SPEECH EMOTION RECOGNITION USING DEEP LEARNING

***Abstract*—Speech emotion recognition is one in all the foremost arduous fields in data science. In Speech Emotion Recognition (SER), emotional aspects frequently emerge in various varieties of energy designs in spectrograms. This project presents the employment Neural networks to acknowledge the emotions from a identified speech, named as Speech Emotion Recognition (SER). In this project, we experimented with the Convolutional Neural Network to recognize emotion from the audio data files. Furthermore, have conjointly scrutinized the efficacy of transfer learning for emotion recognition employing a pre-trained VGG16 model. The audio features have been extracted using MFCC and Mel-Spectrogram of inputted audio datafiles. Project implementation is carried out on Ryerson Audio-Visual Database of Emotion Speech and Song dataset, named as RAVDESS dataset.**

Keywords—speech emotion recognition, convolutional Neural Network, transfer learning, MFCC, Spectrogram.

I. INTRODUCTION

EMOTION recognition plays a crucial part to interlink communication between computers and humans. With the evolution of the artificial intelligence era, it has become quite feasible to recognize human emotions through speech. Speech is a crucial carrier of human interaction, and it makes sense to recognize emotions from speech for a wide range of applications that are emotions in voice chatbots, smart speakers like alexa, google assistant to play music according to current mood through recognizing emotions of users. Deep learning has advanced the contribution for recognizing emotions through human speech, but there are still inadequacies in the research of emotion recognition through speech. Despite the ongoing improvements in this field of speech recognition, there is still a considerable need for progress and innovation in this field to make interaction between humans and machines a more easy and organic process.

Speech emotion recognition refers to survey of vocal behavior as an indicator of physiological changes focusing on the nonverbal outlook of the speech. Its basic presumption is that human speech expression is a collection of various measurable parameters that mirrors affective occurrences in speech, human presently expressing. This presumption is carried by the fact that voice produced is modified by numerous physiological responses. For example, the emotion anger often produces variations in respiration and enlarge muscle tension, influencing the vibration of the vocal folds and vocal tract appearance and affecting the acoustic characteristics of the speech.

SER workflow often carried out by initially extracting the necessary features from the audio speech followed by classification of specific emotions through speech recognition. There are various difficulties and hurdles observed by researchers which incorporate selection of suitable speech features, vigor to tone variations, speaking patterns, and how various emotions are conveyed in different circumstances and situations. One of the crucial and complex barriers is the feature extraction of prejudiced, robust audio speech. Features used for emotion recognition systems are basically categorized into acoustic, linguistic, and hybrid features.

In contrast, Automatic Speech Recognition(ASR), recent SER research faces the problem of lacking high-quality data. Generally stating, it is high-priced to invite professionals to stage emotions in the studio and it is easier to collect noisy data practically. On the other hand, emotions are subjectively governed by various factors, language and culture have an crucial influence on the judgment of emotions in speech, which improves the cost of data labeling.

The organization of this report is as follows. Section II describes some existing notable related work in this field. Section III describes the Proposed Methodology of the project. Section IV describes the proposed experimental results analysis of CNN and VGG16 model in the project. Section V describes the conclusion of project work for SER.

II. RELATED WORK

Before the era of deep learning, for SER, researchers principally used complicated standard options and standard machine learning techniques and strategies. Deep learning accelerated the progress ofSER research. K. Hanet al. first applied deep learning to SER in 2014. Recently, for the same reason, M. Chen et al. merged convolutional neural networks(CNN) and Long Short-Term Memory (LSTM); X. Wu Et al. restored CNN with capsule networks (CapsNet); Y. Xuetal. used Gated Recurrent Unit (GRU) to reckon

options from frame and auditory communication level, and S. Parthasarathy used ladder networks to mix the unsupervised auxiliary jobs and additionally the initial task of predicting emotional attributes.

Noroozi et al. proposed a flexible emotion recognition system that supported the analysis of visual and auditory signals. In his study, the feature extraction stage used 88 features(Mel Frequency Cepstral Coefficients (MFCC), filter bank energies (FBEs))victimization the Principal element Analysis (PCA) in feature extraction to reduce the dimension of options antecedently extracted. Bandela et al. used the fusion of acoustic feature which is the MFCC with Teager Energy Operator (TEO) as a prosodic to identify five emotions by the GMM classifier using the Berlin Emotional Speech database. Zamil et al. also used the spectral characteristics which are the 13 MFCC obtained from audio data in their proposed system to classify the 7 emotions with the Logistic Model Tree (LMT) algorithm Hadhami Aouani et al. / Procedia Computer Science 176 (2020) 251–260253Hadhami Aouani et al / Procedia Computer Science 00 (2020) 000–0003with an accuracy rate 70%.

III. Methodology

In our speech our voice pitch and tone majorly reflects the emotion in our speech. This project aims to recognize the elicit types of emotions such as sad, happy, neutral, angry, disgust, surprised, fearful, and calm. In this project, the emotions in speech are predicted using convolutional neural networks and proposed methodology is discussed in the subsequent sections.

A. Spectrograms

A spectrogram is the visual depiction of power of speech signal at various frequencies over time in a specific waveform. It is depicted by a two-dimensional graph in which time is depicted along the X-axis and frequency along the Y-axis. The various frequency amplitude modules are demonstrated by a range of colors displacements or intensity at a specific time in the graph. Low amplitudes are shown by darker shades of color up through dark blues and noisy amplitudes are shown by the brighter shade of colors up through red determined by adding Fast Fourier Transform to depict the time-frequency plot of the speech signal. In order to uncover the different frequencies present at a specific time in the speech signal, they are splitted into numerous tiny blocks on which Fast Fourier Transform is applied to the signal wave plot. Some of the sample spectrograms are shown in below:

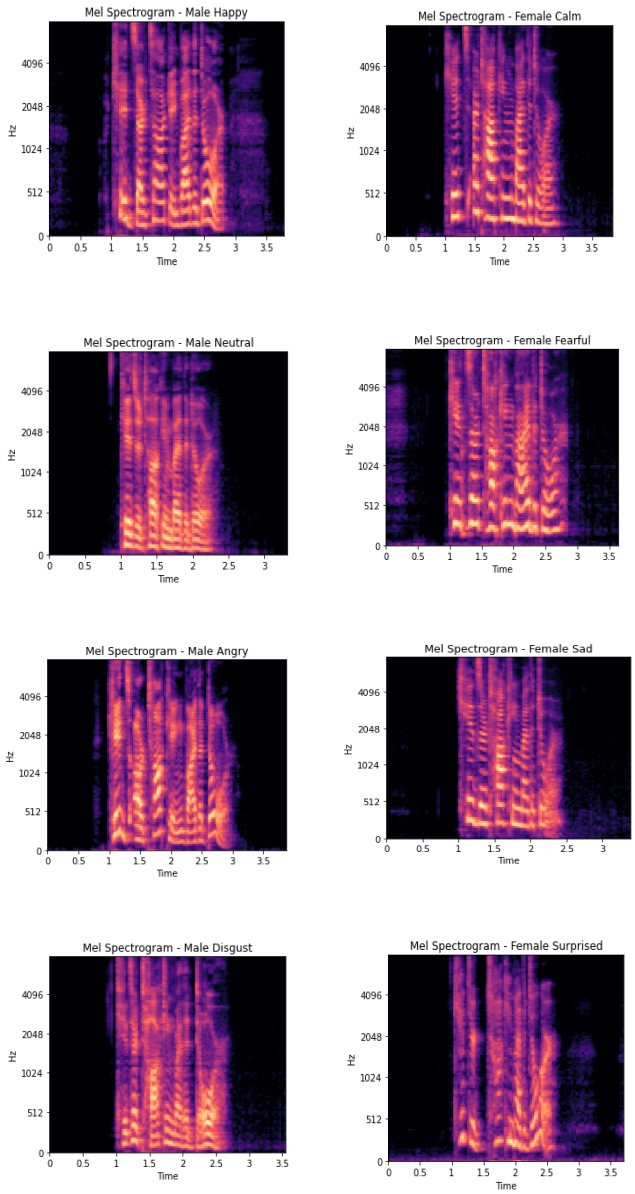

Fig. 1 Sample speech emotion spectrograms

B. Spectrum

Spectrum is a representation of a sound, generally a small sample of a sound representing the fraction of each individual frequency's vibration. It refers to the invisible radio frequencies that wireless signals advance over. It is usually presented as a graph of the power as a function of frequency. The power or pressure is commonly measured in terms of decibels and the frequency is measured in vibrations per second in Hertz.

Sound intensity variations can be measured and can plot these measurements over the time. Then these raw waveplot after applying fast fourier transform can be used to generate power spectrum. Some of the sample emotion waveplot are shown in below:

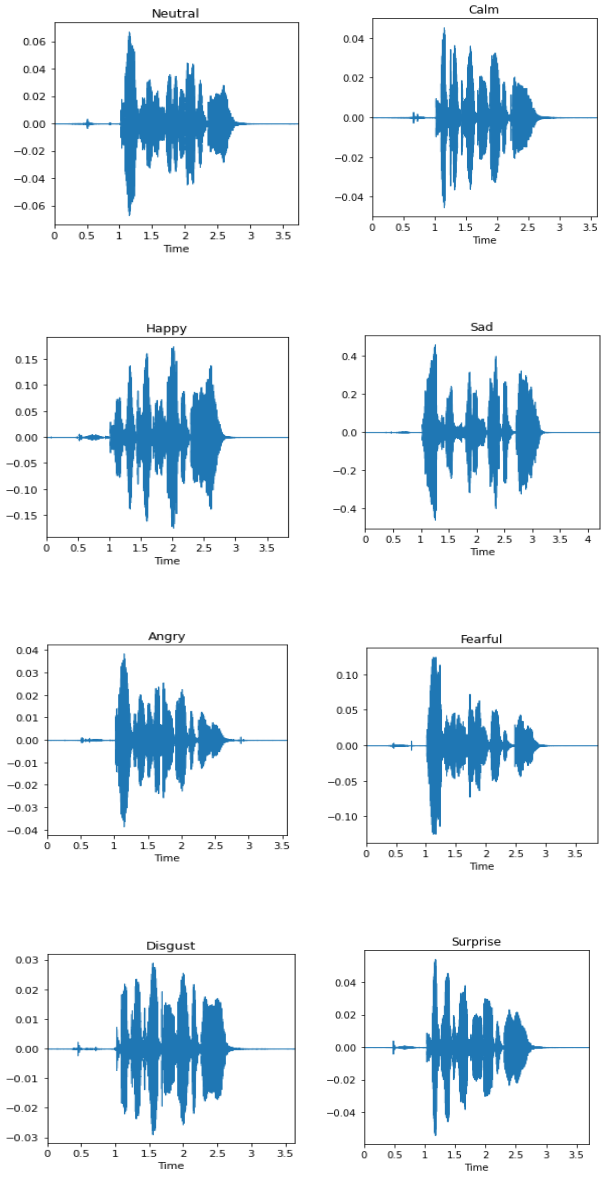

Fig2. Sample emotion wave plots

C. Mel-frequency Cepstral Coefficient

Mel Frequency Cepstral Coefficients (MFCC) is employed to recover the sound from the given wave audio file by utilising distinct hop length and HTK-styles mel frequencies. Pitch of 1 kHz tone and 40 dB over the perceptual visible edge is characterised as 1000 mels, employed as reference point. The MFCC gives a Discrete Cosine Change (DCT) of an original logarithm of the transient vitality depicted on the Mel recurrence scale.

To compute the MFCC coefficients, the inverse Fast Fourier Transform (IFFT) is applied to the logarithm of the Fast Fourier Transform (FFT) module of the signal, filtered according to the Mel scale.

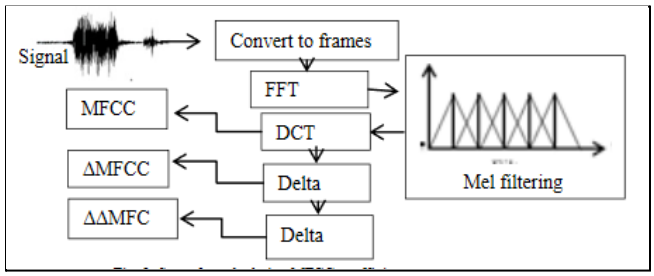

Fig. 3 MFCC coefficient architecture

D. Speech Emotion Recognition

Speech Emotion Recognition is a research area field that states to improve emotion detection making a lot of systems effortless and better by gathering emotions from the speech signal. But it is a challenging task so involves a lack of speech data in many other languages, speech corpus is not sufficient to accurately output emotions and each individual facial expression leads to ambiguous discovery.

Speech Recognition is the technology that trades with techniques and methodologies to recognize speech from speech signals. With numerous technological advancements in the field of artificial intelligence and signal processing techniques, the recognition of emotion is made easier and possible. There are many voice products that have been invented like Amazon Alex, Google Home, Apple HomePod which functions mostly on voice-based commands. It is evident that Voice will be the better channel for communicating with the machines.

E. Convolutional Neural Networks

A convolutional neural network is a hierarchical neural network comprised of assorted layers in sequence. A quintessential model generally consists of various convolutional layers as a collection of features acquired from visual contents of a spectrogram with a huge variety of features that are grasped during the model training part.

The proposed CNN model architecture is designed with

three convolutional layers followed by two fully connected dense layers and a dropout layer. Each of these convolutional layers is followed by ReLU units and lastly by softmax units. In order to avoid overfitting, the initial two fully connected layers are followed by dropout layers having a dropout of 0.4 units.

F. Transfer Learning

Transfer learning a deep learning technique that permits developers to harness a neural network problem using the experience acquired from solving one problem to unravel another problem. It's evident that Transfer learning solves many problems within a brief interval of your time. Transfer Learning is incorporated whenever there's a need to cut back computation cost, achieve accuracy with less training. Transfer learning is best applied where the source task's model trained on huge training data and acquired knowledge to solve the problem of destination tasks.

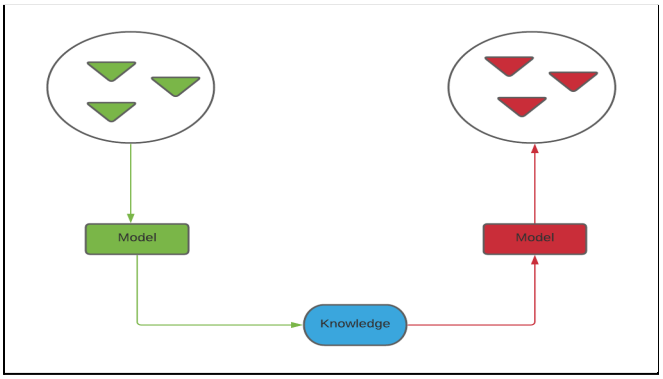

Fig. 4 Transfer Learning Construct

G. VGG16 Model

VGG-16 a convolutional neural network that comprises 16 layers. The model comprises a group of weights pre-trained on ImageNet. VGG-16 a deep convolutional neural network that incorporates 16 layers that are combined by many 3×3 convolutional layers and 2×2 pooling layers repeatedly, and VGG-16 incorporates a remarkable feature extraction capability so it can obtain a decent effect in image classification. VGG-16 just uses a 3×3 convolution layer and 2×2 pooling layer repeatedly. In this project using the VGG16 model with a pretrained weight of "imagenet" and an input shape of 224 x 224. Then on this architecture added the custom-made three dense layers with a dropout of 0.2 units and used as a feature extractor.

IV. Experimental Results and Analysis

A. Dataset Description

The dataset taken for the implementation task is the RAVDESS dataset which is composed of 24 neutral North American accent professional actors, with 12 female and 12 male actors uttering two lexically equaled sentences. The emotions accommodated within the dataset are as sad, happy, neutral, angry, disgust, surprised, fearful, and calm emotions. Since our aim is to recognize emotions from speech, our model is trained on Audio data. Two fixed statements repeated twice are vocalized by all the 24 actors for all 8 emotions, with each statement. Two statements are "Kids are talking by the door" and "Dogs are sitting by the door".

B. Pre-processing

The dataset used utterance and randomly divided speech emotion data into an 80% train data set and a 20% test data set. Each data file has an 80% probability of being picked into the training set and a 20% probability of being picked into the test set and both are mutually exclusive. Since the division of the information set is random and also the number of utterances is large enough, each subset has an identical distribution because of the original data set. Then after this normalized the data and converted it to arrays for Keras and applied one-hot encoding and lastly reshape the data.

C. Implementation

Using the Convolutional neural networks to recognize the emotions from the fed audio dataset. Then using the extracted features using MFCC and Mel-Spectrogram and passing it all together to get a considerable deviation of the prediction emotion, as a single featured parameter is not sufficient to come up with a well-organized prediction. The RAVDESS dataset is made to pass to the neural networks, split the dataset into 80:20 ratio that is the training and testing dataset. The dataset consisted of samples of audio by 24 professional actors with North American accents. Eight types of emotions are used. The fig.5 below shows the workflow architecture:

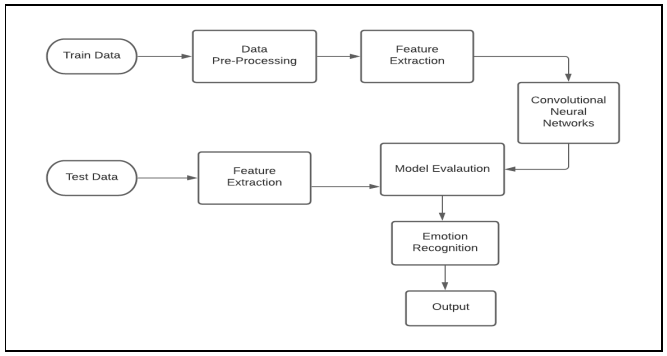

Fig. 5 Workflow Architecture

1. Convolutional Neural Network

The proposed CNN model images from the spectrogram were generated from the RAVDESS dataset and resized to 224 x 224. Eighty percent of this data was used for training purposes and the remaining was used in the testing part. Initially, the training process was run for 15 epochs with a batch size set to 32. After 13 epochs accuracy of 41.67 % was achieved. Then applied the performance tuning increased batch size to 64 and used SGD optimizer with a momentum of 0.9. And ran for 50 epochs then achieved an accuracy of 69.10 %. The predictions from the CNN model are gathered to decide the probabilities for independent emotions by computing average predictions from the collected evidence. Below FIg.6 shows the accuracy and loss plot:

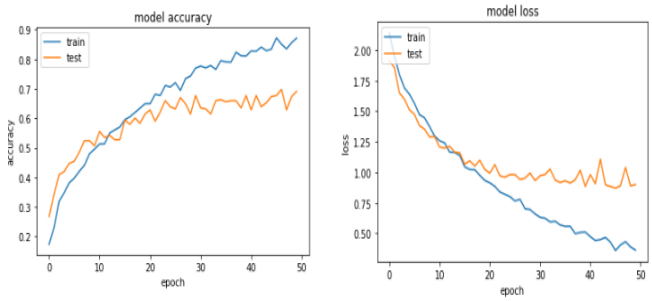

Fig 6. Accuracy and Loss graph plot.

	Precision	Recall	F1-score
Angry	0.82	0.72	0.77
Calm	0.75	0.79	0.77
Disgust	0.71	0.73	0.72
Fearful	0.72	0.55	0.63
Happy	0.63	0.83	0.72
Neutral	0.65	0.54	0.59
Sad	0.47	0.39	0.43
Surprise	0.73	0.92	0.82

Table 1. Classification Report for CNN model

2. VGG16 Model

The proposed pretrained VGG16 model initially trained with pretrained "imagenet" weights using transfer learning. The training process was trained for 10 epochs with a batch size put to 32. And accuracy of 52.31 % and performed slightly better than the previous model. Then applied performance tuning used Mel-Spectrogram feature extractor on the entire audio data files. Below is some sample of images from the training set:

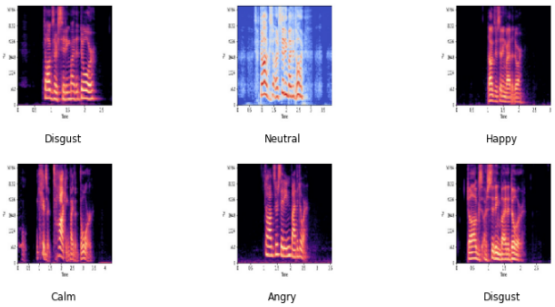

Fig7: Sample training images

Then applied augmentation using the VGG-16 as a feature extractor with Image Augmentation. Applied augmentation on spectrograms generated earlier to add more training data to deal with small audio data files. Then used ImageDataGenerator to augment the spectrogram images generated earlier. Lastly, run the model on 80 epochs and achieved an accuracy of 81.25%. And it was observed that the performance-tuned model improves the robustness of the model and improves the prediction performance to recognize the emotions. Below are accuracy and loss graph plot:

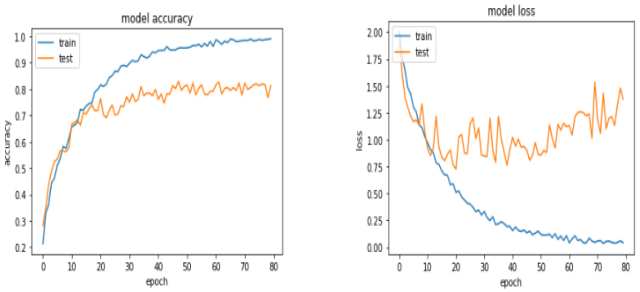

Fig 8 Accuracy and Loss graph plot.

	Precision	Recall	F1-score
Angry	0.94	0.77	0.85
Calm	0.85	0.92	0.89
Disgust	0.78	0.86	0.82
Fearful	0.82	0.74	0.78
Happy	0.85	0.80	0.82
Neutral	0.76	0.79	0.78
Sad	0.63	0.76	0.69
Surprise	0.92	0.85	0.88

Table 2. Classification Report for VGG16 model

V. CONCLUSION

Speech emotion recognition is a complex task, which involves two essential problems: the first feature extraction and the other one emotion classification. This project proposed the framework for speech emotion recognition using a convolution neural network and pre-trained VGG 16 models with the eight different emotions of the RAVDESS dataset. The speech signal is represented as spectrograms using the Mel-spectrogram feature extractor which acts as the input to the neural network and output predictions for the eight different emotion classes.

Future work is required to enhance the overall proposed framework so that all emotions of the speech can be recognized in a constructively and robust manner. Also, outline to take more data with comparably complex models to tweak the speech emotion recognition performance even further.

REFERENCES

[1] K. Han, D. Yu, and I. Tashev, “Speech emotion recognition using deep neural network and extreme learning machine,” in the 15 yearly conferences of the international speech communication association, Singapore, 2014.

[2] M. Chen, X. He, J. Yang, and H. Zhang, “3-d Convolutional recurrent neural-networks with attention model for speech emotion recognition,” IEEE Signal Processing Letters, vol. 25, no. 10, pp. 1440–1444, 2018.

[3] X. Wu, S. Liu, Y. Cao, X. Li, J. Yu, D. Dai, X. Ma, S. Hu, Z. Wu,X. Liu et al., “Speech emotion recognition using ‘capsule networks,” inICASSP 2019-2019 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).Brighton, United Kingdom: IEEE,2019, pp. 6695–6699.

[4] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features-linguistic information in a hybrid support vector machine-belief network architecture," in Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on, 2004, pp. I-577-80 vol. 1.

[5] D. J. France, R. G. Schiavi, S. Silverman, M. Silverman, and M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," IEEE transactions on Biomedical Engineering, vol. 47, pp. 829-837, 2000.

[6] F. Dipl and T. Vogt, “Real-time automatic emotion recognition from speech,” 2010.

[7] M. Xu, F. Zhang, and S. U. Khan, “Improve accuracy of speech emotion recognition with attention head fusion,” in2020 10th Annual Computing and Communication Workshop and Conference (CCWC). Las Vegas, NV USA: IEEE, 2020, pp. 1058–1064.

[8] A. Chatterjee, U. Gupta, M. K. Chinnakotla, R. Srikanth, M. Galley, and. Agrawal, “Understanding emotions in text using deep learning and big data,” Computers in Human Behavior, vol. 93, pp. 309–317, 2019.

[1] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features-linguistic information in a hybrid support vector machine-belief network architecture," in Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on, 2004, pp. I-577-80 vol. 1.

[1] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features-linguistic information in a hybrid support vector machine-belief network architecture," in Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on, 2004, pp. I-577-80 vol. 1.

[2] D. J. France, R. G. Schiavi, S. Silverman, M. Silverman, and M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," IEEE transactions on Biomedical Engineering, vol. 47, pp. 829-837, 2000.

[3] F. Dipl and T. Vogt, “Real-time automatic emotion recognition from speech,” 2010.

[4] M. Xu, F. Zhang, and S. U. Khan, “Improve accuracy of speech emotion recognition with attention head fusion,” in2020 10th Annual Computing and Communication Workshop and Conference (CCWC). Las Vegas, NV USA: IEEE, 2020, pp. 1058–1064.

[5] A. Chatterjee, U. Gupta, M. K. Chinnakotla, R. Srikanth, M. Galley, and. Agrawal, “Understanding emotions in text using deep learning and big data,” Computers in Human Behavior, vol. 93, pp. 309–317,