rather than regarding them as new independent chords.
- Concrete steps: given a chord signature (in major or minor triad) and the feature sequence (normalized for each frame) of corresponding time region (segmentation), first the mathematical mean of the feature value along the dimension of its third, fifth, seventh and major-seventh note, and bass feature value of its root, third and fifth note, is calculated.
- Then the true quality and inversion are determined with an explicit thresholding metric.

References

[1] M. Mauch and S. Dixon, Approximate Note Transcription for the Improved Identification of Difficult Chords, International Society for Music Information Retrieval Conference (ISMIR), pp. 135-140, 2010.

[2] F. Korzeniowski and G. Widmer, A Fully Convolutional Deep Auditory Model for Musical Chord Recognition, IEEE International Workshop on Machine Learning for Signal Processing (MLSP), pp. 13-16, 2016.

Experimental results

WAOR under majmin metric

	Isophonics	RWC-Popular
Chordino[1]	76.4%	74.8%
CNN-CRF[2]	83.2%	79.9%
BLSTM-CRF	83.1%	79.2%
DC-BLSTM-CRF	83.0%	79.1%
MIDI-BLSTM-CRF	**84.1%**	**80.8%**

*The same 8-fold cross validation is performed on the systems other than Chordino plugin. CNN-CRF system is a re-implemented version by ourselves.

MIREX evaluation results

	Majmin	Sevenths	SeventhsBass
Isophonics	79.70	69.01	64.11
Billboard2012	78.54	**62.54**	59.57
Billboard2013	72.53	**57.87**	55.14
JayChou	**82.34**	62.02	58.60
RobbieWilliams	80.65	71.23	68.38
RWC-Pop	82.87	**70.02**	66.91
USPOP2002	**81.20**	**68.91**	65.48

*Underlined cell indicates the best score among the submissions this year and bold-font cell indicates the best score ever. The participated system used RWC-Pop and USPOP2002 for model pre-training.

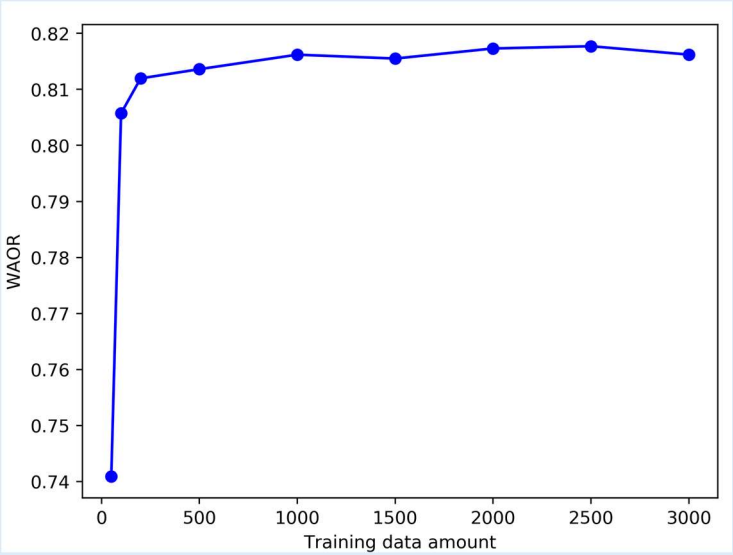

Fig. 3. Relationship between chord recognition accuracy (WAOR) and different size of MIDI dataset.

We are furtherly investigating into Chroma feature extraction with alternative neural network models (like deep CNNs) and even larger MIDI dataset.