EEL6825 Project Report
Speech Emotion Recognition (April 2022)

Nan Mu

UFID: 69253367

Department of Electrical & Computer Engineering

***Abstract*— Emotion recognition is an important but challenging subject in recent years. This project explores the speech emotion system (SER) with several models (MLP, SVM, CNN) to recognize different emotions correctly. The dataset used in this project is The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). Some features such as Mel-frequency Cepstral Coefficients (MFCCs), Mel-scaled spectrogram and Chromogram are extracted to catch time-frequency domain information then be fed into models, data augmentation method is also implemented to improve the training performance. The SER system shows that the MLP have the classification accuracy of 60.85% for total 8 classes, SVM model have the accuracy of 68.67% for total 8 classes and the CNN model acquire the accuracy of 79.52% for total 10 classes.**

***Index Terms*— pattern recognition, speech emotion recognition, machine learning, CNN, feature extraction.**

I. Introduction

emotion recognition is an important subject to explore in long time. In recent years, Speech emotion recognition, which expects to investigate the emotion states through speech signals, has been receiving more attention. Generally, emotions can be divided into positive, negative, natural, and more types of emotions which usually contains various emotions such as angry, boring, disgusted, surprised, fearful, joyful, happiness, neutrality, and sadness [1]. Understanding and responding correctly to other peoples' motions play a fundamental role in our daily life [2]. As the advent of the artificial intelligence age, speech emotion recognition developed into one of the key components of human-computer interaction (HCI) such as speech emotion recognition (SER) can be used in the field of customer service intelligence, smart equipment services, autopilots, medical diagnostics, and psychological assistants.

SER aims to automatically identify a person's current emotion status from their speech voices. There should be easier or subconscious for a people to distinguish different emotions. Nevertheless, for machines or computers, SER remains a challenging task, with introducing the problem of how to extract emotional features effectively and how to let machines to recognize correctly among the different emotions [3]. That is, how we make computers understand and capture people emotion just like human itself.

As discussed above, there are three critical challenges for the SER problem, the first is the datasets choice, second is the technology of the speech feature extraction, Last is different classification methods implemented in emotion recognition.

In this project, the "Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)" datasets are used as the voices source to train and test the SER models. MLP method are used to get the baseline performance, then SVM model and CNN model are built for SER system. The advantage and disadvantage for these models are discussed briefly in the "relate work" section. Also, the comparison about the model accuracy is test on the MLP and SVM model for only Mel-frequency Cepstral Coefficients (MFCCs) and the combination of multitype features on CNN model, as well as some data augmentation method be added to improve the performance.

The paper structure is as follows. Second part is description, there contains several sections, one section is the description of the datasets used in this project; Then, detailed feature extraction techniques and SER model method are introduced in second and third section. Third part is evaluation, it shows the experiments and results about the accuracy acquired in different models, and the analysis for the outcome and performance. Forth part brief describing the related work about SER problem in recent years. At the last part, the conclusion and summary are made for this project.

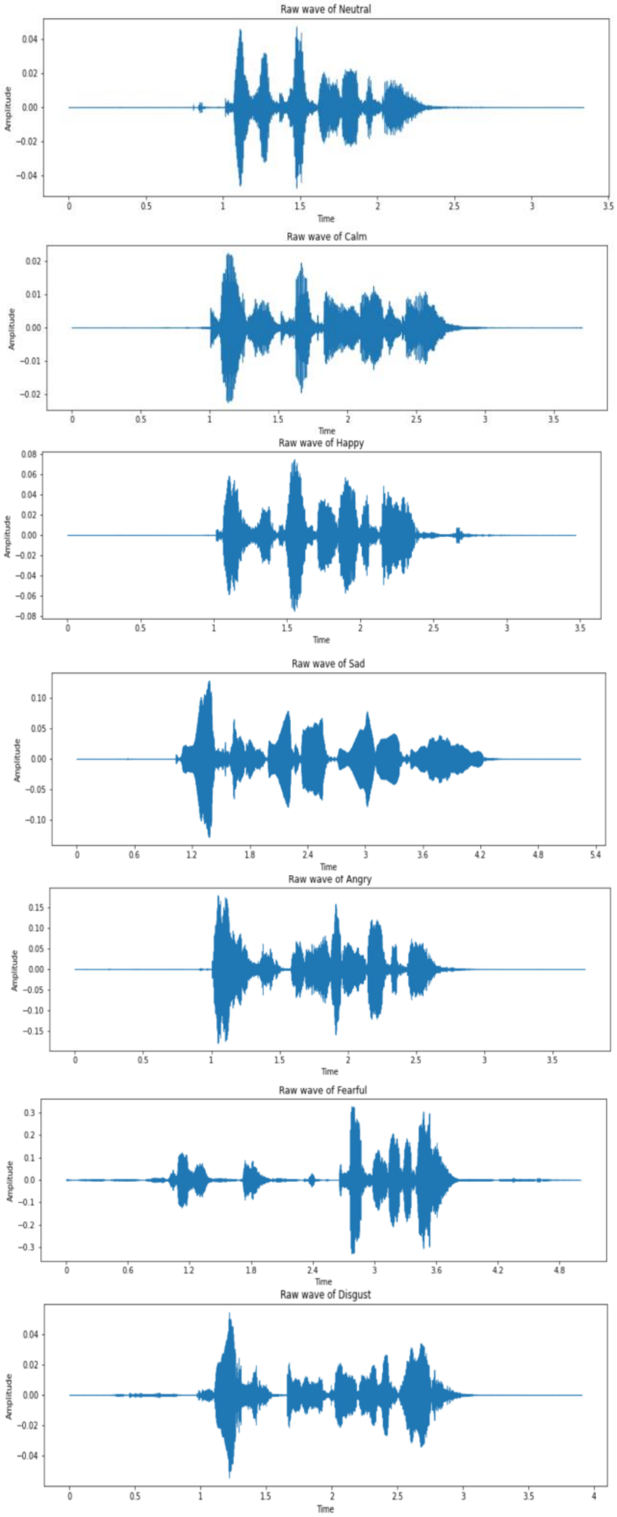
II. Description

A. Datasets

In this project, the dataset be chosen for speech emotion recognition is "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)", the reason why chooses this dataset is that it has a great availability. The RAVDESS is a suitable multi-emotional database of speech, video and song. The database consisting of 24 professional actors, perform the speech voice professionally in a neutral North American accent [4]. 12 male and 12 female actors pronouncing the sentence with eight different emotion expression which includes calm, happiness, sadness, anger, fearful, surprise, disgust and neutral in each Speech file, and Song files contain calmness, happiness, sadness, angry, and fearful emotions. RAVDESS provide three formats:

Audio-only, Audio-Video, and Video-only (no sound). In this project, the Audio-only format will be used for the SER training and testing samples. There are 1440 speech files with 60 trials per actor multiply 24 actors, and 1012 song files with 44 trials per actor multiply 23 actors, the total samples are 1012 add1440 that equal 2452.

Figure1 shows waveform of each of 8 emotions provided in this dataset:



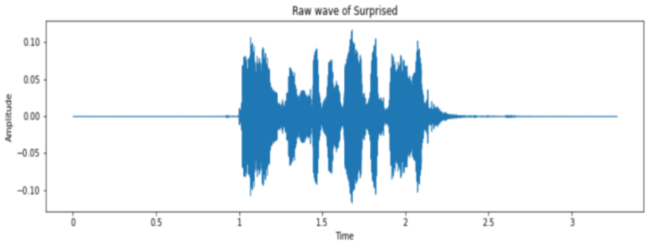


Fig. 1. Eight emotions of raw waveform

B. Feature Extraction Techniques.

The emotion expressed in speech is represented by the large number of features it contains, which vary with emotion. Feature extraction plays an important role in the Speech Emotion Recognition system. It is abundant and complex, there is no commonly acknowledged arrangement of features for extraction [5]. Some prosodic features such as rhythm, pitch, amplitude, energy, or speech rate are applied in the SER system. On the other hand, spectrum features such as “Linear Prediction Coefficients (LPC)”, “Mel-Frequency Cepstrum Coefficients (MFCC)” convey much information from voice which play a very important role in the SER problem.

Suitable features extraction will lead to a good training model, whereas inappropriate features or not enough features will have bad influence for the correct model building. In this project, “Mel-frequency Cepstral Coefficients (MFCCs)” are applied as the speech features used in SER model, then more spectral features that are “Mel-scaled spectrogram” and “Chromogram” be processed as input features for the SER system performance improvement. During the feature extraction stage, the python tool of “LibROSA” is applied for the audio analysis and feature extraction.

(1) chromogram

“Chroma” features will distinguish about 12 different pitches. It is powerful tool for analyzing sounds by clearly and meaningfully categorize those pitches’ information. “Chroma” captures melodic and harmonic information of voice and is robust to large disturbance. In this project, the short time Fourier Transform is used to convert sound file into “chromogram”. These “chromogram” features then are fed into the system model to do recognition for different emotions.

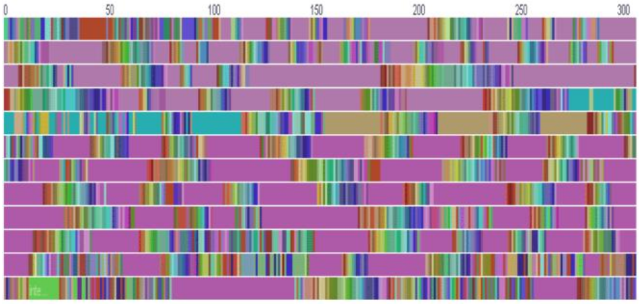


Fig. 2. Chromogram sample graph
http://www.katehollenbach.com/chromogram/

(2) Mel-scaled spectrogram

“Mel spectrogram” is a spectrogram where the frequencies are converted to the Mel scale, and the Mel scale is the result of nonlinear transformations from the frequency. Using "Mel-spectrogram", the voice will be seen and be processed with many scales of speech information. The Hertz scale are split into segments, overlapping triangular filters are used to convert each time segments to the Mel-scale segment then acquire the visualization of voice in Mel spectrum range. Figure 3 shows a sad voice sample of Mel-scaled spectrogram.

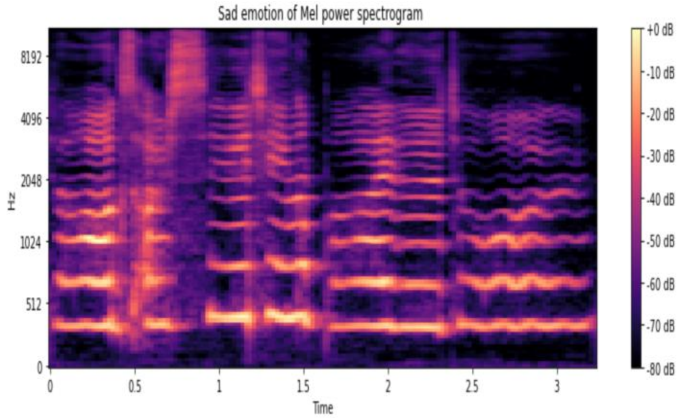


Figure .3 a sad voice sample of Mel-scaled spectrogram

(3) Mel-frequency Cepstral Coefficients (MFCCs)

In Mel scale range, MFCC uses a "quasi-logarithmic" spaced frequency scale, compared with the Mel-spectrogram which use a linear spaced frequency scale, MFCC has a relatively good decorrelation performance, and it is considered as the "spectrum of the spectrum". MFCC represents the short-term power spectrum of a speech frame using a linear cosine transform of the log power spectrum on Mel frequency scale [6]. It is widely used in speech recognition research, reflecting the shape of the voices.

The formular for converting the frequency to Mel-scale is that

$$M(f) = 1125 \ln(1 + f/700)$$

The steps for how to calculate MFCCs is depicted in below:[7]

Firstly, Fourier transform of a signal be processed with a "hamming" window.

Secondly, transfer the spectrum power to the Mel-scale, the "triangular overlapping window" are used to do the cosine overlapping process.

Thirdly, DCT are be implemented to get the Mel log power and Mel frequencies.

The MFCCs will be the amplitudes of the resulting spectrum.

Then the MFCCS features are fed into speech model to do training working.

Figure 4 is a sad emotion sample MFCC

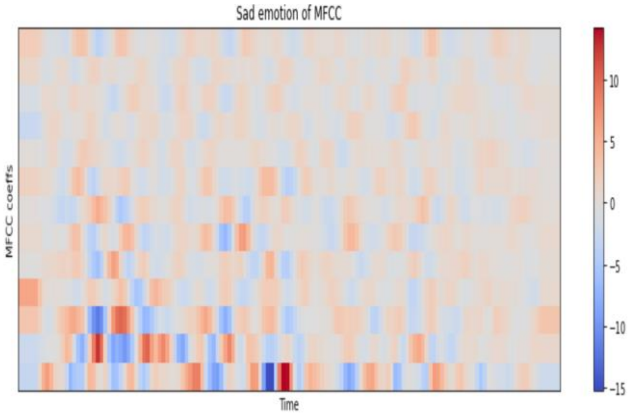


Figure. 4 a sad voice sample of MFCC

"Mel-scaled spectrogram" and "MFCCs" imitate how the human receive the voice information. However, both Mel-scaled spectrogram and MFCCs tend to be weak ability in a distinguishable representation of pitch classes and harmony even though they are useful for identifying and tracking fluctuations in sound [8]. So, the "Chromogram" are implemented as supplementary feature as the description above section, that allows system to receive both time domain and frequency domain information.

During the training and testing stage, it shows that using multiple voice features combined with three methods mentioned above not just one feature (for example, only MFCC) will improve the performance of the emotion recognition model. By using all three-features extraction method, the total features for each sound sample are 160.

C. SER Model Methods

Traditional classifier model and deep learning classifier method are used to do the classification about the multiple emotions which are received from the human speech sources. Referring to SER research and papers, different classifiers are used to explore the SER problem, but it is hard to define which type of classifier is best. In this project, three classifiers are ported and test their performance about the SER problem. They are support vector machine (SVM), multilayer perceptron (MLP) and convolution neural networks (CNN). This part will describe these models' characteristics and how to implement in this project.

(1) *Multilayer Perceptron (MLP)*

Multilayer Perceptron (MLP) is a class of feedforward "artificial neural network (ANN)", it is built by an input layer, at least one hidden layer, and an output layer. MLP using backpropagation algorithm, which has forward computation followed by backward computation step for training, so, the general layer of MLP gets data information from lower layers and its class data from higher layers [9]. In this project, MLP model are implemented for the baseline task. It is built by one hidden layer with activation function "ReLu", solver" Adam", batch size is "128", the hidden layer nodes is "200".

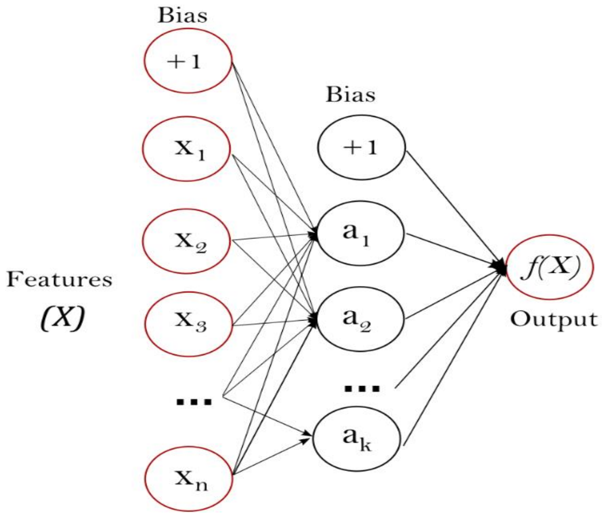
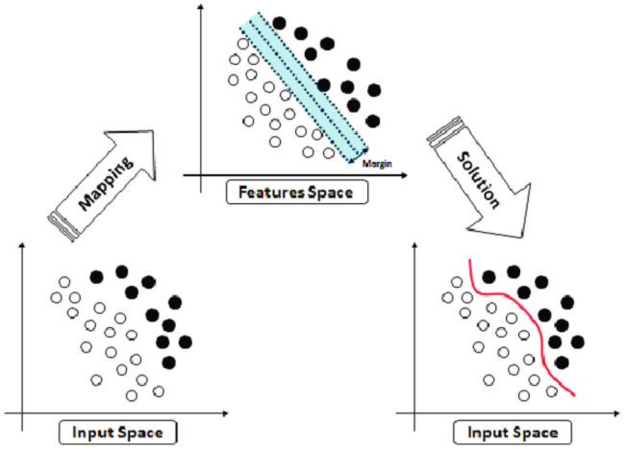


Figure.5 MLP Model
Cited from:
https://scikitlearn.org/stable/modules/neural_networks_supervised.html

(2) *Support Vector Machine (SVM)*

Support Vector Machines are depicted as supervised learning model, it is implemented for analyze classification problem. It has been implemented for a classifier used in the emotion recognition in much research and showing a good performance. SVM separates linear or non-linear features from the input data, and transform the output into high dimensional features, then the high dimension space separates the data into different classes [10].



Cited from: DOI:10.13140/RG.2.1.4573.8325
Figure.6 SVM Model sample

SVM has different types of kernels such as Radial Basis Function (RBF) kernel, or linear kernel, Gaussian kernel, or polynomial kernel and each of kernel produce different results in emotion classification. The two most widely chosen kernel in SER is linear and RBF, selection about kernel depend on the dataset. In general, if the features have a large number but we only have a small training data, using a linear kernel might be a good option. If there is little feature but lots of dataset's input, then using an RBF kernel can be better.[10] There are two parameters "C" (regularization parameter) and "gamma" (parameter of RBF) that should be considered when training SVM model. Larger C is relative to low error tolerance, smaller gamma means more vectors supporting.

In this project, each data sample are treated as a point with the number features (dimension space) extracted from the voice. linear and Radial Basis Function kernel are chosen to do test and parameter C and gamma are set to 10 and 0.0001. Also, standard scaler is used to normalize the dataset.Because the large range scaling feature dimension will influence more for the distance than other small dimensions. Using scaled method can also be helpful for the calculation process.

(3) *Convolution Neural Network (CNN)*

Convolutional Neural Networks (CNN) is the systematic neural network that consists of various layers sequentially. The architecture of Convo net is same as the connection patterns of neurons in the human brain. In general, each neurons in CNN are sensitive to the stimulation within a little space that we call it the "receptive field", each field overlap by a portion of the space then all the neurons will cover the whole feature fields.

CNNs have be seen that it is significant for recognition problem which contains image, video, and sound. the CNN model usually consists of multiple convolution layers, pooling layers, flatten layers, and a SoftMax layer.

In this project, CNN model contains one-dimensional convolutional layers combined with max pooling layer, dropout, batch normalization, and activation function. The feature which comes from the "Mel-scaled spectrogram", "MFCCs" or "Chromagram" are fed into convolution layer, batch normalization layer is used for normalizing the output which are friendly to avoid feature display the different distributions among the training and test dataset. Using downsample method, Max-pooling layers can weak the sensitivity for the location of feature, and collect the most features as much as. The "ReLU" activation function will contribute to the sparsity of the network, thereby try avoiding the inter dependency between parameters and decrease the "vanishing gradient" problem, also, dropout layer will invalid to some features randomly in some layer so that it is helpful to lowdown for the over-fitting problem.

The detail of the structure is depicted as follow:

The first layer of the CNN Model receives 160 *1 features as input data, it composes with 256 filters and the kernel size is 8*8, stride is 1, then one more same convolution layer is added with "ReLu" activation function. After that batch normalization is implemented, and using "ReLU" function to activation the output, dropout layer that be set to the rate of 0.25 are also added, the outputs are fed to the max-pooling layer which has a 8 size of window. Next part contains 4 convolution layers with 128 filters and kernel size 8*8, stride is 1, each layer followed by the ReLu activation and then batch normalization, dropout with 0.25 are implemented then using one max pooling layer receive the feature. Final part contains two convolutional layers with 64 filters and RELU activator, one flatten layer added to make output compatible with next layer, the fully connected layer with SoftMax function are added to changing the output elements to the number of emotions suspected and estimation the probability distribution for each emotion class. This model use SGD as the optimizer with the learning rate of 0.0001. Figure 7 shows the CNN Model structure used in this project.

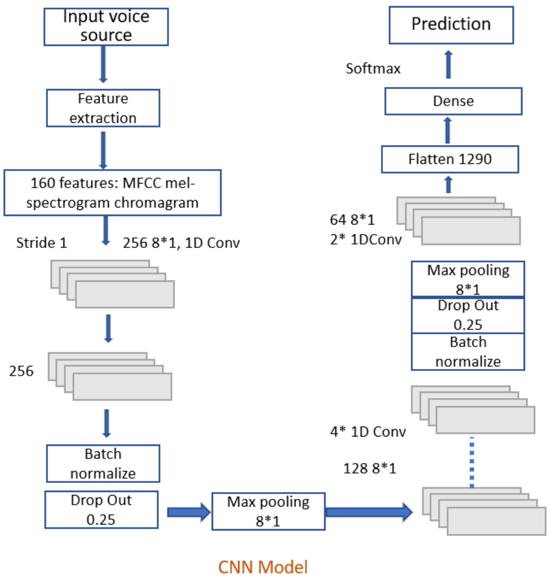


Figure. 7 CNN Model used in this project

III. Evaluation

A. *Overall Results*

After finishing the dataset preprocessing, feature extracting and model building, the dataset is divided for the training dataset, test dataset and final validation dataset, where the ratio is 75% training, 25% testing in MLP and SVM model, For CNN model the train set is 80% for total are used to training model with the 0.8*02=0.16 used for the CNN model to adjust training process and test model accuracy during the training progress. Then the final validation dataset is 20% for total used to test the trained model to see how correctly the model can recognize different emotions.

The experiment progress is divided several stages, at the first stage of this project, only MFCC features are used as the input, MLP, SVM model are trained firstly with 8 emotions. After analyzing the outcome and combine the realistic that mainly emotion used in daily life, adjust classification regulation, test 5 emotions with the distinguish of male and female in CNN Model. In order to try to find the improvement method, more features mentioned in section description (Mel scale and chroma) are introduced the feature extraction set. Then more data augmentation methods are implemented to try to improve the correctness. Table 1 shows the results of experiments and a performance summary table for each experiment. We can see the CNN model perform best in these three models which get nearly 80% average accuracy, SVM model with scaled method perform better than one other SVM model, the MLP model which is used for baseline has relative low correctness.

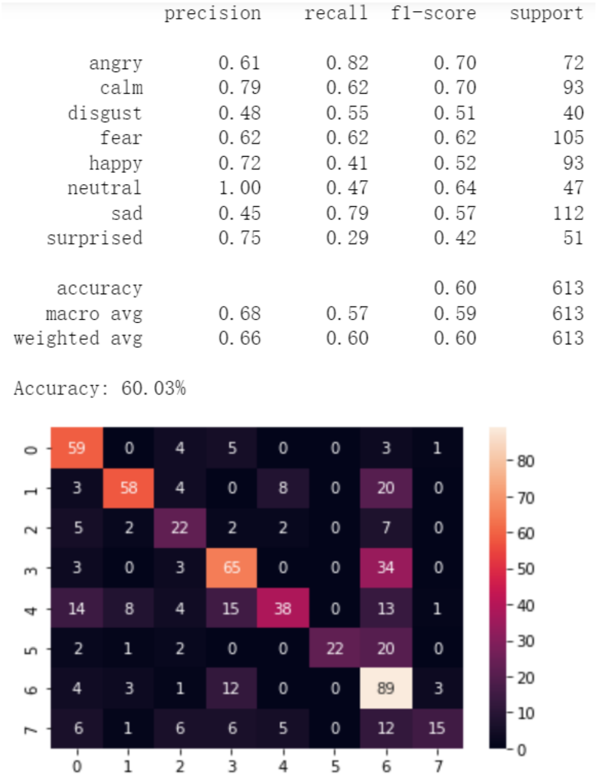
model	feature	Emotion classification	Test Accuracy
MLP	MFCC	8 classes	60.85%
SVM1 (‘rbf’ kernel)	MFCC	8 classes	65.25%
SVM2 (linear kernel + scale)	MFCC	8 classes	68.67%
CNN	MFCC +Mel scale + chroma	10 classes (5male+5 female)	69.91%
CNN	MFCC +Mel scale + chroma +data augmentation	10 classes (5male+5 female)	79.59%

Table 1 performance summary table

B. *Experiment for MLP and SVM Model*

Firstly, MLP Model are used as the performance baseline which get highest accuracy about 60.85%. SVM model with an ‘rbf’ kernel has a 75.36% training accuracy and 65.25% testing accuracy, then test another method for SVM, by adjusting and adding scaled method with linear kernel, it acquires the has an 80.09% training accuracy and 68.67% testing accuracy.

Below figure 8 shows the details of every emotion classification accuracy for MLP we can see the emotion of angry present the highest correctness, but the emotion surprise performs bad on MLP model which only get 0.42 f1-score and the surprise, sad and happy are also perform not good enough which sad emotion get 0.57 f1-score and happy emotion have 0.52 f1-score, disgust only get 0.51 f1-score. From the figure, we see that most wrong classify emotion are mainly trended to be attributed to sad emotion, 20 cases of calm, 34 cases of fearful and 20 cases of neutral emotions are attributed to the sad emotion. That shows a confusion for these emotion classifications.



```
              precision    recall  f1-score   support

       angry       0.61      0.82      0.70        72
        calm       0.79      0.62      0.70        93
     disgust       0.48      0.55      0.51        40
        fear       0.62      0.62      0.62       105
       happy       0.72      0.41      0.52        93
     neutral       1.00      0.47      0.64        47
         sad       0.45      0.79      0.57       112
   surprised       0.75      0.29      0.42        51

    accuracy                           0.60       613
   macro avg       0.68      0.57      0.59       613
weighted avg       0.66      0.60      0.60       613

Accuracy: 60.03%
```

	0	1	2	3	4	5	6	7
0	59	0	4	5	0	0	3	1
1	3	58	4	0	8	0	20	0
2	5	2	22	2	2	0	7	0
3	3	0	3	65	0	0	34	0
4	14	8	4	15	38	0	13	1
5	2	1	2	0	0	22	20	0
6	4	3	1	12	0	0	89	3
7	6	1	6	6	5	0	12	15

Figure. 8 MLP model prediction results

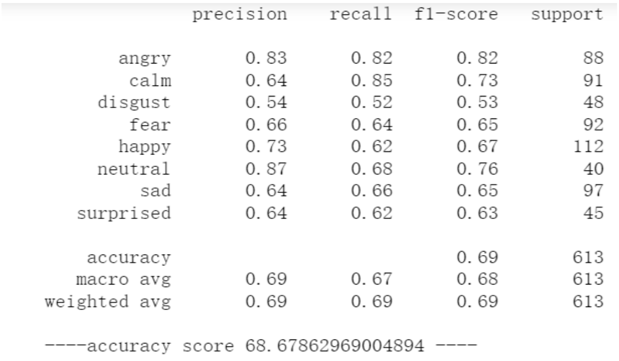
Figure 9 shows the details of every emotion classification accuracy for SVM model. In SVM model, the emotion of angry perform great which get the 0.82 f1-score. The worst result for SVM comes from surprise emotion witch only get 0.53 f1-score, the sad and disgust are not performed well, which the sad emotion get 0.65 f1-score, and the disgust emotion get 0.53 f1-score.

Compared with the MLP Model, some emotion confusion problem has improved such like calm, fearful and neutral emotions have been classified more correctly that only 7 cases of sad, 12 cases of fearful and 4 cases of neutral are attributed to sad emotions. below table shows the changes in these three emotion classifications.

Emotion Wrong to sad	calm	fearful	neutral
MLP	20	34	20
SVM	7	12	4

Table 2 confusion emotion in MLP and SVM model

.

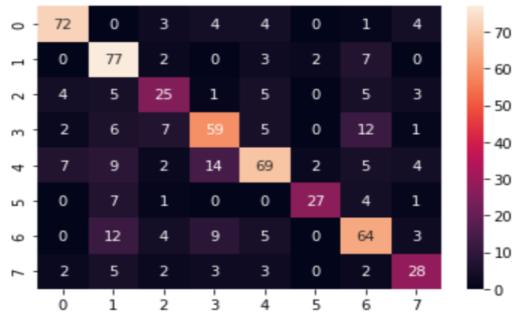


```
              precision    recall  f1-score   support

       angry       0.83      0.82      0.82        88
        calm       0.64      0.85      0.73        91
     disgust       0.54      0.52      0.53        48
        fear       0.66      0.64      0.65        92
       happy       0.73      0.62      0.67       112
     neutral       0.87      0.68      0.76        40
         sad       0.64      0.66      0.65        97
   surprised       0.64      0.62      0.63        45

    accuracy                           0.69       613
   macro avg       0.69      0.67      0.68       613
weighted avg       0.69      0.69      0.69       613

----accuracy score 68.67862969004894 ----
```



	0	1	2	3	4	5	6	7
0	72	0	3	4	4	0	1	4
1	0	77	2	0	3	2	7	0
2	4	5	25	1	5	0	5	3
3	2	6	7	59	5	0	12	1
4	7	9	2	14	69	2	5	4
5	0	7	1	0	0	27	4	1
6	0	12	4	9	5	0	64	3
7	2	5	2	3	3	0	2	28

Figure. 9 SVM model prediction results

According to the real requirements, I think that neutral, disgust and surprise these three emotions are not typical used in daily emotion recognition, so to study how to implement a suitable model, I will focus on the mainly 5 emotions that is angry, calm, fear, happy and sad these positive negative and neutral types of emotions. Also, I think whether classification for the male and female with different emotion is important. In CNN Model, the emotion classification task is 10 types with female-calm, female-happy, female-fear, female-angry, female-sad, and male-calm, male-happy, male-fear, male-angry, male-sad.

C. *Experiment for CNN Model*

According to the realistic application, the emotion classification is adjusted. In this part, two experiment for CNN Model are test, one three type features which are Mel-frequency Cepstral Coefficients (MFCCs), Mel-scaled spectrogram and Chromogram be extracted into model for training, other is combing three types of features with the data augmentation method, layer structure adjusting, dropout rate changing are used to change the performance in next experiment and then only the data augmentation method enhance the accuracy.

For three types of features, 400 epochs are trained to get the CNN test model, then using the final test dataset to test the performance of this model, the results is that 79.09% for training accuracy test 80.85% for validate accuracy and 69.91% for final testing accuracy. From the model loss curve figure, we can see that the training loss and validation loss has same trend and has a good convergence, without too much gap between them. That shows the parameter and structure of this model are suitable for speech emotion test. Figure 10 shows the CNN model loss curve and figure 11 shows the classification prediction for each motion.

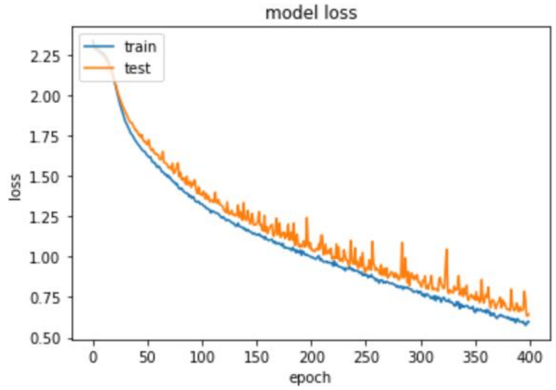
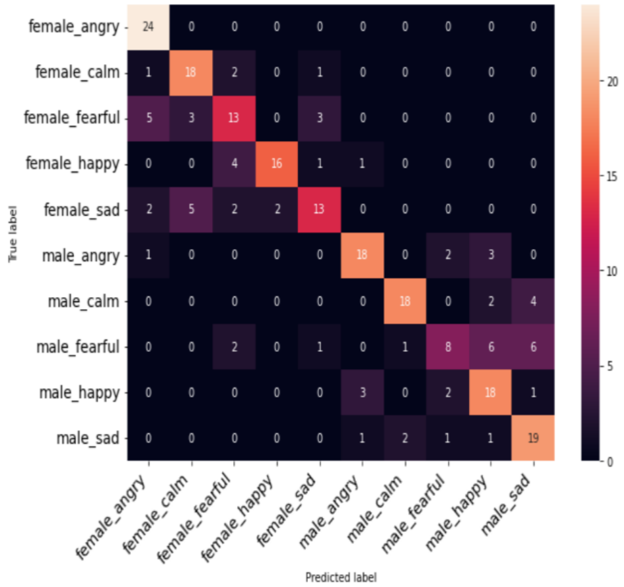


Figure. 10 CNN model training and validation loss curve



	female_angry	female_calm	female_fearful	female_happy	female_sad	male_angry	male_calm	male_fearful	male_happy	male_sad
female_angry	24	0	0	0	0	0	0	0	0	0
female_calm	1	18	2	0	1	0	0	0	0	0
female_fearful	5	3	13	0	3	0	0	0	0	0
female_happy	0	0	4	16	1	1	0	0	0	0
female_sad	2	5	2	2	13	0	0	0	0	0
male_angry	1	0	0	0	0	18	0	2	3	0
male_calm	0	0	0	0	0	0	18	0	2	4
male_fearful	0	0	2	0	1	0	1	8	6	6
male_happy	0	0	0	0	0	3	0	2	18	1
male_sad	0	0	0	0	0	1	2	1	1	19

Figure. 11 CNN model classification results

From the prediction label figure about the final testing, most emotions get a correctness classification, this model can distinguish male and female voice greatly which has more than 95% accuracy only has 1 error shown in female group and 4 errors shown in male group. And the female-angry emotion performs best in these ten types of emotions.

The outcome of CNN is better than the MLP Model and SVM two model, which means that CNN model with more types of features may improve the performance.

Next step, I try other method such as data augmentation method or adjusting CNN Model parameters to study how to improve the SER classification accuracy.

D. *Experiment for CNN Model with Data Augmentation*

From the part C, the CNN Model get 69.91% for final testing accuracy, this outcome is near to some research's result. Because CNN model needs larger datasets to train, but the RAVDESS dataset only contains 2452 voices, and only has 1504 sound sources for training the model (376 for final testing), so in this part, the data augmentation method is used for amplifying numbers of training data. I use two methods by adding white noise and shifting the voice segment that broaden the training dataset from 1504 to 4512. These five emotions are total 3609 samples. Then using these datasets to train the CNN model. After several times experiment, below is the best outcome I get.

After 400 epoch training, the CNN Model has the following performance:

88.86% for training accuracy test 90.25% for validate accuracy and 79.52% for final testing accuracy.

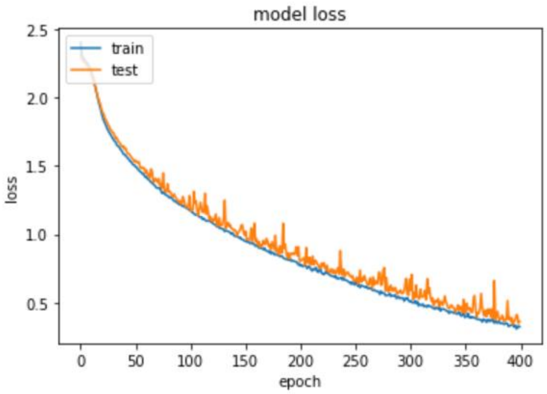
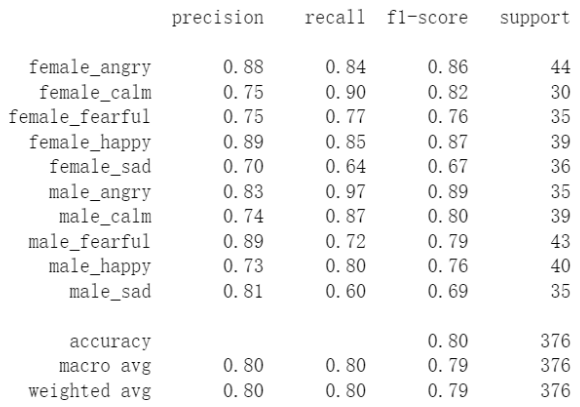


Figure.12 CNN Model training and validation loss curve



```
               precision    recall  f1-score   support

 female_angry       0.88      0.84      0.86        44
  female_calm       0.75      0.90      0.82        30
female_fearful      0.75      0.77      0.76        35
 female_happy       0.89      0.85      0.87        39
   female_sad       0.70      0.64      0.67        36
   male_angry       0.83      0.97      0.89        35
    male_calm       0.74      0.87      0.80        39
 male_fearful       0.89      0.72      0.79        43
   male_happy       0.73      0.80      0.76        40
     male_sad       0.81      0.60      0.69        35

     accuracy                           0.80       376
    macro avg       0.80      0.80      0.79       376
 weighted avg       0.80      0.80      0.79       376
```

Figure.13 CNN Model classification with data augmentation performance table

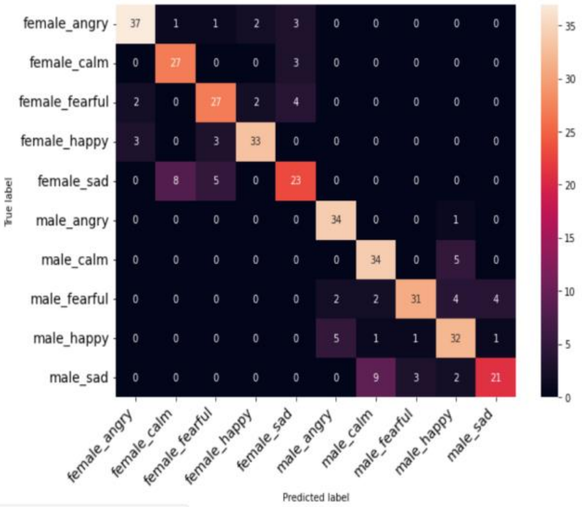


Figure.14 CNN model prediction results with data augmentation

From the figure 12 of model loss, the training loss and validation loss has same trend and has a good convergence just like the result got from part 2. Compared with experiment 2, the loss value become lower. Figure 13 & 14 showing the CNN prediction results with data augmentation method. From the final predicted label figure, we can see, this experiment has better performance than part2 in classify the male and female which get 100% accuracy.

Some emotion correctness such like fearful has got improvement. Compared with the MLP and SVM model, the confusion between sad and other emotions has been improved. But sad emotion for both female and male still has relative bad performance, as shown in figure, we can see the emotion sad is easy confused with emotion calm, for female, there are 8 cases of sad are thought to female sad, 3 cases of calm are thought into female sad. For male, there are 9 cases of calm are classified in to the male sad. That may be these two emotions is similar such like low pitch, not excited expression which causes the confusion.

IV. Related work

Emotion recognition has been explored in long time as an important subject, especially in these 10 years. As mentioned above speech emotion recognition has mainly two challenges which are if there is a great method to extract useful emotion features to distinguish emotions, and if there have an optimization method to recognize correctly for each emotion. Apart from these two tasks, the datasets also have an important influence for SER system that different characteristic speech datasets may be suitable with different model.

According to search papers and websites, we can see in recent years, many speech databases related to speech emotion research have been established. For example, the "Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)" dataset, "Berlin (EMO-DB)" dataset, and "Interactive Emotional Dyadic Motion Capture (IEMOCAP)" datasets and A "Mandarin Emotional Speech Corpora (MESC)" are used in much relative research. each datasets have its characteristics such like they can be divided into acted, natural, elicited and simulation type [3]. Using these datasets, researchers will capture different kinds of features from source of voices. The most important feature for the SER system is the speech feature, specific feature extraction combined with many kinds of model algorithm have improved the classification accuracy. However, there is no recognized feature arrangement for exactly accurate classification, nor a perfect optimal model algorithm, from now on, the existing research is experimental [11]. There has many research using different feature extraction technics with classification model to acquire good performance. there are mainly two types of methods which are traditional machine learning method (ML) and deep learning algorithms (DL) be implemented to recognize correct emotion.

For SER problem, many traditional classifiers have been implemented in this area such as and Decision Tree, the Support Vector Machine (SVM) and Maximum Likelihood Principle (MLP). These methods are valid for SER problem. For instance, Multilayer Perceptron often used as a non-linear classifier to solve the classification tasks. [10], the advantage is that it provides a high-speed predictions in testing procedure and be suitable for a relative small dataset, so it is good for SER problem as we know the voice datasets usually has not enough sources. However, it has some disadvantage such like it is not easy to define how degree each independent variable is affected [11]. SVM is widely used for pattern classification problems, the advantage for the SVM model is that it is more effective in high dimensional spaces, and it can prevent the overfitting, under the conditions of limited training data, it can have a very good classification performance [9]. But it also has some disadvantage that it may perform bad when the datasets contain many noises, how to select a suitable kernel for specific problem is not easy to define. Deep Learning models have been used to build system and has gained more attention in recent years [12], For the CNN model, applying it to SER system, CNN will learn more stable voice features and it is more robust for the noise disturbance. However, overfitting problem are supposed to be considered when training a CNN model, also, CNN model usually needs a larger dataset to do the training work.

By studying some references about the SER problem, some models are used in SER system and acquire a good performance. For example, Liu et al. [13] combining spectral and prosodic emotional features then using CNN and DNN method to build the model, they acquire 86% accuracy on "CASIA Chinese Emotion Corpus" dataset. Zamil et al [14]. Conveyed "Mel Frequency Cepstrum Coefficient (MFCC)" feature from two datasets; Emo-DB and RAVDESS, building a LMT classifier to recognize different emotions r and shows the best performance accuracy with 70% for seven different emotions. Shen et al [15]. build model based on SVM algorithm with mainly total 6 types of features such as energy, pitch, Mel Frequency cepstrum coefficients (MFCC), Linear Prediction coefficients and Mel cepstrum coefficients (LPCMCC), and the performance tested by the Berlin emotional database is 82.5%. Issa et al. [8] implement a 1D-CNN model and extracts Mel-frequency cepstral coefficients, chromogram, Mel-scale spectrogram, Tonnetz representation, and spectral contrast features, acquiring 71.61% for RAVDESS with 8 classes, 86.1% for EMO-DB in 7 classes. Apart from these approaches above, many model algorithm are used to SER problem such as Hidden Markov Model (HMM), K-Nearest Neighbor (KNN), Deep Belief Networks (DBN), Restricted Boltzmann Machine (RBM), Long Short-Term Memory (LSTM) [3].

V. SUMMARY AND CONCLUSIONS

Even though many kinds of feature technics and learning models are considered and reward some great grades, as we known, speech emotion recognition is not easy to realize because the emotion for everyone is subjective and is possible to be influenced by the individual speakers. So, how to extract suitable features and how to choose right model is critical for building an optimal SER system

Here is a summary about the work this project done.

Firstly, by reading paper and studying research of related works about speech emotion recognition (SER), I summarize related methods that SER problem uses recently.

Secondly trying to find a dataset which are suitable for this project, statistic the datasets information and preprocessing datasets.

Thirdly, using three different techniques to extract the voice feature and putting these features into the classification model.

Fourth step is experimenting and analyzing three different classification models which are MLP, SVM, and CNN model. Testing and analyzing different model performance include each model testing accuracy, f1-score, and the correctness for each emotion class. First experiment is training MLP and SVM models with MFCCs feature method, adjusting model parameters such as linear or "rbf" SVM kernel, adjusting SVM with scaled pipeline, then analyzing the results for these two models. Adjusting the classification type and broaden to three types of feature, training the CNN model get an improved result. Then attempting data augmentation method to broaden datasets, also attempting other method to enhance the CNN model and analyze the changes on the system performance.

In this project, by several experiments, I find that multiple feature types will enhance the accuracy of model, scale method is useful for SVM model, it performs better than non-scaling model. When training the CNN model, overfitting problem need to be drawn attention. Also, in the circumstance of relatively small datasets, some exploration of the data augmentation for datasets is useful for training a better model. From the section of evaluation, I present the overall performance for each model, it shows that the different model performs different correctness results, and from now on, the best result I get is comes from the CNN Model which has nearly 80% accuracy for the final testing. Most of emotions can be recognized greatly. However, there still has many aspects that can be used to try to improve the performance, such like the adjustment about the model structure and parameters, some attempt that I had done such like larger dropout ratio, changing another optimizer is less helpful for improving the performance. So, in the future works, there still needs more analysis and attempting to optimize the SER Model, and for some relatively easy confused emotion such like calm and sad, there needs to find improvement method to distinguish them more clearly.

VI. REFERENCE LIST

[1] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," in IEEE Access, vol. 7, pp.

[2] Lausen, A., Hammerschmidt, K. Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters. Humanit Soc Sci Commun 7, 2 (2020)

[3] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi and E. Ambikairajah, "A Comprehensive Review of Speech Emotion Recognition Systems," in IEEE Access, vol. 9, pp. 47795-47814, 2021, doi: 10.1109/ACCESS.2021.3068045.

[4] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391.

[5] M. Swain, A. Routray and P. Kabisatpathy, "Databases features and classifiers for speech emotion recognition: A review", Int. J. Speech Technol., vol. 21, no. 1, pp. 93-120, Mar. 2018.

[6] Y Han, G Wang and Y Yang, "Speech Emotion Recognition Based on MFCC", Journal of Chong Qing University of Posts and Telecommunication Natural Science Edition, vol. 20, no. 5, 2008.

[7] Sahidullah, Md.; Saha, Goutam (May 2012). "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition". Speech Communication. 54 (4): 543 – 565.

[8] Dias Issa, M. Fatih Demirci, Adnan Yazici, Speech emotion recognition with deep convolutional neural networks, Biomedical Signal Processing and Control, Volume 59, 2020, 101894, ISSN 1746-8094.

[9] Machine Vision Theory, Algorithms, Practicalities A volume in Signal Processing and its Applications Book • Third Edition • 2005

[10] P. P. Dahake, K. Shaw and P. Malathi, "Speaker dependent speech emotion recognition using MFCC and Support Vector Machine," 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), 2016, pp. 1080-1084, doi: 10.1109/ICACDOT.2016.7877753.

[11] Conference: y-BIS 2019 Conference: ISBIS Young Business and Industrial Statisticians Workshop on Recent Advances in Data Science and Business AnalyticsAt: Istanbul, Turkey

[12] R. Chinmayi, N. Sreeja, A. S. Nair, M. K. Jayakumar, R. Gowri and A. Jaiswal, "Emotion Classification Using Deep Learning," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), 2020, pp. 1063-1068, doi: 10.1109/ICSSIT48917.2020.9214103

[13] G. Liu, W. He and B. Jin, "Feature fusion of speech emotion recognition based on deep learning", 2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC), pp. 193-197, Aug 2018.

[14] A. A. A. Zamil, S. Hasan, S. M. Jannatul Baki, J. M. Adam and I. Zaman, "Emotion detection from speech signals using voting mechanism on classified frames", 2019 International Conference on RoboticsElectrical and Signal Processing Techniques (ICREST), pp. 281-285, Jan 2019.

[15] P. Shen, Z. Changjun and X. Chen, "Automatic Speech Emotion Recognition using Support Vector Machine," Proceedings of 2011 International Conference on Electronic & Mechanical Engineering and Information Technology, 2011, pp. 621-625, doi: 10.1109/EMEIT.2011.6023178.

[16] https://numpy.org/doc/stable/index.html

[17] https://www.tensorflow.org/resources/learn-ml/basics-of-machine-learning

[18] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0196391&fbclid=IwAR0pMF9vaxEvCucqjm2DJ1TH6CUv7JpBD79vi8qJcCAdHzJjJ4X2pFGDv_E

[19] https://tspace.library.utoronto.ca/handle/1807/24487

[20] https://zenodo.org/record/1188976#.YmRQ_dqZM2x