Speech Emotion Recognition Using Machine Learning

Preethi Jeevan[1], Kolluri Sahaja[2], Dasari Vani[3], Alisha Begum[4]

[1]*Professor,Dept.ofComputerScienceandEngineering,SNIST,Hyderabad-501301,India*
[2,3,4]*B.TECHScholars,Dept.ofComputerScienceandEngineeringHyderabad-501301,India*

---***---

Abstract: Speech *is the ability to express thoughts and emotions through vocal sounds and gestures. Speech emotion recognition (SER), is the act of attempting to recognize human emotion and affective states from speech. This is capitalizing on the fact that voice often reflects underlying emotion through tone and pitch..In this paper we analyze the emotions in recorded speech by analyzing the acoustic features of the audio data of recordings .From the Audio data Feature Selection and Feature Extraction is done. The Python implementation of the Librosa package was used in their extraction. Feature selection (FS) was applied in order to seek for the most relevant feature subset. We analyzed and labeled with respect to emotions and gender for six emotions viz. neutral, happy, sad, angry, fear and disgust, the number of labels was slightly less for surprise.*

Keywords: MFCC, Chroma, Mel Spectrogram, Feature selection

1. INTRODUCTION

Emotion plays a significant role in interpersonal human interactions. For interaction between human and machine use of speech signal is the fastest and most efficient method. Emotional displays convey considerable information about the mental state of an individual. For machine emotional detection is a very difficult task, on the other hand, it is natural for humans. So, knowledge related to emotion is used by an emotion recognition system in such a way that there is an improvement in communication between machine and human.

1.1 LITERATURE SURVERY

A majority of natural language processing solutions, such as voice-activated systems etc. require speech as input. Standard procedure is to first convert this speech input to text usingAutomatic Speech Recognition (ASR) systems and then run classification or other learning operations on the ASR text output. ASR resolves variations in speech transcriptions being speaker independent. ASR systems produce outputs with high accuracy but end up losing a significant amount of a speech from different users using probabilistic acoustic and language models which results in information that suggests emotion from speech. This gap has resulted in Speech-based Emotion Recognition (SER) systems becoming an area of research interest in the last few years .Three key issues for successful SER system

- Choice of a good emotional speech database.
- Extract effective features.
- Design reliable classifiers using machine learning algorithms.

The emotional feature extraction is a main issue in the SER system. A typical SER system works on extracting features such as spectral features, pitch frequency features, formant features and energy related features from speech, following it with a classification task to predict various classes of emotion . Speech information recognized emotions may be speaker independent or speaker dependent. Bayesian Network model, Hidden Markov Model (HMM), Support Vector Machines (SVM), Gaussian Mixture Model (GMM) and Multi-Classifier Fusion are few of the techniques used in traditional classification tasks.

In this work, we propose a robust technique of emotion classification using speech features and transcriptions. The objective is to capture emotional characteristics using speech features, along with semantic information from text, and use a deep learning based emotion classifier to improve emotion detection accuracies. We present different deep network architectures to classify emotion using speech features and text. The main contributions of the current work are:

1.2 PROPOSED METHOD

The speech emotion recognition system is performed as a Machine Learning (ML) model. The steps of operation are similar to any other ML project, with supplementary fine- tuning systems to make the model function adequately.

The fundamental action is data collection, which is of prime importance.

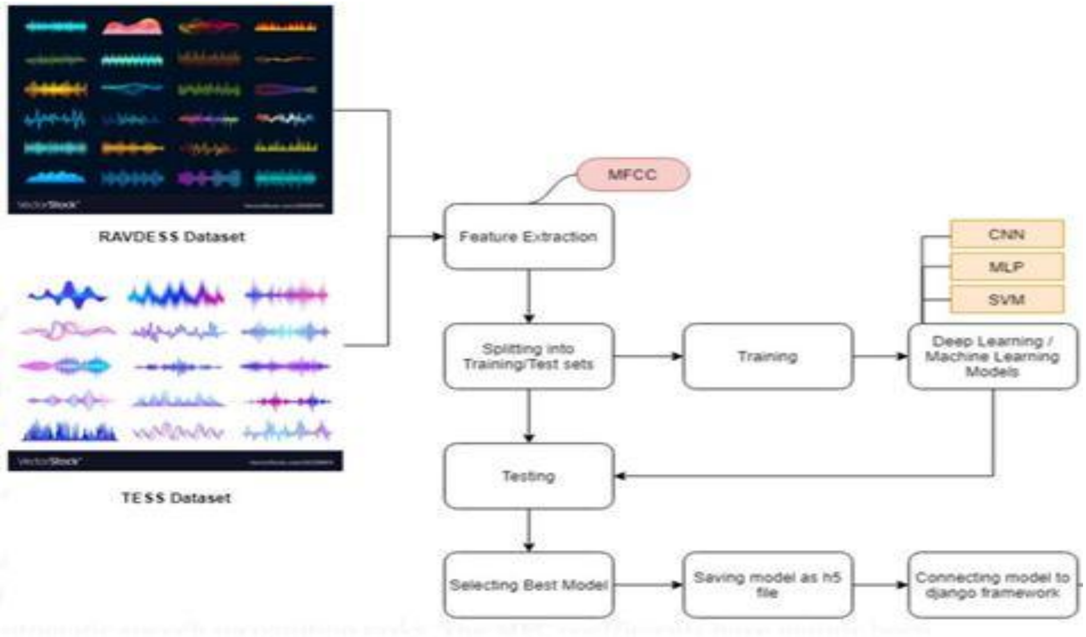
The model being generated will acquire from the data contributed to it and all the conclusions and decisions that a progressed model will produce is supervised data. The secondary action, called as feature engineering, is a combination of various machine learning assignments that are performed over the gathered data. These systems approach the various data description and data quality problems. The third step is often explored the essence of an ML project where an algorithmic based prototype is generated. This model uses an ML algorithm to determine about the data and instruct itself to react to any new data it is exhibited to. The ultimate step is to estimate the functioning of the built model. Very frequently, developers replicate the steps of generating a model and estimating it to analyze the performance of various algorithms. Measuring outcomes help to choose the suitable ML algorithm most appropriate to the p **Dataset*:***

English Language Dataset, namely the Toronto Emotional Speech Set (TESS) was taken into consideration. This is a dataset which consists of 200 target words and were spoken by two women, one younger and the other older, and the phrases were recorded in order to portray the following seven emotions happy, sad, angry, disgust, fear, surprise and neutral state. There are 2800 files in total, in this dataset.

CREMA-D is a data set of 7,442 original clips from 91 actors. These clips were from 48 male and 43 female actors between the ages of 20 and 74 coming from a variety of races and ethnicities (African America, Asian, Caucasian, Hispanic, and Unspecified). Actors spoke from a selection of12 sentences. The sentences were presented using one of six different emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad) and four different emotion levels (Low, Medium, High, and Unspecified).

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech emotions include calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.

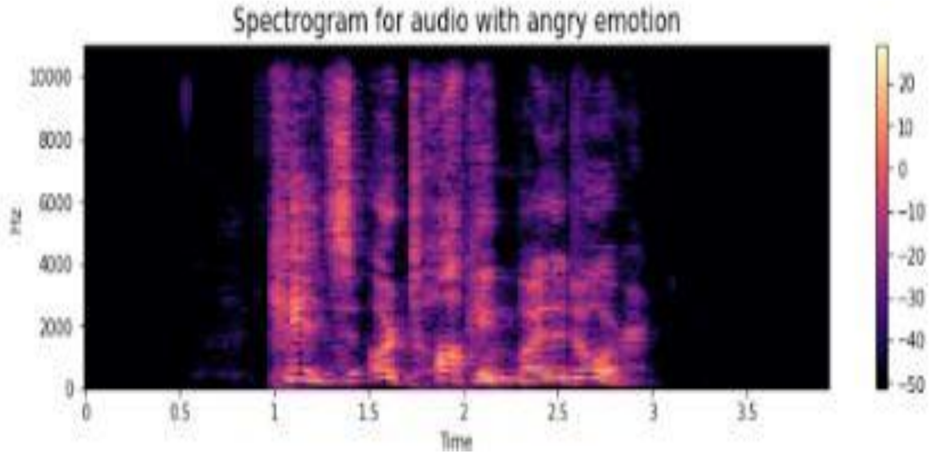
2.SPEECH EMOTION RECOGNITION SYSTEM



2.1 Data Set & Data Visualization:

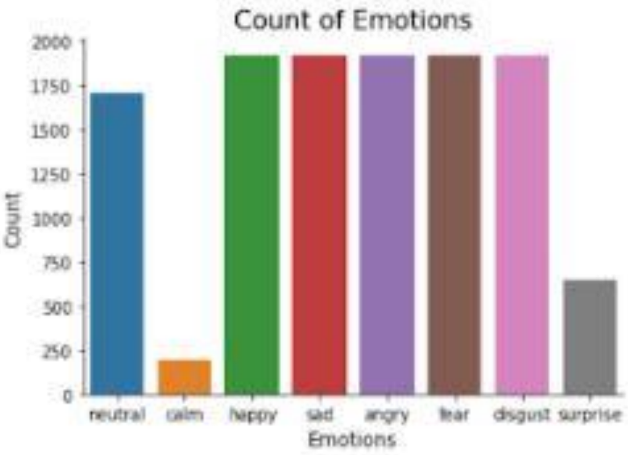
In this Speech Emotion Recognition Project, Audio File is taken from the TESS Dataset, and that will be uploaded in .wav file format before the file upload process is validated, which relates to the file format and empty file input, and will be connect directly to python files where the output generated in the form of Emotional Labels. Data visualization provides information about the given audio data in redicament.

Graphic and pictorial form. Here, initial dataset is divided into its emotional labels, and then the whole data is pictured into spectrogram graph and wave plot diagram. Spectrogram may be a visual representation of the spectrum of frequencies of a sign because it varies with time. Wave-plot is employed to plot waveform of amplitude vs time where the primary axis is amplitude and the second axis is time.



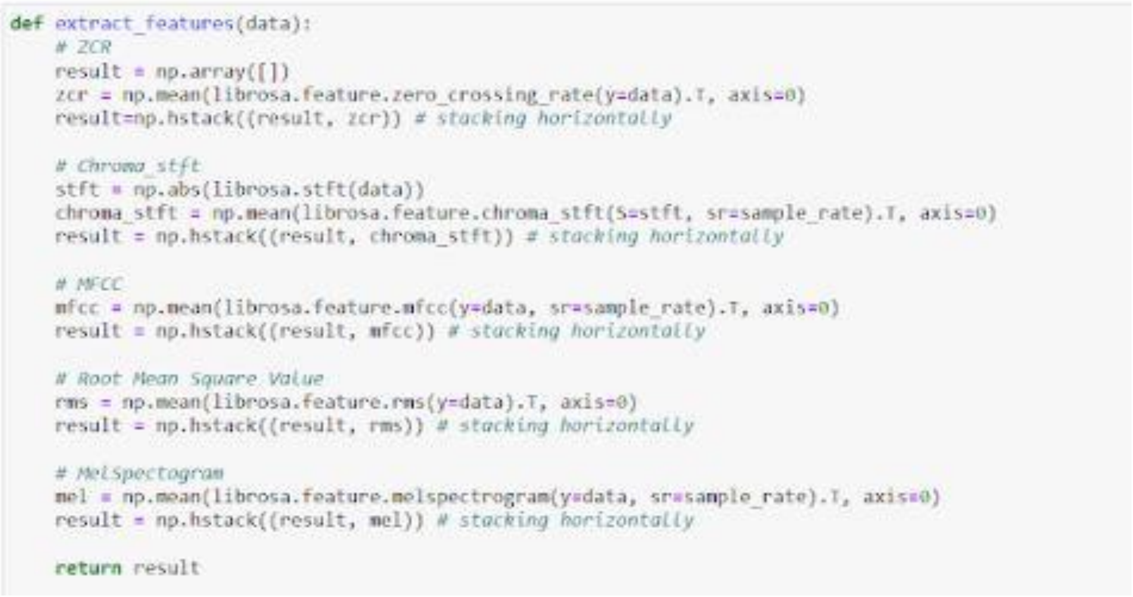
2.2 Features Selection:

In this Project mainly comprise features like Zero Crossing rate, Chroma Shift, Root Mean Square Value, Mel Spectrogram and MFCC (Mel information retrieval in the music category; these extract the crux of given audio. The zero- crossing rate is when a significant change from positive to 0 Frequency Cepstral Coefficient). These are few predominantly used audio features for emotional-related audio content, acoustic recognition, and to negative or from negative to 0 to positive. Mel Spectrograms are spectrograms that visual sounds on the Mel scale as against the frequency domain. The Mel Scale must be a logarithmic transformation of a signal's frequency.



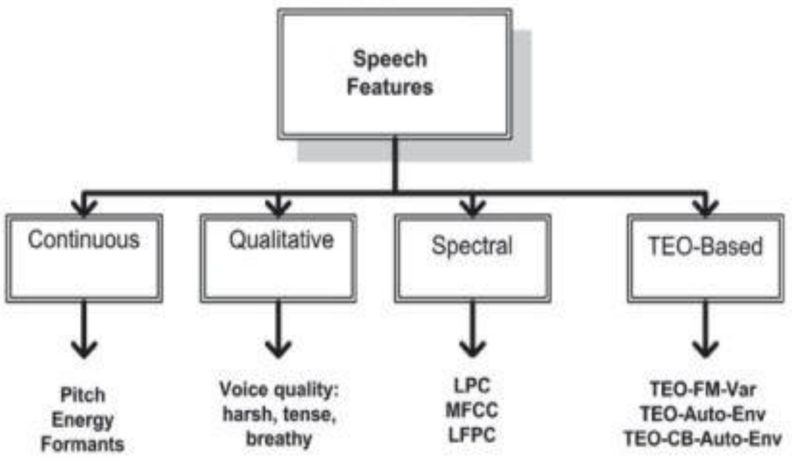
2.3 Feature extraction:

The speech signal contains a large number of parameters that reflect the emotional characteristics. One of the sticking points in emotion recognition is what features should be used. In recent research, many common features are extracted, such as energy, pitch, formant, and some spectrum features such as linear prediction coefficients (LPC), mel-frequency cepstrum coefficients (MFCC), and modulation spectral features. In this work, we have selected modulation spectral features and MFCC, to extract the emotional features.



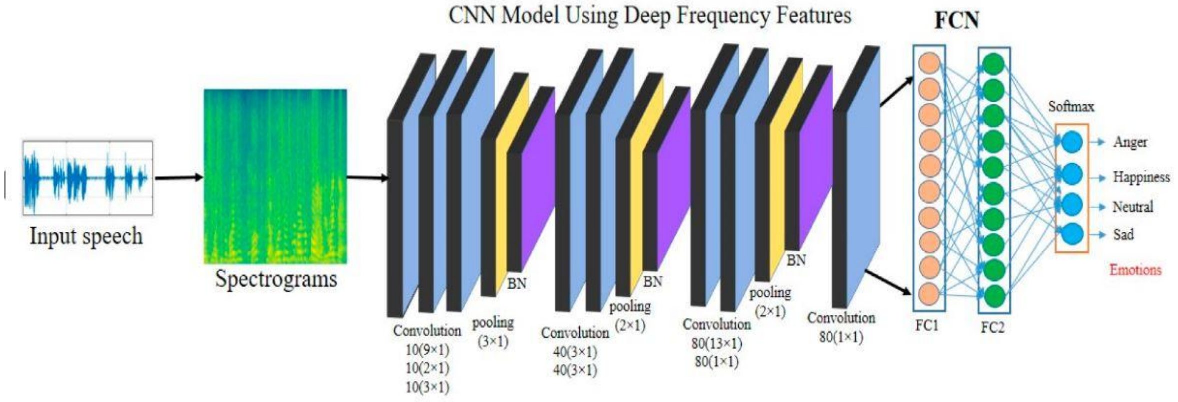
```
def extract_features(data):
    # ZCR
    result = np.array([])
    zcr = np.mean(librosa.feature.zero_crossing_rate(y=data).T, axis=0)
    result=np.hstack((result, zcr)) # stacking horizontally

    # Chroma_stft
    stft = np.abs(librosa.stft(data))
    chroma_stft = np.mean(librosa.feature.chroma_stft(S=stft, sr=sample_rate).T, axis=0)
    result = np.hstack((result, chroma_stft)) # stacking horizontally

    # MFCC
    mfcc = np.mean(librosa.feature.mfcc(y=data, sr=sample_rate).T, axis=0)
    result = np.hstack((result, mfcc)) # stacking horizontally

    # Root Mean Square Value
    rms = np.mean(librosa.feature.rms(y=data).T, axis=0)
    result = np.hstack((result, rms)) # stacking horizontally

    # MelSpectogram
    mel = np.mean(librosa.feature.melspectrogram(y=data, sr=sample_rate).T, axis=0)
    result = np.hstack((result, mel)) # stacking horizontally

    return result
```

Mel-frequency cepstrum coefficient (MFCC) is the most used representation of the spectral property of voice signals. These are the best for speech recognition as it takes human perception sensitivity with respect to frequencies into consideration. For each frame, the Fourier transform and the energy spectrum were estimated and mapped into the Mel-frequency scale. In our research, we extract the first 12 order of the MFCC coefficients where the speech signals are sampled at 16 KHz. For each order coefficients, we calculate the mean, variance, standard deviation, kurtosis, and skewness, and this is for the other all the frames of an utterance. Each MFCC feature vector is 60-dimensional.



3. CONVOLUTIONAL NEURAL NETWORK

Convolutional neural network (CNN) is used to train the speech data in the training set. Fig. 2 shows the accuracy of CNN model RNN model and MLP model training set and verification set. It can be seen from Figure 2 that with the increase of iteration times, the accuracy of training set and verification set tends to increase, especially the training set. Finally, after 200 iterations, the accuracy of the training set is over 97%, and that of the verification set is over 80%. CNN model developed in this paper has high accuracy than RNN and MLP in both training set and verification set.



The convolution operation represents the hierarchical extraction of speech features, while the maximum pooling operation removes redundant information in the previous layer features, and the operation is simplified. An activation layer is set after each convolution layer, and the best activation function determined by experiments. Two fully connected neurons are set in the output layer to divide the speech signals into two categories. In addition, considering the complexity of the network structure, random Dropout is set after each hidden layer to prevent the network from over-fitting in the training process. After adding Dropout, the neurons in each hidden layer will have a certain probability of not updating the weights in the training process, and the probability of each neuron not updating the weights is equal.

Algorithm

Step 1: The sample audio is provided as input.

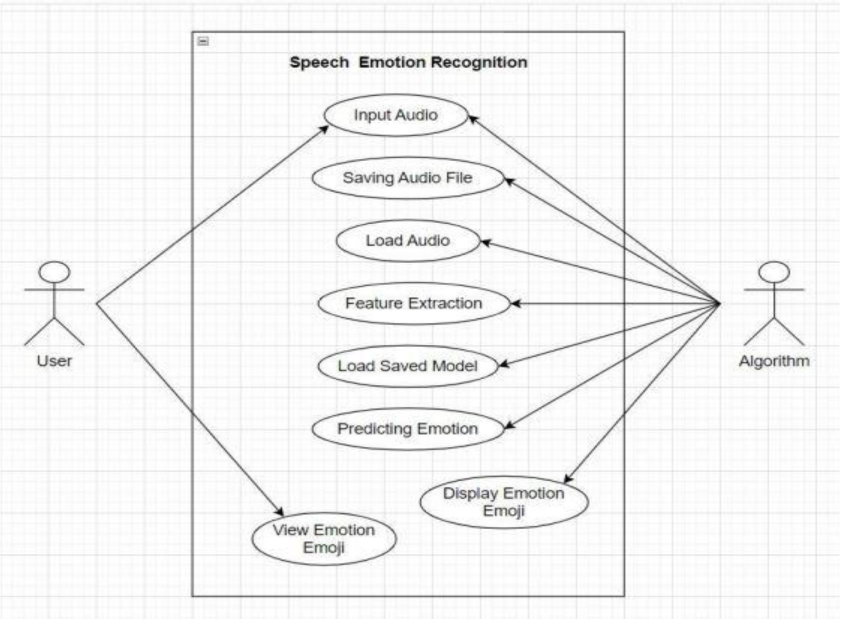
Step 2: The Spectrogram and Waveform is plotted from the audio file.

Step 3: Using the LIBROSA, a python library we extract the MFCC (Mel Frequency Cepstral Coefficient) usually about 10–20.

Step 4: Remixing the data, dividing it in train and test and there after constructing a CNN model and its following layers to train the dataset.

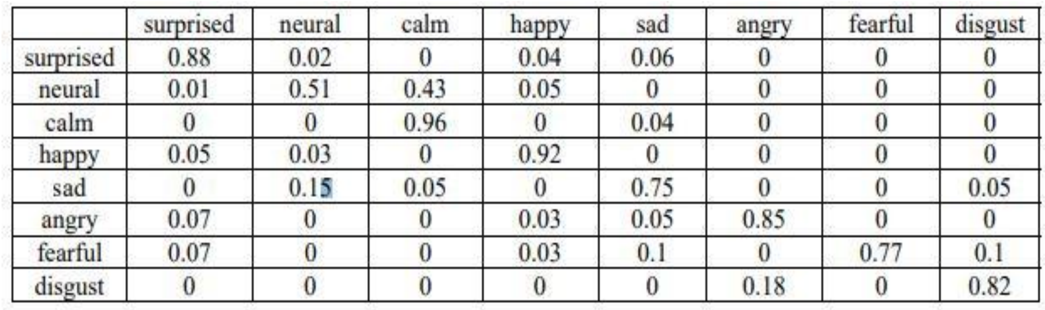
Step 5: Predicting the human voice emotion from that trained data (sample no.- predicted value - actual value)

UML DIAGRAM*:*

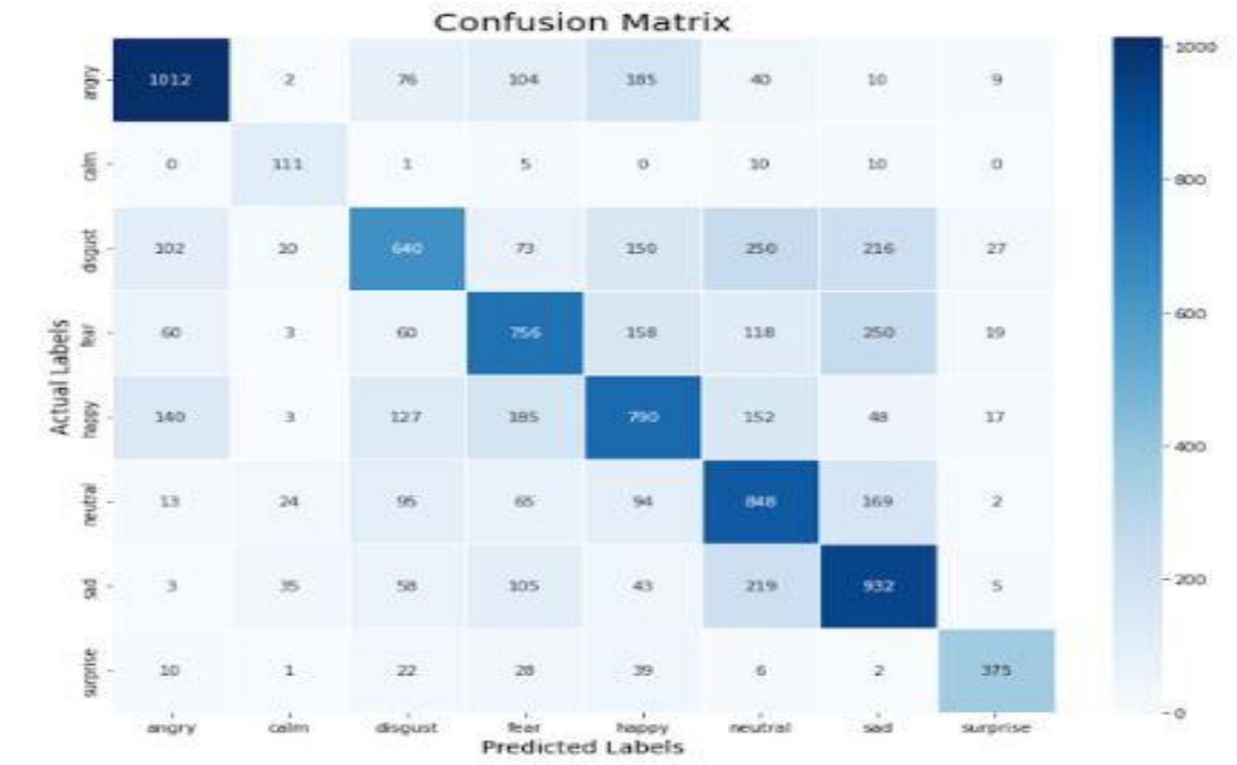


5. EXPERIMENTAL RESULTS AND ANALYSIS

In this paper, the obtained models are tested in the test set following the experimental procedure. Table 1 shows the normalized confusion matrix for SER. The recognition accuracy of expression recognition achieved by the proposed method on the RAVDESS dataset is better than existing studies on SER. The average classification accuracy on RAVDESS is listed in Table



	surprised	neural	calm	happy	sad	angry	fearful	disgust
surprised	0.88	0.02	0	0.04	0.06	0	0	0
neural	0.01	0.51	0.43	0.05	0	0	0	0
calm	0	0	0.96	0	0.04	0	0	0
happy	0.05	0.03	0	0.92	0	0	0	0
sad	0	0.15	0.05	0	0.75	0	0	0.05
angry	0.07	0	0	0.03	0.05	0.85	0	0
fearful	0.07	0	0	0.03	0.1	0	0.77	0.1
disgust	0	0	0	0	0	0.18	0	0.82



Some discriminations of 'sad' for ' neural ' fail. Perhaps surprisingly, the frequency of confusion between 'fear' and 'happiness' is as high as the frequency of confusion between 'sadness' or 'disgust'- perhaps this is because fear is a true multi-faceted emotion. Based on this, we will compare the characteristics of confounding emotions more carefully to see if there are any differences and how to capture them. For example, translating spoken words into text and training the network for multimodal prediction, and combining semantics for sentiment recognition.

6. CONCLUSION

In recent years, SER technology as one of the key technologies in human-computer interaction systems, has received a lot of attention from researchers at home and abroad for its ability to accurately recognize emotions and thus improve the quality of human-computer interaction. In this paper, we propose a deep learning algorithm with fused features for SER. in terms of data processing, we quadruple-processed the RAVDESS dataset with 5760 audio. For the network structure, we constructed two parallel convolutional neural networks (CNNs) to extract spatial features and a transform encoder network to extract temporal features to classify emotions from one of the eight categories. The TESS dataset that we considered is fine-tuned. Since it has noiseless data, it was easy for us to classify and feature .Taking advantage of CNNs in spatial feature representation and sequence coding transformation, we obtained an accuracy of 80.46% on the holdout test set of the RAVDESS dataset. Based on the analysis of the results, the recognition of emotions by converting speech into text combined with semantics is considered.

The combined Spectrogram-MFCC model results in an overall emotion detection accuracy of 73.1%, an almost 4% improvement to the existing state-of-the-art methods. Better results are observed when speech features are used along with speech transcriptions. The combined Spectrogram-Text model gives a class accuracy of 69.5% and an overall accuracy of 75.1% whereas the combined MFCC-Text model also gives a class accuracy of 69.5% but an overall accuracy of 76.1%, a 5.6% and an almost 7% improvement over current benchmarks respectively. The proposed models can be used for emotion-related applications such as conversational, social robots, etc. where identifying emotion and sentiment hidden in speech may play a role in the better conversation.

REFERENCES

1. Surabhi V, Saurabh M. Speech emotion recognition: A review. International Research Journal of Engineering and Technology (IRJET). 2016;03:313-316

2. Nicholson, J., Takahashi, K. &Nakatsu, R. Emotion Recognition in Speech Using Neural Networks. NCA 9, 290–296(2000). https://doi.org/10.1007/s005210070006.

3. ‘Han, Kun / Yu, Dong / Tashev, Ivan (2014): “Speech emotion recognition using deep neural network ”

4. A. Rajasekhar, M. K. Hota, ―A Study of Speech, Speaker and Emotion Recognition using Mel Frequency Cepstrum Coefficients and Support Vector Machines||, International Conference on Communication and Signal Processing, pp. 0114-0118, 2018.

5. Zheng, W. L., Zhu, J., Peng, Y.: EEG-based emotion classification using deep belief networks. In: IEEE International Conference on Multimedia & Expo, pp. 1-6 (2014).

6. Parthasarathy S, Tashev I.Convolutional neural network techniques for speech emotion recognition. In: 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC). IEEE 2018. pp. 121-125.

7. Matilda S. Emotion recognition: A survey. International Journal of Advanced Computer Research. 2015;3(1):14-19

8. Lim W, Jang D, Lee T. Speech emotion recognition using convolutional and recurrent neural networks. Asia- Pacific. 2017:1-4

9. G. Liu, W. He, 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, 2018.

10. Koolagudi SG, Rao KS. Emotion recognition from speech: A review. International Journal of Speech Technology 2012.