**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 Speech Emotion Recognition (SER) comes from the fact that human voice usually reflects underlying emotion through tone and pitch. The voice can be represented by data matrix (MFCC or Mel-spectrogram) and put into a neural network for training. CNN can capture the features of tone and pitch and classify an audio into the emotion categories, thus realizing SER through neural networks.

In this project, a dataset contains short audios speaking some sentences from different speakers. The dataset is processed into matrix representation. Then it is used for training a CNN. Efforts are spent on improving the CNN with techniques like network with attention.

**[write a bit more details on the techniques here if feel appropriate]**

The goal is to achieve reasonable accuracy when the neural network model does classification on test data. At the same time, analysis is done on the misclassified part.

**[write a bit more about the analysis you did here if feel appropriate]**

# 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 feature extraction method is needed.

**[write a bit more about how feature extraction matches what kind of nn if feel appropriate]**

**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

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

1. 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

1. 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 dataset has 8 commonly expressed emotions.

In theory, the emotions are recognized through pitch and tone. The common sense is that the pitches of male and female voices are different. To exclude the effect of gender during emotion recognition, each emotion label is attached with gender. In the end there are 14 labels.

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 |
| male\_surprise | 96 |
| female\_surprise | 96 |

1. emotions in the RAVDESS dataset

Comparing sentences:

As shown from the figure below, different sentences uttered have different wave plots. However, the two wave plots 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

1. female, intense angry, speaking "Kids"

Chart

Description automatically generated

1. 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

1. 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

1. male, intense angry, speaking "Dogs"

Text, whiteboard

Description automatically generated

1. 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 CREMA-D dataset is used to train the autoencoder which later will be used to improve the CNN model.

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

1. female, intense angry

|  |  |
| --- | --- |
| female\_neutral | 512 |
| male\_neutral | 575 |
| female\_sad | 600 |
| male\_sad | 698 |
| female\_happy | 600 |
| male\_happy | 698 |
| female\_disgust | 600 |
| male\_disgust | 698 |
| female\_angry | 600 |
| male\_angry | 698 |
| female\_fear | 600 |
| male\_fear | 698 |

1. emotions in CREMA-D dataset

## Baseline 2D CNN

**MFCC**

**Chart

Description automatically generated**

1. RAVDESS dataset, MFCC, train & test acc plot

Model accuracy: 71.11%

**Chart, histogram

Description automatically generated**

1. RAVDESS dataset, MFCC, train & test loss plot

**Background pattern

Description automatically generated**

1. RAVDESS dataset, MFCC, testing result confusion matrix

**Melspec**

Chart, line chart

Description automatically generated

1. RAVDESS , melspectrogram, train & test acc

Model testing accuracy: 73.61%

Chart, line chart, histogram

Description automatically generated

1. RAVDESS, melspectrogram, train & test loss

Background pattern

Description automatically generated

1. RAVDESS + CREMA, melspectrogram, confusion matrix

**Discussion:**

**MFCC vs Melspectrogram**

In the experiment, feature extraction from both MFCC and melspectrogram reached accuracy over 70%, with melspectrogram performing better. As discussed in the methodology section, melspetrogram performs better in CNN models. Therefore, melspetrogram is used for feature extraction in the improved CNN networks.

**Analysis on the Confusion Matrix**

As shown in the confusion matrix, the most well classified categories are strong emotions, like angry, fear, sad and happy. Neutral are well classified. Others like disgust are not well classified, easy to mix with fear, happy neutral and sad.

There are some other points which deserves notice: neutral is often misclassified as sad, fear is often misclassified as sad.

There are some other points which deserves notice: neutral is often misclassified as sad, fear is often misclassified as sad.

Comparing male and female, the emotions from male speaker do not classify as well as female speaker. Further analysis about how gender affects the emotion classification is done on the improved model.

## CNN + LSTM

start

# conclusion

## Baseline 2D CNN

In the standard 2D CNN model, the experiment achieved accuracy over 70% for two feature extraction methods, MFCC and melspectrogram. From analysing the result using confusion matrix, the misclassified labels and how gender affects the result are discussed.

# References

[1] sample

[2] sample