**Homework #3: Automatic Polyphonic Piano Transcription**

The Robotics Program

20195595

Dooyoung Hong

1. **Introduction**

Automatic music transcription (AMT) refers to an automated process that converts musical signals into a piano roll. Polyphonic piano transcription is a specific AMT task for piano music.

Because of the sequential aspect of piano transcription, the recurrent neural network (RNN) module is commonly used for the task. However, the problem of AMT is already notoriously difficult even for humans because of the polyphonic nature of piano music. The transcription problem is made more harder by the fact that there is an enormous number of possible outputs. HW#3 deals with Onsets, Frames for data inputs, and various types of neural networks to perform AMT.

1. **Algorithm description, Experiments, and results**

LSTM (Long Short-Term Memory) is a special type of recurrent neural network (RNN) that aims to solve the forgetting of neural network. A common RNN contains loops that allow the network to better train sequential data. However, it contains problem that arises when the location gap between the sequential data. An LSTM solves this by essentially allowing the neural network to learn what to keep and throw away. Namely, LSTM plays a role in ensuring that the network retains the necessary information for a long time by appropriately adjusting the information to be discarded and the information to be transmitted.

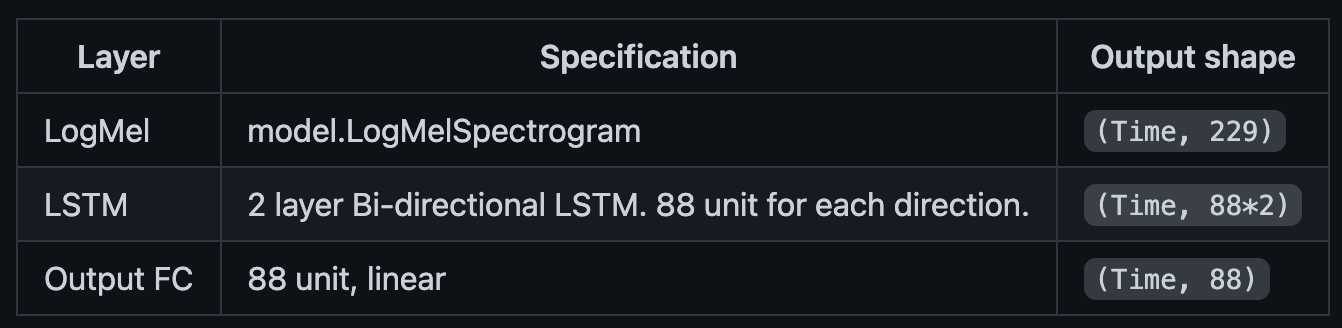


Figure 1. Structure of LSTM-based model.

Question 1 requires implementing LSTM-based model for performing the automatic polyphonic piano transcription. The model uses mel spectrogram normalized through log function and then it uses BiLSTM that runs in both directions, meaning the algorithm is fed the data both front-to-back and back-to-front, which decrease training time of model. The input data replace loss function of Frame and Onset by the fully connected layer after LSTM layer. Figure.2 shows the performance of LSTM-based model. It uses the performance matric of F1-Score.

F1 Score is a ratio average of Precision@K and Recall@K. It is mainly used when data imbalance between classification classes is severe. The machine learning performance cannot be properly represented when if the data classification class is not uniform. So, F1 Score are derived by the harmonic average of precision and recall. The model has the higher value, the better the model.

LSTM-based model has about 0.46 F1 score. I will compare it with other models.

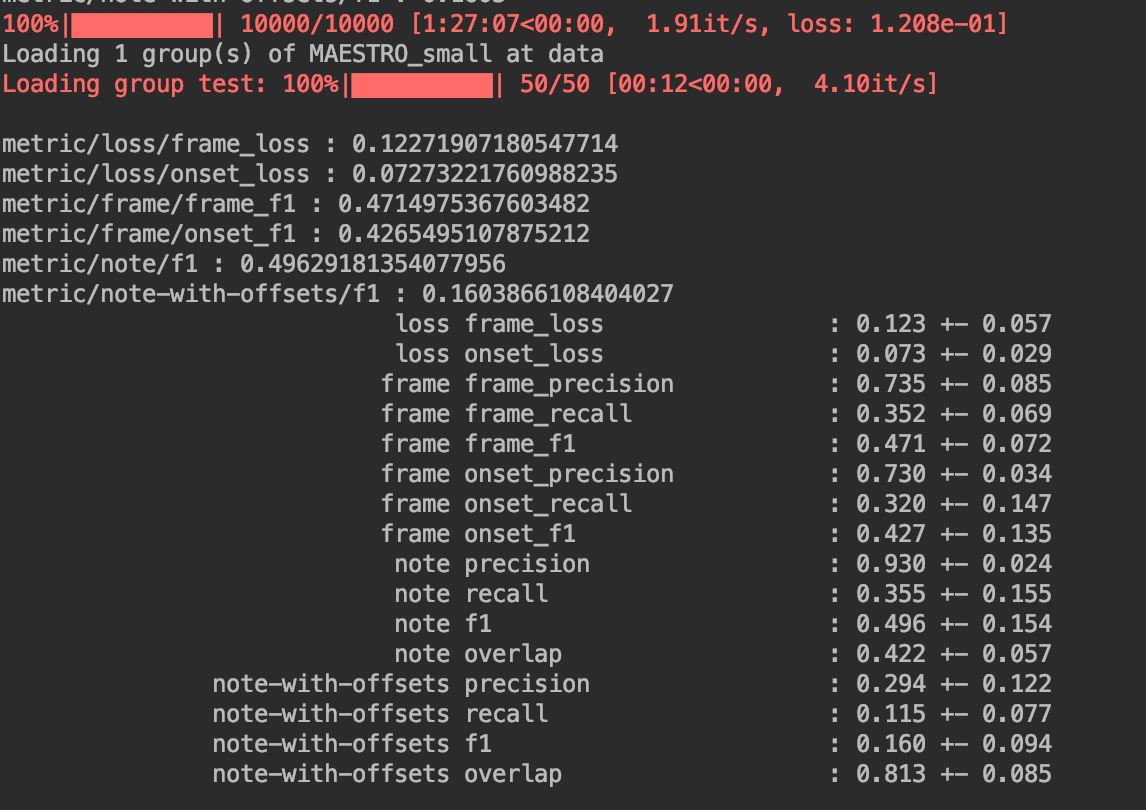


Figure 2. Result of LSTM-based model.

Question 2 requires sequential model with Convolutional neural networks (CNN). CNN is a special type of neural network that uses alternating convolutional and polling layers to gain compress information for recognition, and they are primarily used for image classification.

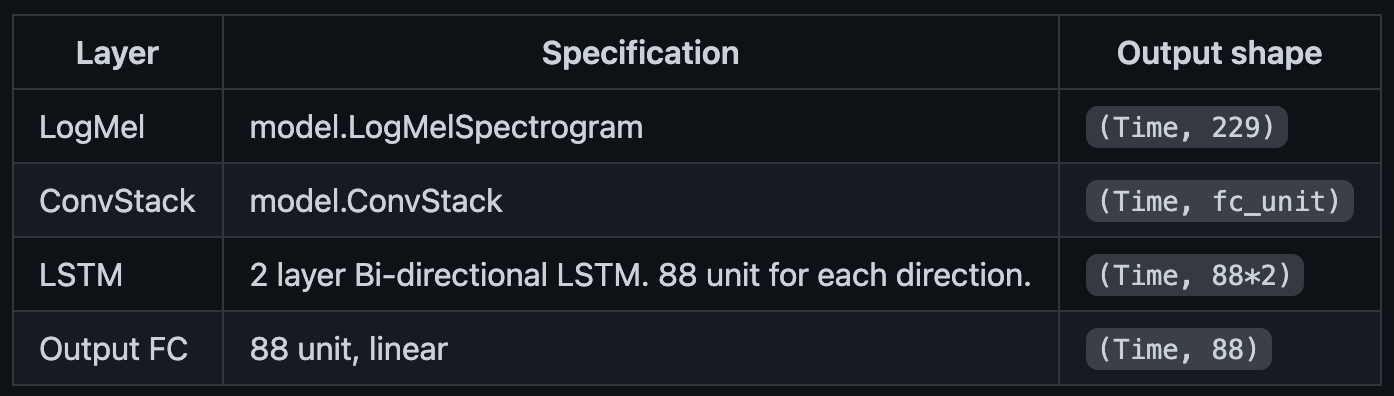


Figure 3. Structure of CRNN model.

CNN commonly takes input data and analyze certain sections of that at a time. It is essentially to make multi-dimensional data like images or spectrograms easier to handle while still preserving the important features. CNN works by alternating convolution and pooling layers.

Convolution layers analyze little parts of whole data at a time and combine the feature of data.

The pooling layers aim to further reduce the spatial size of the data by only keeping dominant feature.

The model also uses mel spectrogram normalized through log function and BiLSTM. Instead, it uses convolutional layer before LSTM layer. Convolutional filters have a positive effect on the model by consolidating information between frames in the spectrogram. Figure.4 shows the performance of CRNN model. It shows a significant performance improvement compared with the LSTM-based model.

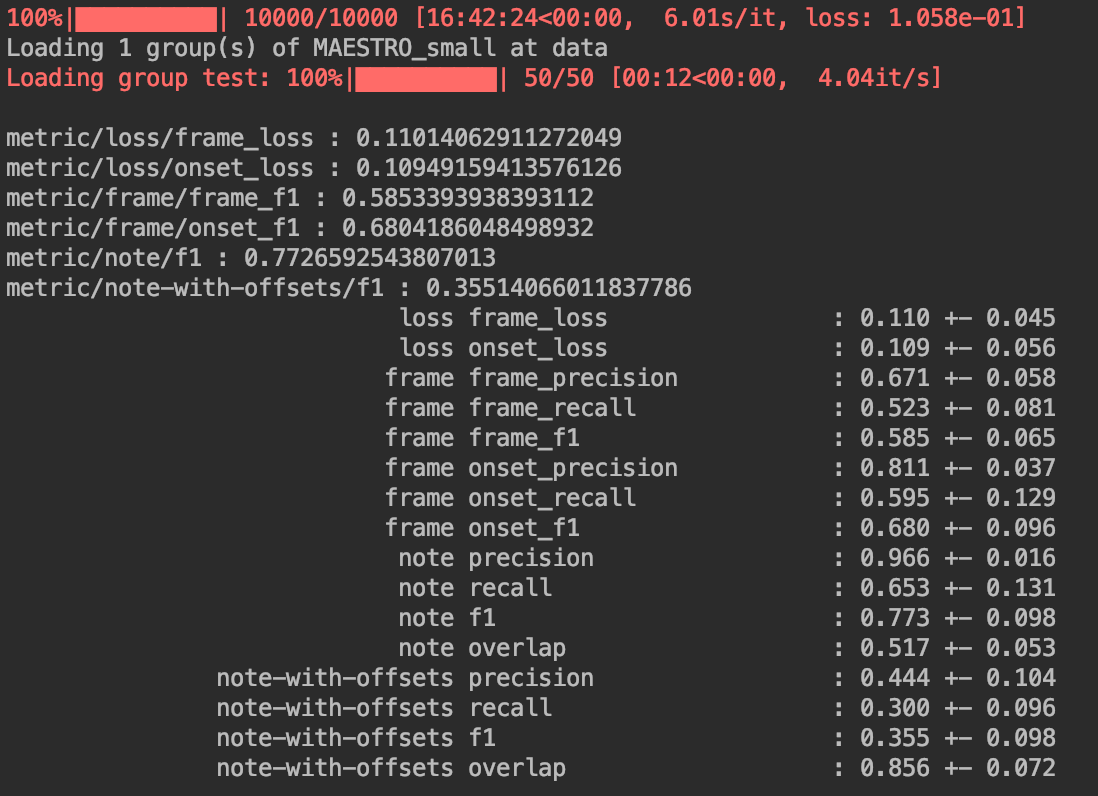


Figure 4. Result of CRNN model.

Question 3 is implementing Onsets-and-Frames model. This model uses an inter-connection between the onsets and frames. This is a method of using the logit information of Onsets before going through the sigmoid function as basic information to derive the frame logits.

Separating the task of transcription into note onset and framewise activation prioritizes more important musical moments, ensuring a musically useful transcription. This model directly links the derived relationship of onset and frame to give certainty to the model’s putative relationship.

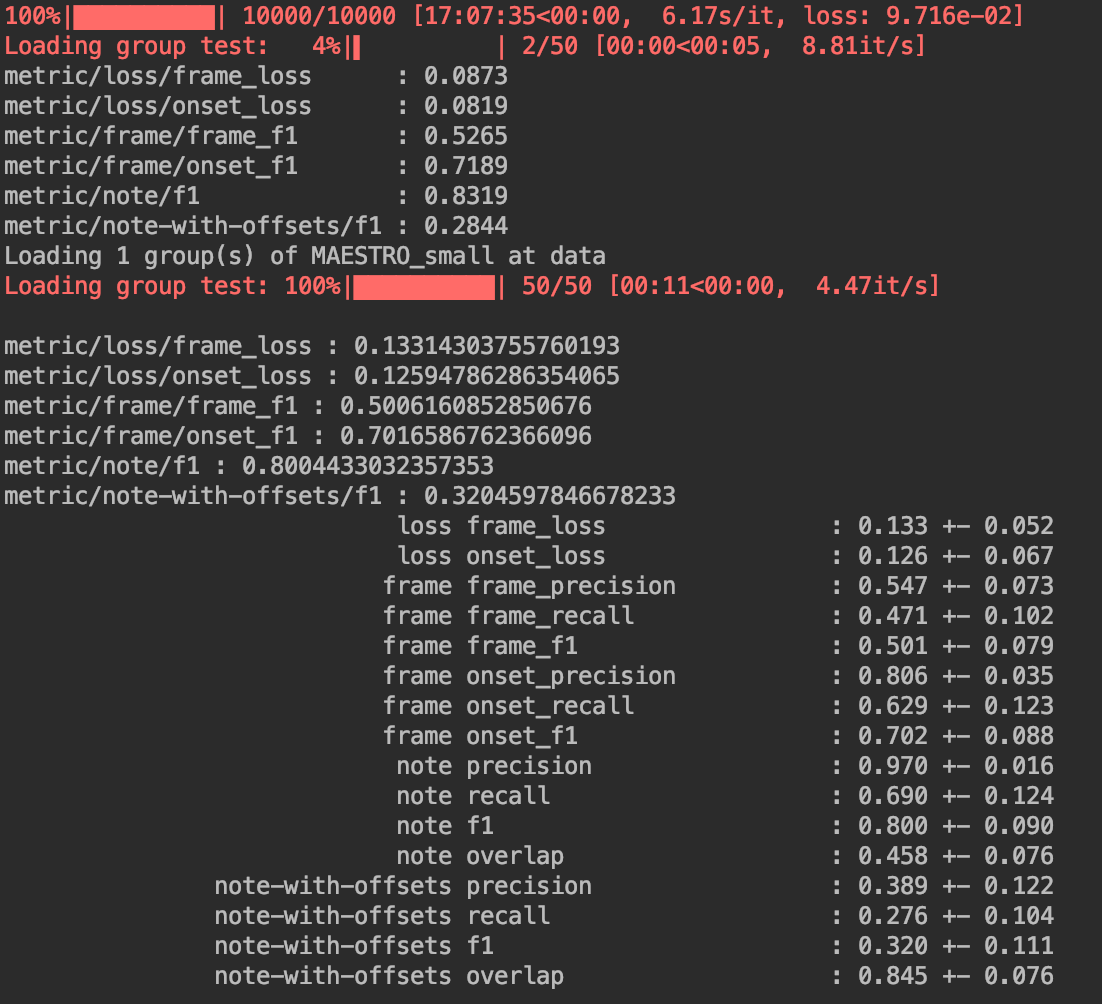


Figure 5. Result of Onsets-and-Frames model.

Figure 5 shows the performance of Onsets-and-Frames model. This model does show some performance improvement over the CRNN model although there is no such dramatic performance improvement between the LSTM-based model and CRNN model.

Question 4 requires the visualization of model prediction results as the piano roll format. Figure.6 is the sample dataset of MAESTRO (it presented dataset.py). It shows audio sample and its frames and onsets. Finally, figure.7 shows the model prediction results of AMT.

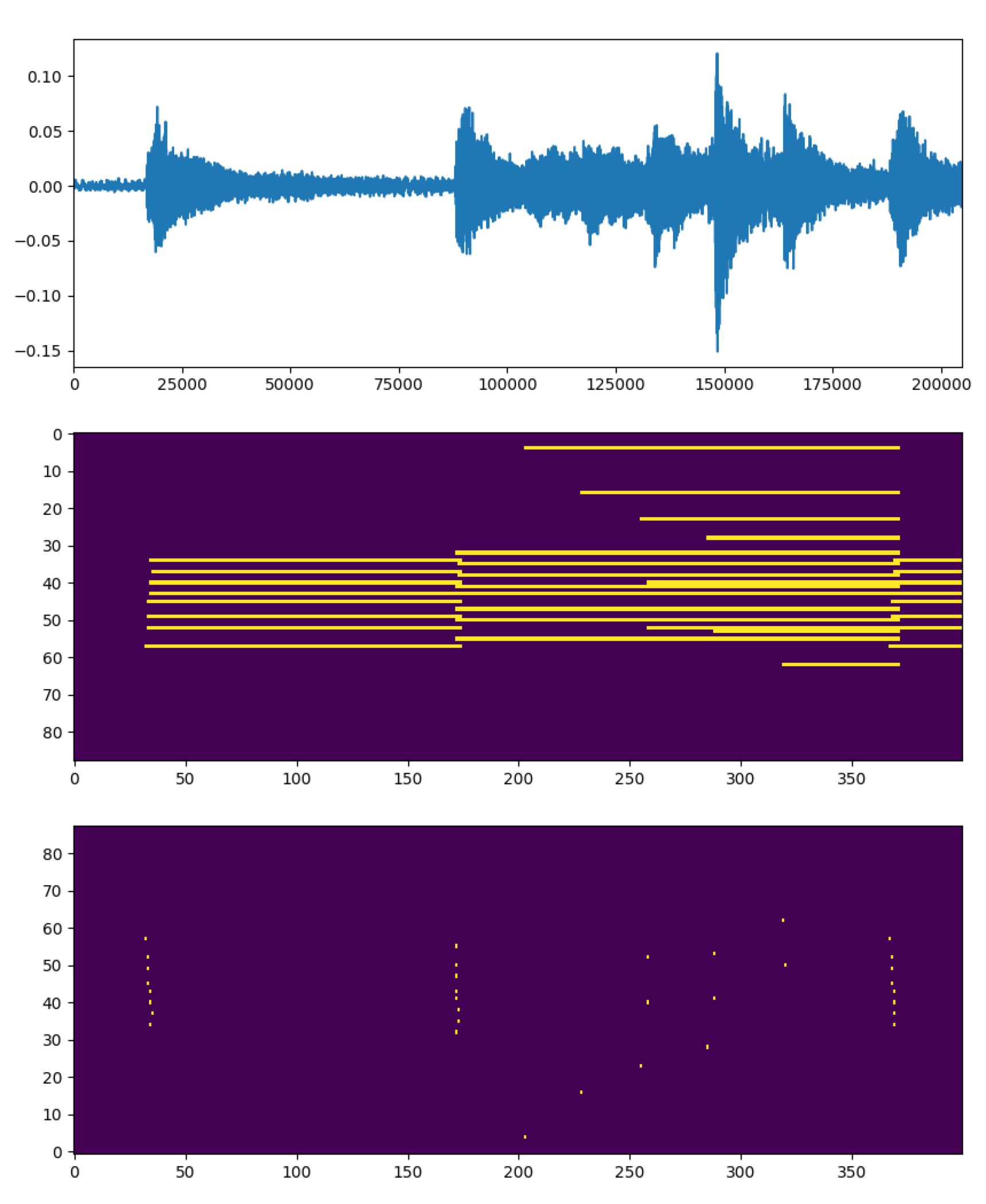


Figure 6. Sample audio and its frames and onsets.

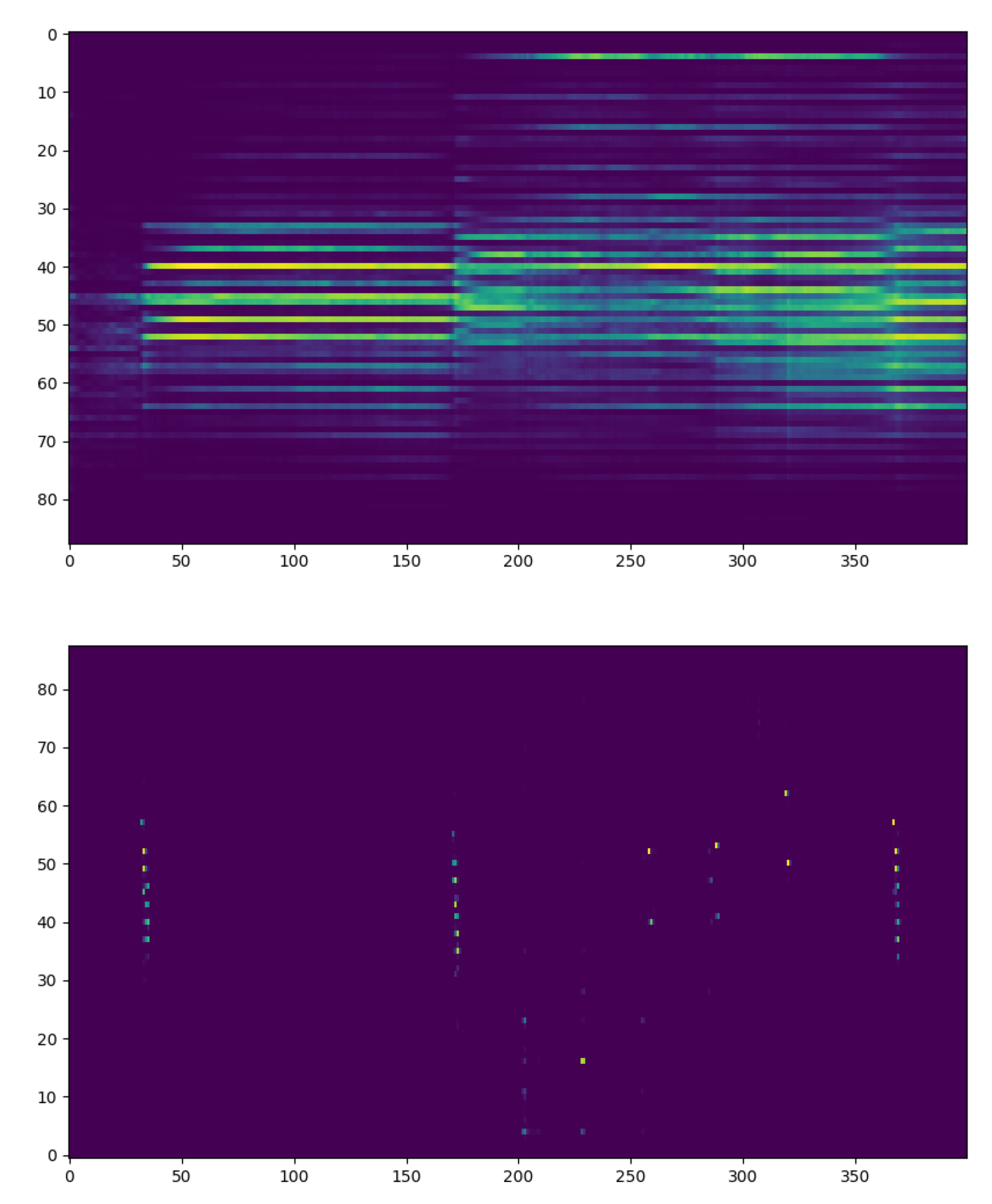
****

Figure 7. Prediction of frames and onsets by ONF model.

Both frame and onset show visibility comparable to the naked eye compared with the basic data. However, in the frame the noise that occurs between predictions of each tone is clearly recognized. It through that approximation of the model affects the accuracy when the model predicts the variability of the input signal for the computer evaluate all sounds.

1. **Conclusion**

Homework #3 performs the automatic music transcription using machine learning techniques.

It contains convolutional filter and sequential network for analyzing the overlap of harmonics in the acoustic signal. AMT is already notoriously difficult even for humans due to the polyphonic sound of piano. However, Onsets and Frames can convert piano recordings into a MIDI sequence. It uses a system similar to speech recognition in that it uses acoustic models in conjunction with a music language model and show the higher classification performance.