MUSICAL POLYPHONY ESTIMATION

Saarish Kareer and Sattwik Basu
Department of Electrical and Computer Engineering

ABSTRACT

Knowing the number of sources in a mixture is useful for many computer audition problems such as polyphonic music transcription, source separation and speech enhancement. In this project we present a deep learning framework for musical polyphony estimation.

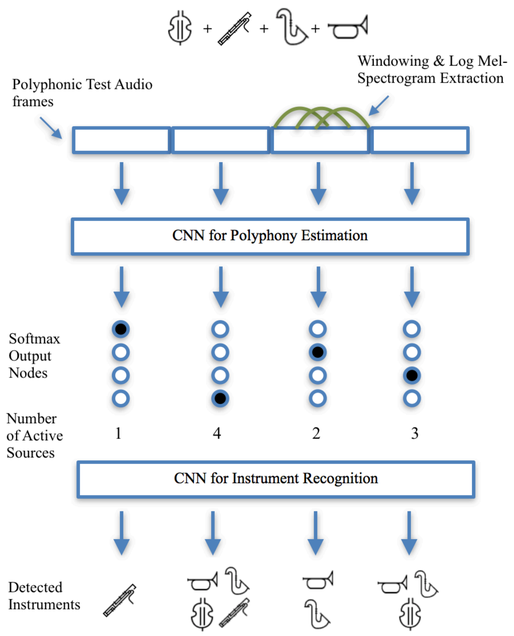
The results are first compared to a multi-pitch estimation algorithm and then used to improve the performance of an instrument classifier. Our method seems to be a promising starting point for further research in unsupervised source counting and separation models for music and speech.

INTRODUCTION

A Convolutional Neural network (CNN) is a class of deep, feed forward neural networks that has successfully been applied to multimedia information retrieval tasks. They are especially suited to work on data that have distinctive local characteristics like images and spectrograms.

Level of Polyphony: In a given piece of music, we define the level of polyphony to be a frame wise count of active notes, pitches or instruments.

Method Outline:



Given a piece of music, the system is able to estimate the level of polyphony and name the constituent instruments active per frame in real time.

IMPLEMENTATION DETAILS

The Bach10 Dataset - Recordings of ten Bach chorales played on a violin, clarinet, saxophone and bassoon

The IOWA Dataset - 24 bit/44.1 KHz chromatic notes from the entire range of numerous instruments.

CNN 1for Polyphony estimation used:

- All combinations of instruments from Bach10 resulting in 40 solo, 60 duets, 40 trios and 10 quartets
- 33 tetrads, triads, dyad and single notes using non-repeating notes of the violin, bassoon, saxophone and clarinet in all combinations

CNN 2 for Instrument Recognition used:

- 4 solo instruments from both datasets giving 40 and 33 solo notes each resulting in 18 minutes of data per instrument

Data pre-processing: The MIR library librosa was used to convert 0.3 second long sub-segments of each audio file into normalized log-mel spectrograms.

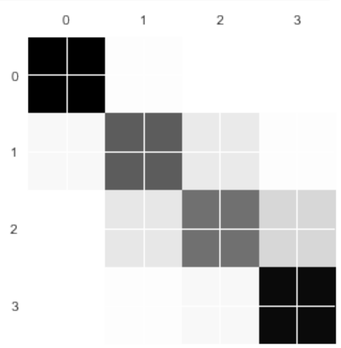
Dimensions	CNN 1 and CNN 2 Architecture
4344 x 128 x 52 3612 x 128 x 52	CNN 1 input CNN 2 input
128 x 52 x 1	Log-mel spectrogram (frame size 1024, hop size 256)
126 x 50 x 32	3 x 3 Conv, 32 filters (ReLU)
63 x 25 x 32	2 x 2 max pool
61x 23 x 64	3 x 3 Conv, 64 filters (ReLU)
30 x 11 x 64	2 x 2 max pool
28 x 9 x 128	3 x 3 Conv, 128 filters (ReLU)
14 x 4 x 128	2 x 2 max pool
7168	Flatten
128	Dense (ReLU)
4	RMSprop optimizer Dense (softmax), categorical-crossentropy Dense (sigmoid), binary-crossentropy

Python Libraries used:
TensorFlow, Keras, Scikit-learn

Training and Testing Results for CNN 1 after 20 epochs:

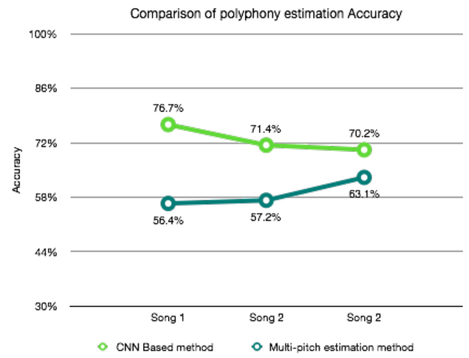
Validation accuracy - 76.81%
Testing accuracy - 81.37%

The confusion matrix is shown here.

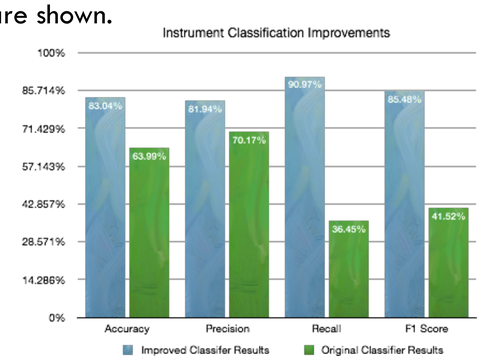


RESULTS

Polyphony estimation of three new audio tracks using CNN 1 and multi-pitch estimation showed the following results.



CNN 2 initially performed poorly on a multi-label instrument recognition task as it was trained on monophonic music. To correct erroneous outputs, the predicted level of polyphony from CNN 1 was used to select the most probable instruments in a mixture. A comparison of the classification results are shown.



DISCUSSIONS

We successfully tackled the problem of polyphony estimation for instruments playing one note in a time frame. Our next step would be to extend the model to inherently polyphonic instruments like the piano. Using this method in concert with probabilistic models can help us realize an unsupervised source counting technique.

REFERENCES

[1] Yoonchang Han, Jaehun Kim, and Kyogu Lee, “Deep convolutional neural networks for predominant instrument recognition in polyphonic music”, *2016*

[2] Zhiyao Duan, Bryan Pardo and Changshui Zhang, “Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions”, 2010