**CZ4042 Group Project Report**

Audio Emotion Recognition

Li Bingzi, Zeng Yanxi, Zhang Yuehan

# *Abstract-*The aim of this project is to build a deep learning model that can classify the audios into several emotions.

**Keywords:** CNN, Classification, Audio Emotion Recognition

# Introduction

The idea of this project…

# Literature view

Start…

# Methodology

## Neural Network and Keras

Keras is a well-known deep learning API in Python. It wraps frequently used methods and algorithms in deep learning to enable easy model building and access to the history of model training. It also contains some data pre-processing API, but not as many as sklearn. Therefore, sklearn will be used as well in this project alongside Keras.

## Data Pre-processing

In this project, the data is audio files in .wav format. The input of the CNN network is images. Thus a pre-processing method is needed.

**MFCC**

In this project, two audio pre-processing techniques are applied and compared. One is MFCC, which is derived from a type of cepstral representation of the audio clip. It is a time sequence with MFCC bands on the y-axis. For this dataset the bands are defined to 30.

Chart

Description automatically generated

Chart

Description automatically generated

Fig 1: Audio wave transformed into MFCC ( visualized by color bar)

**Melspectrogram**

Another is log-melspectrogram. Mel-spectrogram is less compressed than MFCC. MFCC representation is from taking the logs and compute DCT from mel-spectrogram. In this project we’ll take both method into experiment and compare the accuracies.

A picture containing monitor, computer, clock

Description automatically generated

Fig 2: The same audio wave transformed into melspec (without log transform)

## 2D CNN

When processing dataset like images, ConvNet is able to capture the Spatial and Temporal dependencies in an image through relevant filters.

In this experiment, a simple and commonly used architecture of ConvNet is designed: Conv blocks followed by fully connected layers.

The standard Conv block includes: 2DConv layer, batch normalization, MaxPooling layer and dropout.

In this experiment, there are 32 channels for each layer. Dropout is applied to the layers.

A picture containing diagram

Description automatically generated

Fig 3: Convolutional Network [from lecture slides]

## CNN+LSTM

# experiments and Results

## Data Exploration

**RAVDESS**

The recording of RAVDESS dataset is very clear. The upside of this dataset is that it includes both genders.

The downside is that all of the speakers vocalize the content in North American accent. And there’re only two sentences uttered. These might cause the model to fit in well in this dataset but performs worse on another dataset.

The dataset includes:

|  |  |
| --- | --- |
| female\_neutral | 144 |
| male\_neutral | 144 |
| female\_sad | 96 |
| male\_sad | 96 |
| female\_happy | 96 |
| male\_happy | 96 |
| female\_disgust | 96 |
| male\_disgust | 96 |
| female\_angry | 96 |
| male\_angry | 96 |
| female\_fear | 96 |
| male\_fear | 96 |

Table 1: emotions in the RAVDESS dataset

Comparing sentences:

As shown from the figure below, different sentences uttered have different waveplots. However, the two saveplots have similar feature, which is huge spike in the sound waves, which might be a prominent feature for angry emotion. Different contents uttered does not affect the feature.

Chart

Description automatically generated

Fig3: female, intense angry, speaking "Kids"

Chart

Description automatically generated

Fig4: female, intense angry, speaking "Dogs"

Comparing emotions:

Comparing figure 4 and figure 5, same gender speaking same content in angry and neutral emotion. Clear feature can be distinguished. For angry emotion, there are spikes and the volume is high. Neutral voice has low volume and consistent pace.

Chart

Description automatically generated

Fig 5: female, neutral, speaking "Dogs"

Comparing genders:

Comparing figure 4 and figure 6, female and male speaking the same sentence with same emotion, similar patterns can be seen in the sharp spikes.

The feature which needs attention is that male usually has much lower pitch than female. And more interesting that female tends to express the emotion more strongly.

To prevent genders from affecting the result of emotions, emotions and genders are both in the output label in this project. Therefore, there are 14 class labels, as shown in table 1.

Chart

Description automatically generated

Fig 6: male, intense angry, speaking "Dogs"

Text, whiteboard

Description automatically generated

Fig 7: male vs female, intense angry, speaking "Dogs" (taking average across 30 mfcc bands)

**CREMA-D**

CREMA-D dataset is much larger than RAVDESS.

The downside of this dataset is that the audios are taken from movie. They are not the best of quality.

The upside is that this dataset covers various accents, genders, age. In this project, the two datasets are combined to provide a gender age and accent invariant dataset.

In the baseline 2D CNN build in the next part, it will be shown that adding CREMA-D dataset greatly reduce the model testing accuracy. However, this lower-quality data with high variety shows how well the model will perform in real life.

Comparing with RAVDESS dataset:

Comparing figure 3 and figure 8, from different sentences uttered in these two dataset, the sharp spikes in the angry emotion are still present. This is the base for the classification model to be possible.

Chart, histogram

Description automatically generated

Fig 8: female, intense angry

|  |  |
| --- | --- |
| female\_neutral | 144 |
| male\_neutral | 144 |
| female\_sad | 96 |
| male\_sad | 96 |
| female\_happy | 96 |
| male\_happy | 96 |
| female\_disgust | 96 |
| male\_disgust | 96 |
| female\_angry | 96 |
| male\_angry | 96 |
| female\_fear | 96 |
| male\_fear | 96 |

Table 2: emotions in the combined dataset

## Baseline 2D CNN

**RAVDESS vs RAVDESS + CREMA**

**Chart, line chart

Description automatically generated**

**Chart, histogram

Description automatically generated**

Model accuracy: 76.92%

**A picture containing background pattern

Description automatically generated**

**Chart, line chart

Description automatically generated**

**Chart, histogram

Description automatically generated**

**A picture containing treemap chart

Description automatically generated**

model accuracy: 57.71% (highest 59%)

**MFCC vs Melspec**

Chart, line chart

Description automatically generated

Chart, line chart, histogram

Description automatically generated

A picture containing treemap chart

Description automatically generated

Accuracy: 58.81%

## CNN + LSTM

start

# Discussion

start

# References

[1] Cs.toronto.edu. 2020. CIFAR-10 And CIFAR-100 Datasets. [online] Available at: <https://www.cs.toronto.edu/~kriz/cifar.html> [Accessed 1 November 2020].- sample

[2] sample