Group 1 Final Project

Jingjing Xu
Madison Turano

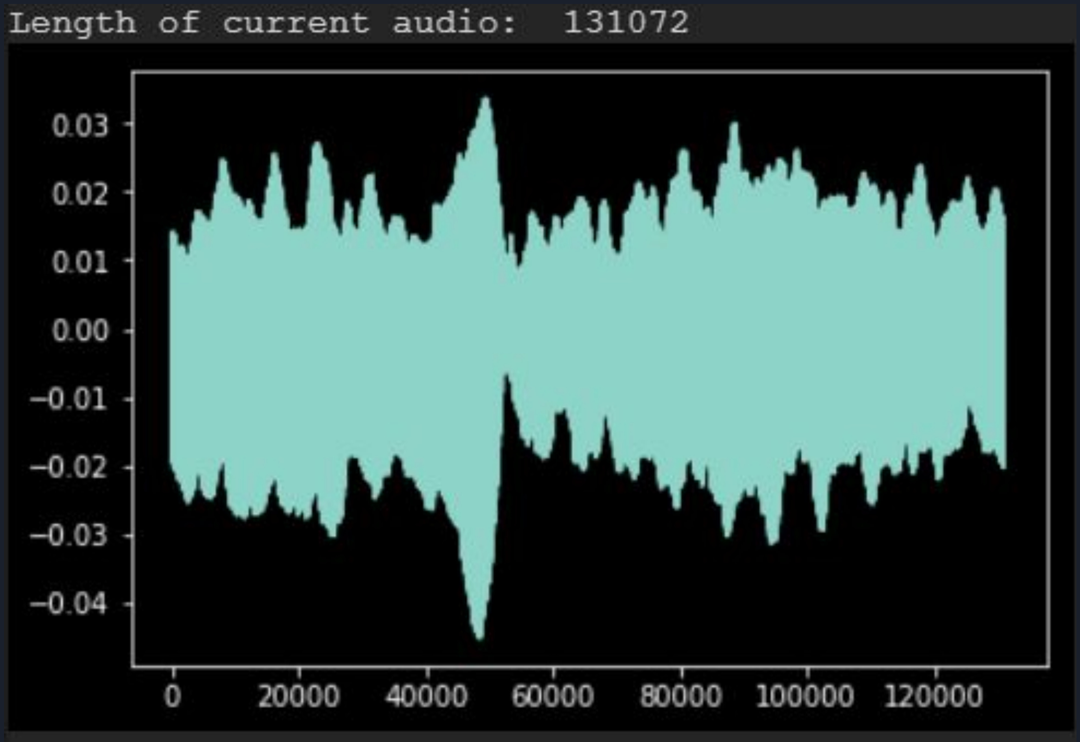
Project Overview

- Use time-series data and LSTM model
- Use audio data from Medley-solos-DB
- Tag audio data using LSTM

Data Overview

- Classes:
  - Clarinet
  - Distorted electric guitar
  - Female singer
  - Flute
  - Piano
  - Tenor saxophone
  - Trumpet
  - Violin
- Number of inputs: 131,072
- Audio file names are based on meta-data
- Created function for file names and file loading

Example of Audio File



- Plots time (ms) vs amplitude

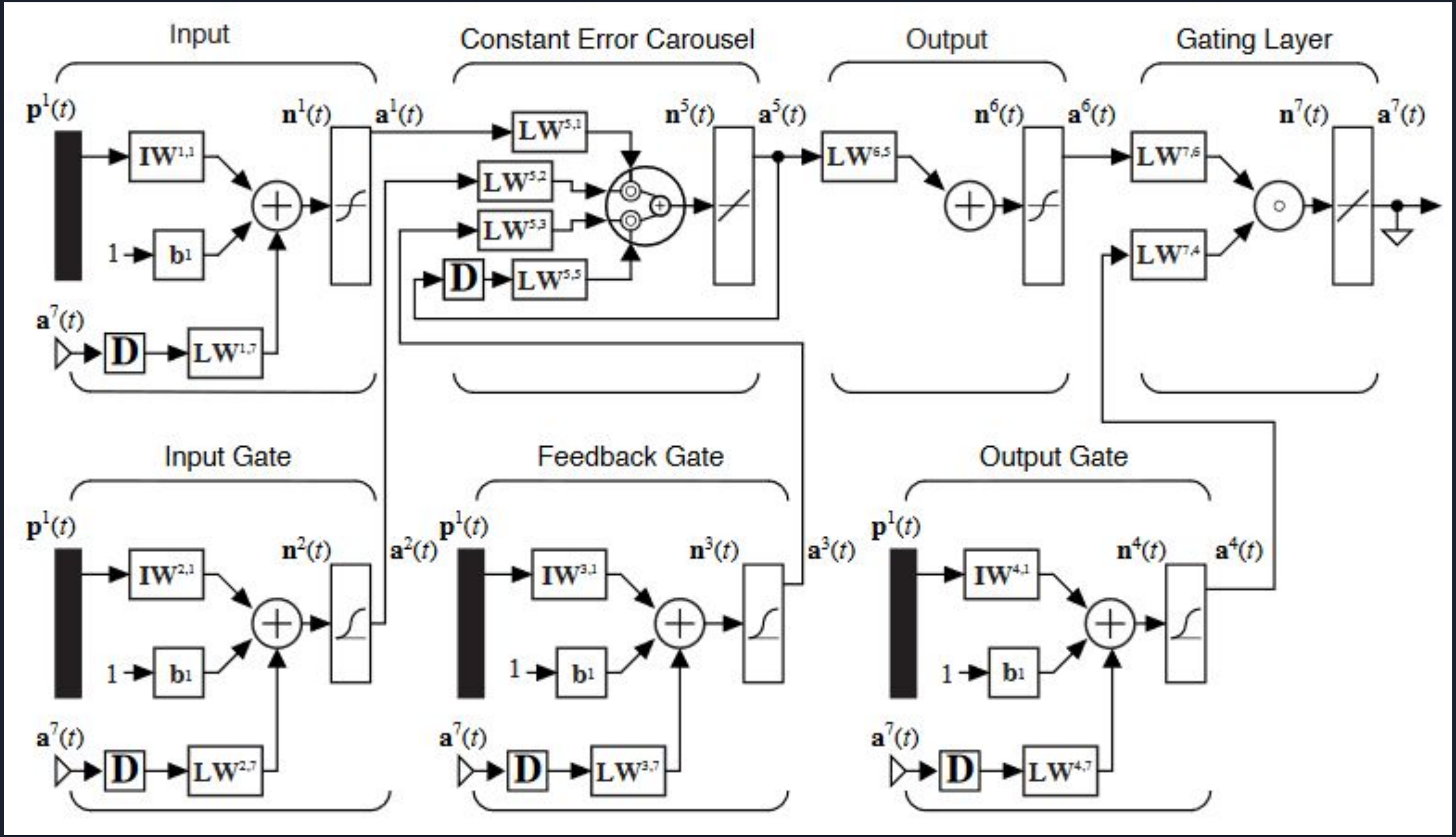
File Functions

```
#Load audio file linked to the uuid
def full_name(file):
    correspding_row = Medley.loc[Medley['uuid4'] == file].iloc[0]
    subset = str(correspding_row.loc['subset'])
    instrument_id = str(correspding_row.loc['instrument_id'])
    parts = ['Medley-solos-DB_', str(subset), '-', str(instrument_id), '_', file, '.wav.wav']
    s = ''
    file_name = s.join(parts)
    return file_name

def load_file(file):
    file_name = full_name(file)
    path = '/home/ubuntu/Final-Project-Group1/Medley-solos-DB/'
    parts = [path, file_name]
    s = ''
    link = s.join(parts)
    return link
```

LSTM Overview

- Recurrent neural network that “remembers” previous inputs
- Dynamic network:
  - Contains delays
  - Works on sequences
- Includes input, feedback, and output gates


Input
Constant Error Carousel
Output
Gating Layer
Input Gate
Feedback Gate
Output Gate
$p^1(t)$
$IW^{1,1}$
b^1
$n^1(t)$
$a^1(t)$
$a^7(t)$
D
$LW^{1,7}$
$LW^{5,1}$
$LW^{5,2}$
$LW^{5,3}$
$LW^{5,5}$
$n^5(t)$
$a^5(t)$
$LW^{6,5}$
$n^6(t)$
$a^6(t)$
$LW^{7,6}$
$LW^{7,4}$
$n^7(t)$
$IW^{2,1}$
$n^2(t)$
$a^2(t)$
$LW^{2,7}$
$IW^{3,1}$
$n^3(t)$
$a^3(t)$
$LW^{3,7}$
$IW^{4,1}$
$n^4(t)$
$a^4(t)$
$LW^{4,7}$

Data Extraction Comparison

<u>Time Domain</u>

- Audio Data

<u>Advantage:</u>

- Fast

<u>Disadvantage:</u>

- Low accuracy

<u>Frequency Domain</u>

- MFCC
- Spectral Centroid
- Spectral Contrast

<u>Advantage:</u>

- High accuracy

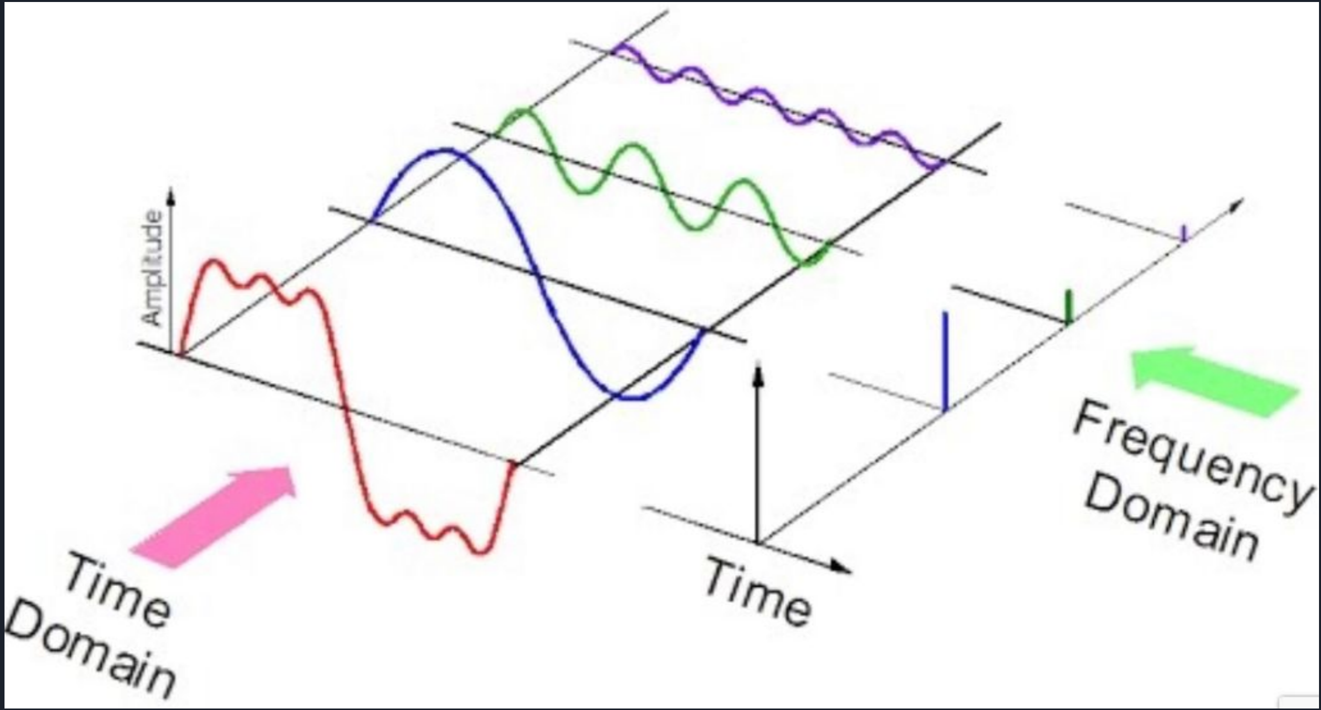
<u>Disadvantage:</u>

- Slow

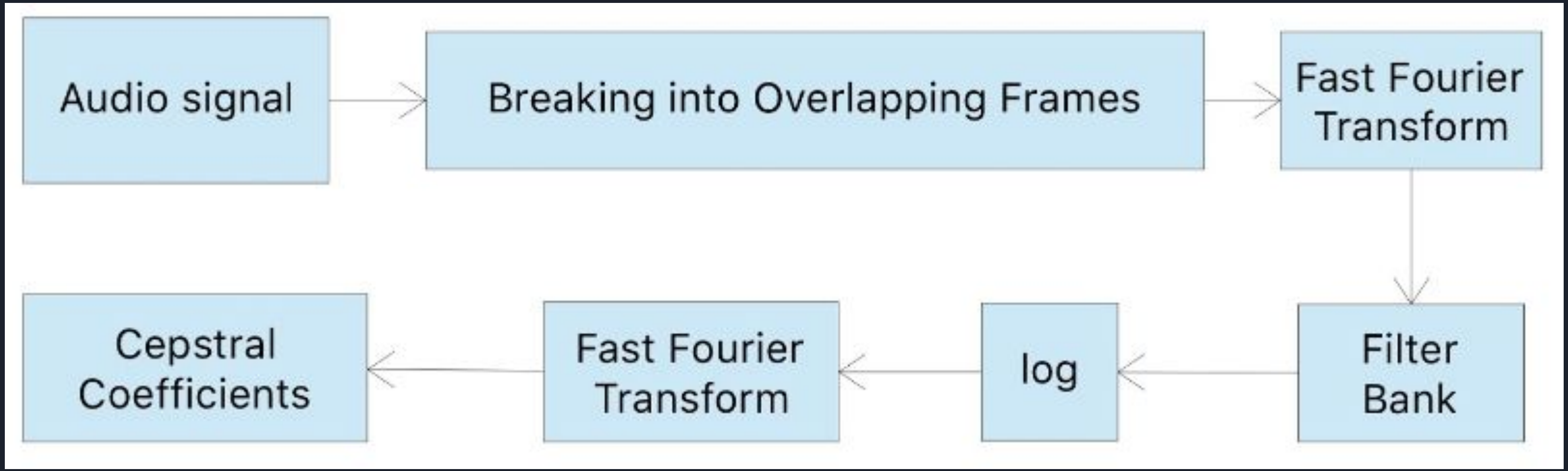
Feature Extraction

- Captures more differences between classes then time series
- Uses Fourier Transform
  - Extracts cosine and sine waves with different features
  - Expresses in frequency domain
- Features used:
  - Spectral centroid: “Brightness” of sound
  - Spectral contrast: Level differences between peaks and valleys
  - MFCCs: Mel frequency cepstral coefficients

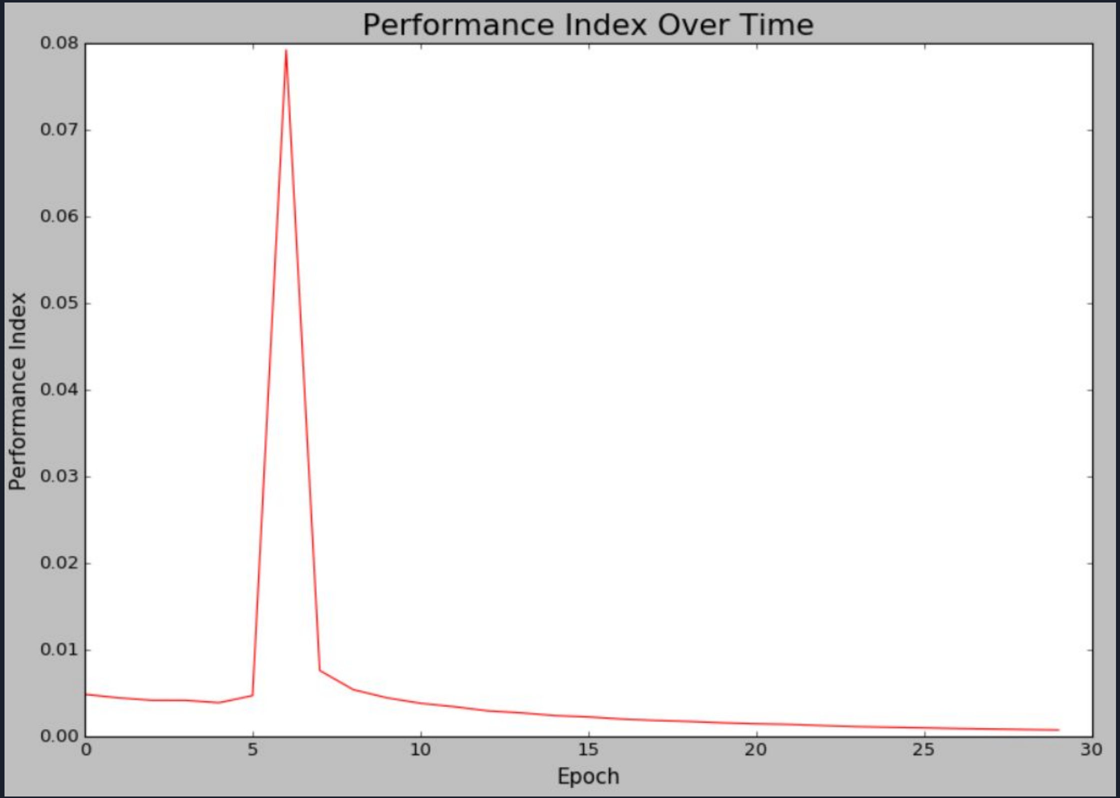
Fourier Transformation



Archive MFCCs



Results: Time Domain



Results

Frequency domain:

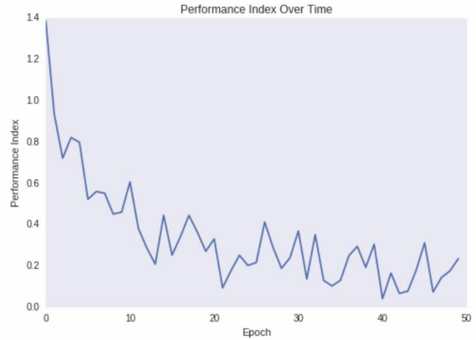


Figure 5-1. Epoch 50, Batch Size 100, Learning Rate 0.0001, Dropout 0.1, Adam, LSTM Layer 2

```
Accuracy of the network on the 3494 validation audio clips: 78 %
Accuracy of clarinet : 48 %
Accuracy of distorted electric guitar : 83 %
Accuracy of female singer : 75 %
Accuracy of flute : 30 %
Accuracy of piano : 89 %
Accuracy of tenor saxophone : 24 %
Accuracy of trumpet : 84 %
Accuracy of violin : 86 %
```

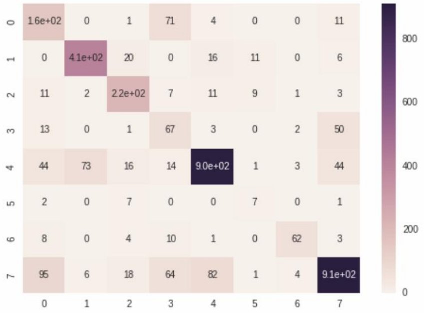


Figure 5-2. Epoch 50, Batch Size 100, Learning Rate 0.0001, Dropout 0.1, Adam, LSTM Layer 2

Results

Batch size:

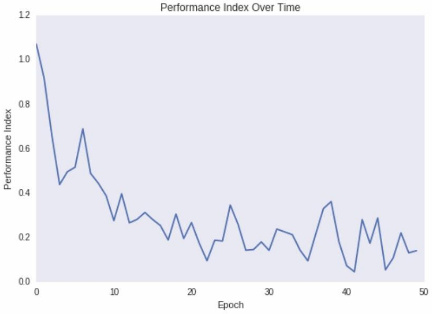


Figure 6-1. Epoch 50, Batch Size 50, Learning Rate 0.0001, Dropout 0.1, Adam, LSTM Layer 2

```
Accuracy of the network on the 3494 validation audio clips: 80 %
Accuracy of clarinet : 52 %
Accuracy of distorted electric guitar : 84 %
Accuracy of female singer : 73 %
Accuracy of flute : 33 %
Accuracy of piano : 88 %
Accuracy of tenor saxophone : 24 %
Accuracy of trumpet : 84 %
Accuracy of violin : 90 %
```

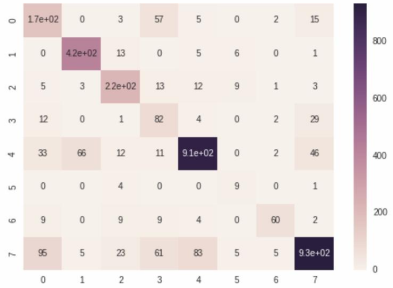


Figure 6-2. Epoch 50, Batch Size 50, Learning Rate 0.0001, Dropout 0.1, Adam, LSTM Layer 2

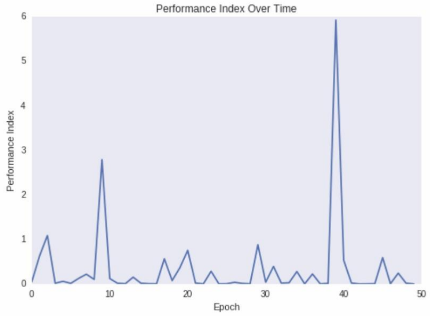


Figure 7-1. Epoch 50, Batch Size 20, Learning Rate 0.0001, Dropout 0.1, Adam, LSTM Layer 2

```
Accuracy of the network on the 3494 validation audio clips: 78 %
Accuracy of clarinet : 44 %
Accuracy of distorted electric guitar : 71 %
Accuracy of female singer : 75 %
Accuracy of flute : 29 %
Accuracy of piano : 89 %
Accuracy of tenor saxophone : 17 %
Accuracy of trumpet : 77 %
Accuracy of violin : 92 %
```

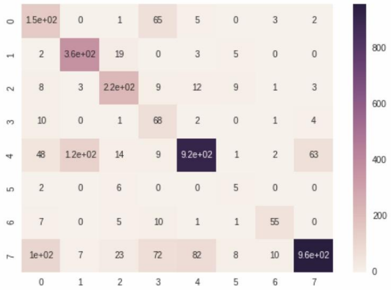


Figure 7-2. Epoch 50, Batch Size 20, Learning Rate 0.0001, Dropout 0.1, Adam, LSTM Layer 2

Results

Learning rate:

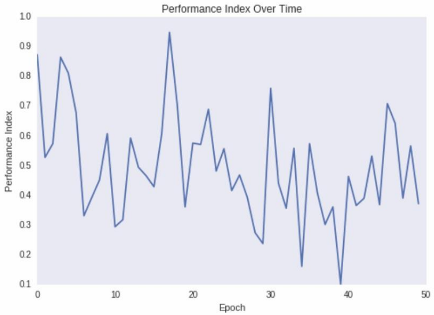


Figure 8-1. Epoch 50, Batch Size 100, Learning Rate 0.001, Dropout 0.1, Adam, LSTM Layer 2

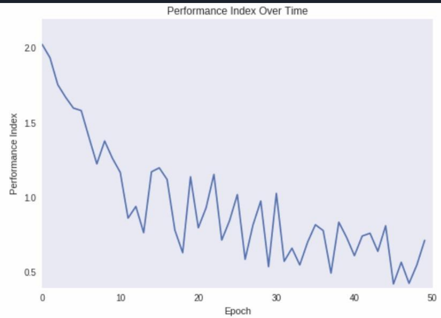


Figure 9-2. Epoch 50, Batch Size 100, Learning Rate 0.00001, Dropout 0.1, Adam, LSTM Layer 2

```
Accuracy of the network on the 3494 validation audio clips: 75 %
Accuracy of clarinet : 35 %
Accuracy of distorted electric guitar : 79 %
Accuracy of female singer : 77 %
Accuracy of flute : 14 %
Accuracy of piano : 86 %
Accuracy of tenor saxophone : 17 %
Accuracy of trumpet : 69 %
Accuracy of violin : 93 %
```

```
Accuracy of the network on the 3494 validation audio clips: 67 %
Accuracy of clarinet : 11 %
Accuracy of distorted electric guitar : 58 %
Accuracy of female singer : 45 %
Accuracy of flute :  0 %
Accuracy of piano : 88 %
Accuracy of tenor saxophone :  0 %
Accuracy of trumpet : 13 %
Accuracy of violin : 95 %
```

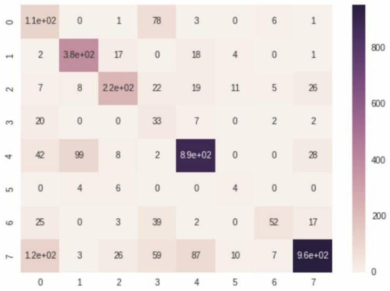


Figure 8-2. Epoch 50, Batch Size 100, Learning Rate 0.001, Dropout 0.1, Adam, LSTM Layer 2

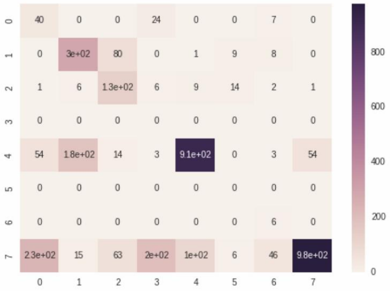


Figure 9-2. Epoch 50, Batch Size 100, Learning Rate 0.00001, Dropout 0.1, Adam, LSTM Layer 2

Results

Dropout:

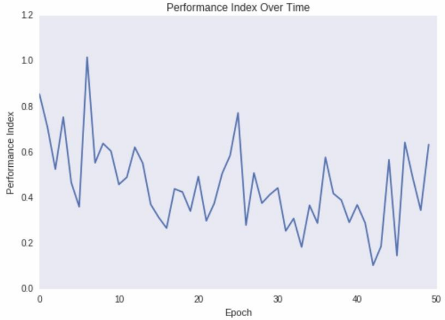


Figure 10-1. Epoch 50, Batch Size 100, Learning Rate 0.0001, Dropout 0.05, Adam, LSTM Layer 2

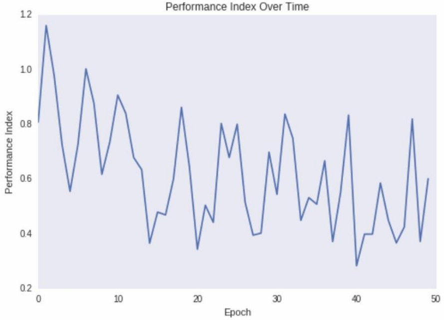


Figure 11-1. Epoch 50, Batch Size 100, Learning Rate 0.0001, Dropout 0.5, Adam, LSTM Layer 2

```
Accuracy of the network on the 3494 validation audio clips: 75 %
Accuracy of clarinet : 39 %
Accuracy of distorted electric guitar : 76 %
Accuracy of female singer : 61 %
Accuracy of flute : 27 %
Accuracy of piano : 85 %
Accuracy of tenor saxophone : 17 %
Accuracy of trumpet : 76 %
Accuracy of violin : 94 %
```

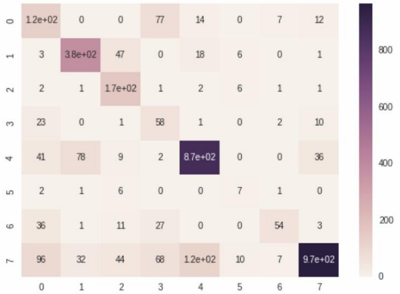


Figure 10-2. Epoch 50, Batch Size 100, Learning Rate 0.0001, Dropout 0.05, Adam, LSTM Layer 2

```
Accuracy of the network on the 3494 validation audio clips: 70 %
Accuracy of clarinet : 23 %
Accuracy of distorted electric guitar : 68 %
Accuracy of female singer : 50 %
Accuracy of flute : 15 %
Accuracy of piano : 86 %
Accuracy of tenor saxophone :  3 %
Accuracy of trumpet : 59 %
Accuracy of violin : 88 %
```

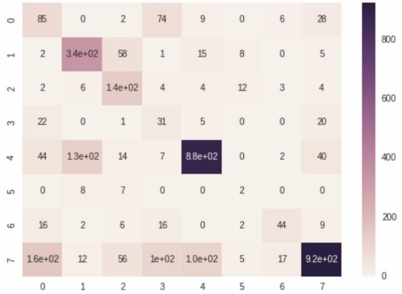


Figure 11-2. Epoch 50, Batch Size 100, Learning Rate 0.0001, Dropout 0.5, Adam, LSTM Layer 2

Results

Optimizer:

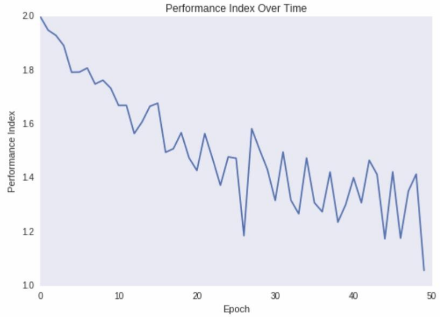


Figure 12-1. Epoch 50, Batch Size 100, Learning Rate 0.0001, Dropout 0.1, SGD, LSTM Layer 2

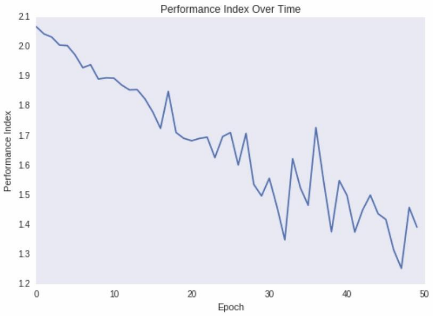


Figure 13-1. Epoch 50, Batch Size 100, Learning Rate 0.0001, Dropout 0.1, Adadelta, LSTM Layer 2

```
Accuracy of the network on the 3494 validation audio clips: 52 %
Accuracy of clarinet :  0 %
Accuracy of distorted electric guitar :  0 %
Accuracy of female singer :  0 %
Accuracy of flute :  0 %
Accuracy of piano : 87 %
Accuracy of tenor saxophone :  0 %
Accuracy of trumpet :  0 %
Accuracy of violin : 90 %
```

```
Accuracy of the network on the 3494 validation audio clips: 49 %
Accuracy of clarinet :  0 %
Accuracy of distorted electric guitar :  0 %
Accuracy of female singer :  0 %
Accuracy of flute :  0 %
Accuracy of piano : 93 %
Accuracy of tenor saxophone :  0 %
Accuracy of trumpet :  0 %
Accuracy of violin : 75 %
```

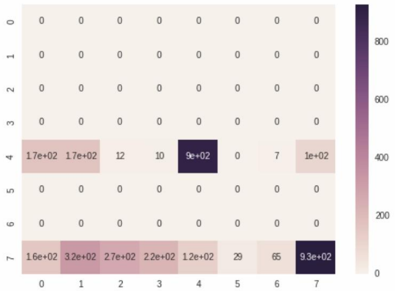


Figure 12-2. Epoch 50, Batch Size 100, Learning Rate 0.0001, Dropout 0.1, SGD, LSTM Layer 2

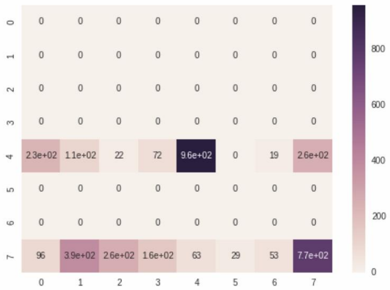


Figure 13-2. Epoch 50, Batch Size 100, Learning Rate 0.0001, Dropout 0.1, Adadelta, LSTM Layer 2

Results

Extra LSTM layer:

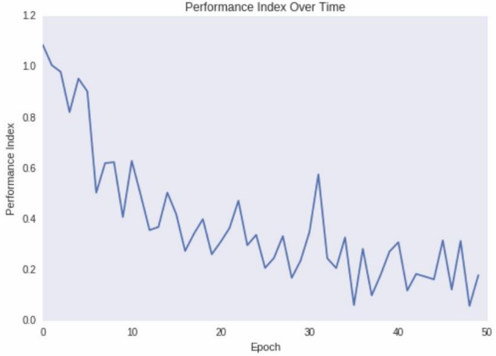


Figure 14-1. Epoch 50, Batch Size 100, Learning Rate 0.0001, Dropout 0.1, Adam, LSTM Layer 2+1

```
Accuracy of the network on the 3494 validation audio clips: 79 %
Accuracy of clarinet : 45 %
Accuracy of distorted electric guitar : 92 %
Accuracy of female singer : 74 %
Accuracy of flute : 30 %
Accuracy of piano : 88 %
Accuracy of tenor saxophone : 55 %
Accuracy of trumpet : 79 %
Accuracy of violin : 91 %
```

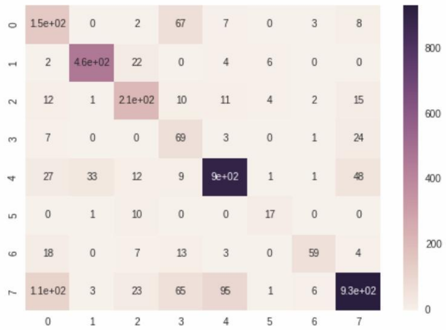


Figure 14-2. Epoch 50, Batch Size 100, Learning Rate 0.0001, Dropout 0.1, Adam, LSTM Layer 2+1

Results on Test Set

```
Accuracy of the network on the 3494 validation audio clips: 53 %
Accuracy of clarinet : 29 %
Accuracy of distorted electric guitar : 83 %
Accuracy of female singer : 69 %
Accuracy of flute :  6 %
Accuracy of piano : 98 %
Accuracy of tenor saxophone :  6 %
Accuracy of trumpet : 50 %
Accuracy of violin : 56 %
```

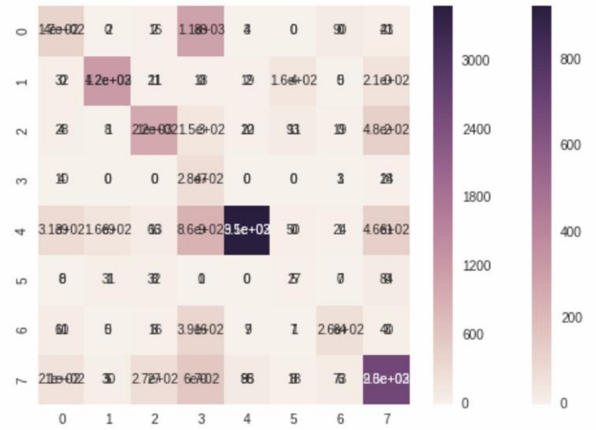


Figure 15. Epoch 50, Batch Size 50, Learning Rate 0.0001, Dropout 0.1, Adam, LSTM Layer 2; Test Set

The Issue

The network seems to confused between these pairs:

- Flute with Clarinet;
- Saxophone, Distorted Electric Guitar and Female Singer.

Possible Improvement

- Use larger proportion of training set;
- Training set more balancedly distributed among classes.
- Extract extra kind of useful data;
- With LSTMs, can possible predict music (another project).

Reference

1. YouTube.com. *Time domain and frequency domain*. Retrieved from https://m.youtube.com/watch?v=tMPDe7z7ERE;
2. Nair, Pratheeksha. (2018). *The dummy's guide to MFCC*. Retrieved from https://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd;
3. Lostanlen, Vincent; Cella, Carmine-Emanuele; Bittner, Rachel; Essid, Slim., Medley solos-DB: A Cross-Collection Dataset for Musical Instrument Recognition. Retrieved from https://zenodo.org/record/;
4. Nogueira W., Rode T., Büchner A. (2016) Optimization of a Spectral Contrast Enhancement Algorithm for Cochlear Implants Based on a Vowel Identification Model. In: van Dijk P., Başkent D., Gaudrain E., de Kleine E., Wagner A., Lanting C. (eds) Physiology, Psychoacoustics and Cognition in Normal and Impaired Hearing. Advances in Experimental Medicine and Biology, vol 894. Springer, Cham.