Speech Emotion Recognition

Srijha Kalyan, Ketaki Kolhatkar
kalyan.sr@northeastern.edu, kolhatkar.k@northeastern.edu

***Abstract*—Emotions play a significant role in human mental life. It is a medium of expression of one's perspective or one's mental state to others. This is essential to our rational as well as intelligent decisions. Automatic Speech Emotion recognition has become a promising area of study in the field of human-computer interaction. The following paper provides an overview of Speech Emotion Recognition and related work. A comparison study of classifying emotions from speech using various deep learning models (CNN, CNN-Transfer Learning Model, LSTM) is also presented. The datasets used were: RAVDESS, SAVEE, TESS.**

***Index Terms*—Speech Emotion Recognition(SER), Convolutional Neural Networks(CNNs), LSTMs(Long Short Term Memory)**

I. INTRODUCTION

In this rapidly advancing AI world, human computer interactions (HCI) are of extreme importance. We live in a world where Siri and Alexa are physically closer to us than other humans. Soon the world will get more populated with physical and virtual service robots to accomplish tasks that range from caring for the elderly to assessing the effectiveness of your marketing campaign. Understanding human emotions paves the way to understanding people's needs better and, ultimately, providing better service. Speech Emotion Recognition as the name suggests is identifying a person's emotion using their speech. Recognizing a person's emotion can be based on a number of factors like the words they use, facial expressions, body language, etc. With the advancement of human computer interaction and artificial intelligence, a machine can be trained to do all of the above tasks. In our project, we focus on the speech part of emotion detection. The motive behind making machines understand human emotions is to improve customer experience in every sector. Emotion being a very subjective term most datasets have been categorized into seven emotions namely anger, disgust, fear, happiness, sadness, surprise and neutral. Although there are so many emotions, the datasets have only these seven emotions since many of the emotions lie very close to one another and it is hard to distinguish between emotions many a times. Hence we try to represent emotions on 3 dimensions. The 3 dimensions across which these emotions are represented are valence, activation and dominance. Valence ranges from negative to positive, activation ranges from low to high and the dominance ranges from dominated to dominant. These on a graph are from -1 to +1. The emotions are then plotted on this graph based on the 3 factors. Neutral comes right in the middle of the graph, while anger and disgust lie close due to the similarity in their features. Joy and happiness lie close by in the same way. Apart from the acoustic features, the lexical features matter in identifying emotions, since understanding the context of a sentence is important in interpreting emotions. Speech also can be used to retrieve the age and gender of the person interacting with the machine, so it can give more relevant results to queries. Although, the scope of our paper would be limited to identifying the emotions of a person based on the speech.

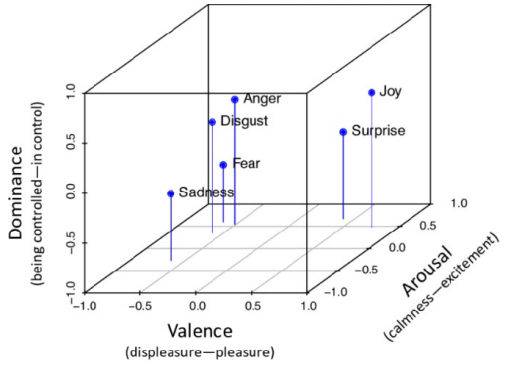

Fig. 1. Emotion Representation

II. RELATED METHODS

The following section provides a literature review related to the development of various deep learning models used to detect emotions from speech. Due to the importance of SER in human–computer interaction and the development of artificial intelligence systems, there are multiple other recent publications and surveys on SER. In "DeepEMO: Deep Learning for Speech Emotion Recognition" by Enkhtogtokh Togootogtokh and Christian Klasen, the proposed speech recognition framework, called DeepEMO, consists of two main pipelines such that preprocessing to extract efficient main features and deep transfer learning model to train and recognize.[1]

The development of the models was based on the methods discussed in "Speech Emotion Recognition Overview and Experimental Results" by Eva Lieskosva et.al. This paper provides a comparison of using recurrent neural networks such as LSTM, GRU and recognition accuracy on IEMOCAP database was also presented.[2]

Further development was done by using ideas from "Speech Emotion Recognition using Semantic Information" by Panagiotis Tzirakis, Anh Nguyen, Stefanos Zafeiriou1, Bjorn W. Schuller .The framework comprised of a semantic feature extractor, that captures the semantic information, and a paralinguistic feature extractor, that captures the paralinguistic information. Both semantic and paraliguistic features were then combined to a unified representation using a novel attention mechanism. The unified feature vector is passed through a

LSTM to capture the temporal dynamics in the signal, before the final prediction.[3]

In the paper "Speech emotion recognition using combination of features" by Qingli Zhang et al. speech features numbers and statistical values such as Mel Frequency Cepstrum Coefficients (MFCCs) and Auto Correlation Function Coefficients (ACFC) extracted directly from speech signal that impact recognition accuracy of emotions present in speech are studied using Gaussian Mixture Model (GMM) that provides an accuracy of 74.41%.[4]

The development of the models was based on the methods in the paper "Speech Emotion Recognition usingDeep Learning Techniques:" by Ruhul Amil Khalil et al. decribes methodologies of detecting emotions from speech using various deep learning techniques such as CNNs, LSTMs, Deep Belief Network etc. [5]

III. METHODOLOGY

A. Dataset

There are three main components to designing a SER: choosing an emotional speech database, feature selection from audio data, and the classifiers to detect emotion.

RAVDESS

The Ryerson Audio-Visual Database of Emotional Speech and Music (RAVDESS) dataset is a multi-modal database of emotional speech and song that has been validated. RAVDESS has 1440 speech files and 1012 song files. This dataset contains recordings of 24 professional actors (12 females, 12 males) speaking in a neutral North American accent and vocalizing two lexically-matched phrases. There are calm, happy, sad, angry, afraid, surprise, and disgust expressions in speech, and there are calm, happy, sad, angry, and fearful emotions in song. On emotional validity, intensity, and genuineness, each file was assessed ten times. 247 people who were typical of untrained adult study volunteers from North America supplied ratings. A total of 72 people participated in the test-retest study.

SAVEE

Surrey Audio-Visual Expressed Emotion (SAVEE) Four native English malespeakers (known as DC, JE, JK, KL), postgraduate students, and researchers at the University of Surrey, aged 27 to 31, contributed to the SAVEE database. Anger, contempt, fear, pleasure, sadness, and surprise are some of the psychologically distinct kinds of emotion.

TESS

A set of 200 target words were spoken in the carrier phrase "Say the word' by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). There are 2800 stimuli in total. Two actresses were recruited from the Toronto area. Both actresses speak English as their first language, are university educated, and have musical training. Audiometric testing indicated that both actresses have thresholds within the normal range.

B. Feature Extraction

To study and work on audio signals like speech, we have to first convert into a format that could be processed by the computer. Speech is a continuous wave. A continuous signal is difficult to be processed and worked on, hence we convert it into the digital form which is a sequence of discrete signals. Next we move to feature extraction from this digital audio. There are three levels of abstraction for feature extraction namely high level, mid level and low level. High level features are the ones that are can be understood by humans, like melody, rhthym, etc. Low level features like amplitude envelope, energy, spectral centroid, zero crossing rate are not understood by humans but easily understood by computers. Mid level features are partially understandable by humans. MFCCs, pitch and beat descriptors are a few examples. We will be working on these features in our project. Another feature we need to focus on is what domain the signal is represented. Although speech is commonly represented in the time domain, sound is characterized by frequency, hence we use the time frequency representation using spectograms, mel-spectograms, etc.

An ideal feature should have time frequency, amplitude representation that can be perceived by humans and perceptually relevant frequency representation. While vanilla spectogram satisfies the first two requirements, it can not represent frequency that is perceptual. Here is when we shift to mel-spectograms. These are based on the mel-scale which is logarithmic, which is coherent to how humans perceive frequency.

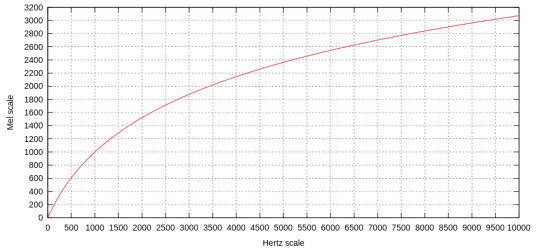

Fig. 2. Mel Scale

$$m = 2595log(1 + f/500) \tag{1}$$

$$f = 700(10^{m/2595} - 1) \tag{2}$$

The above two equations show how we convert frequency to the mel scale and vice versa. Using the mel frequncy we get, we also calculate the mel frequency cepstral coefficient, the MFCC as another feature for analyzing our speech signal. Cepstrum is a concept developed by scientists in the 1960s while studying echoes and has been used for speech recognition from the 1970s. Cepstrum is basically calculating the inverse fourier transform of the fourier transform of the time domain signal.

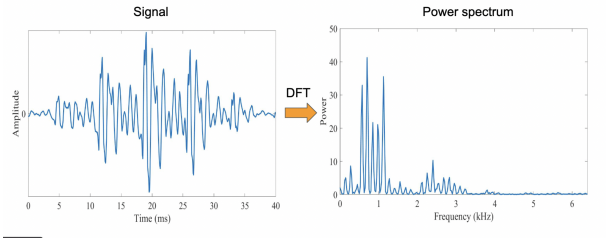

Fig. 3. Caption

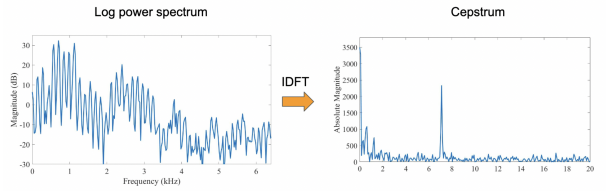

Fig. 4. Caption

$$C(x(t) = F'[log(F[x(t)])] \quad (3)$$

The coefficents we get from the mel filter banks is what we use to as our features in the project. MFCCs have proved to be a very promising feature in the field of speech recognition over the years. We use discrete cosine transform instead of the inverse fourier transform since it gives us the coefficients that we need. Here is the flow chart of the steps we follow

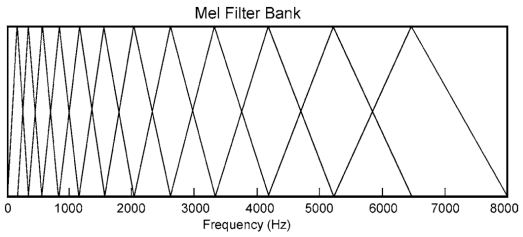

Fig. 5. Caption

for feature extraction. For implementing the feature extraction in our code, we have used the python library, 'librosa'. The sampling frequency is set to 22,050 Hz since that is the human frequency hearing limit. We have also tried extracting the low level features like pitch, etc for increasing our accuracy.

Zero-Crossing Rate:

It is basically the rate at which a signal changes from positive to zero to negative or from negative to zero to positive.

Spectral Centroid:

It is basically the center of 'gravity' of the spectrum. It is a measure used in digital signal processing to characterise a spectrum. It indicates where the center of mass of the spectrum is located.

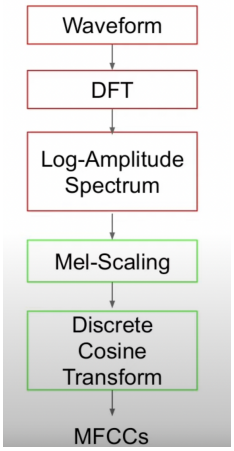

Fig. 6. Caption

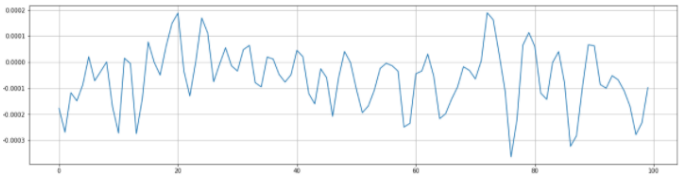

Fig. 7. Zero Crossing Rate

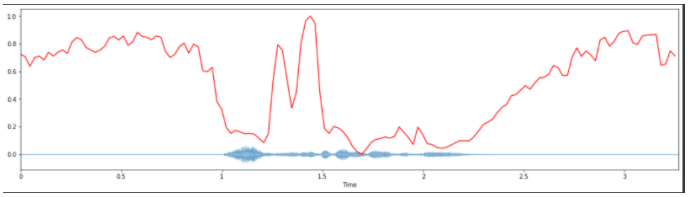

Fig. 8. Spectral Centroid

Spectral Contrast

Spectral contrast considers the spectral peak, the spectral valley, and their difference in each frequency subband. $librosa.feature.spectral.contrast$ computes the spectral contrast for six subbands for each time frame:

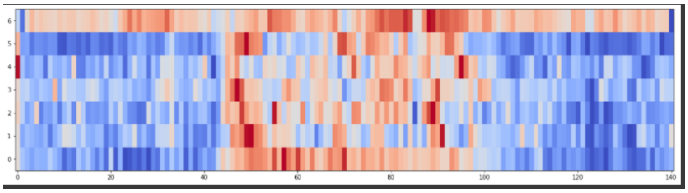

Fig. 9. Spectral Contrast

MFCC (Mel-Frequency Cepstral Coefficients)

In sound processing, it is a representation of the short term power spectrum of a sound based on a linear cosine transform of a log power spectrum on a non-linear mel scale of frequency. It provides us enough frequency channels to analyze the audio.

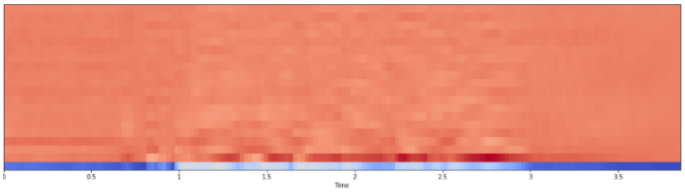

Fig. 10. MFCC

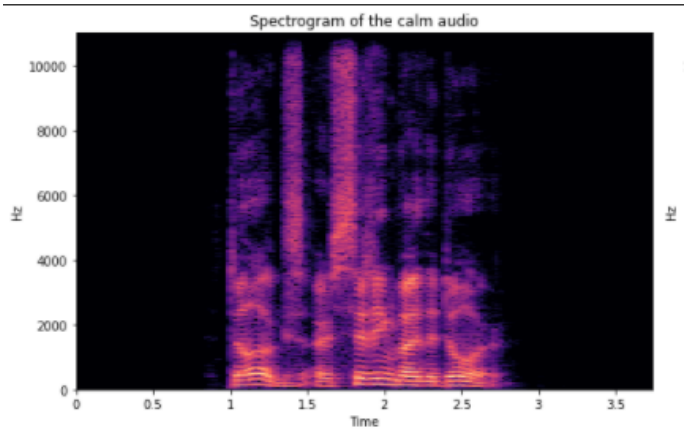

Fig. 11. Spectrogram

C. *Models*

We have implemented speech emotion recognition using various models to provide a comparison of the performance obtained. Intially, Decision trees, Support Vector Classifier has been used, which is one of the most commonly used model in the area of speech recognition. Vocal-based emotion recognition method applies the RF decision making algorithm to the speech signals comprising six emotion categories, namely, happiness, sad, disgust, fear, neutral and surprise. The corresponding emotion labels are then assigned to each voice signal by means of multi-class classification. The fact that the RF utilizes multiple randomly generated decision trees enables it to take advantage of all the virtues of decision trees, ensemble methods and bagging approaches. The RF classifier predicts the class label of an input data by majority voting on the predictions made by a set of tree classifiers. A portion of the data is kept for the test. The rest is used for training. Deep learning is known to perform better than machine learning techniques. Multi-layer perceptron has been implemented to detect various emotions along with Convolutional Neural Networks, LSTMs(Long short term memory) and Fine-tuned VGG16 Transfer learning technique.

Support Vector Classifier

SVM is a very simple and efficient classifying algorithm which is used for classification and pattern recognition. SVMs are systems that discriminate values based on a certain specification by using hyper planes in a high-dimensional feature space. Hyper planes are taught to apply statistical learning through the use of certain algorithms. The SVM classification approach is similar to supervised learning in that it incorporates feature extraction and produces desirable results. SVM has the advantage of being quite simple to learn. It is simple to predict emotional states using testing datasets once the training model has been created. The emotions are automatically categorized as Happy, Angry, Sad, or Fear using features taken from voice signals and SVM model values obtained by training models.

Random Forest Classifier

Random Forest is a tree-based machine learning technique that uses the power of numerous decision trees to make judgments. A forest of decision trees that have been constructed at random. In order to quantify the output, each node in the decision tree uses an arbitrary selection of features. The random forest then combines the number of distinct decision trees to obtain the final result.

Multi Layer Perceptron

A Multi-Layer Perceptron (MLP) is a perceptron-based network. It has an input layer that receives the input signal, an output layer that predicts or makes judgments for a given input, and a hidden layer that sits between the input and output layers. There can be a large number of hidden layers, and the number of hidden layers can be altered as needed. The input data is acted upon and processed by the hidden layer using an activation function. The logistic activation function was employed as the activation function. The output layer outputs the knowledge that the network has learned. According to the computation performed by the hidden layer, this layer classifies and outputs the expected emotion. To perform classification, the Multi-layer Perceptron Classifier (MLP Classifier) uses an underlying Neural Network. MLP Classifier uses Backpropagation to train the Neural Network using the Multi-Layer Perceptron (MLP) technique.

Building the MLP Classifier involves the following steps.

• Initialise the MLP Classifier by defining and initiating the required parameters.

• Data is given to the Neural Network to train it.

• The trained network is used to predict the output.

• Calculate the accuracy of the predictions. [h!]

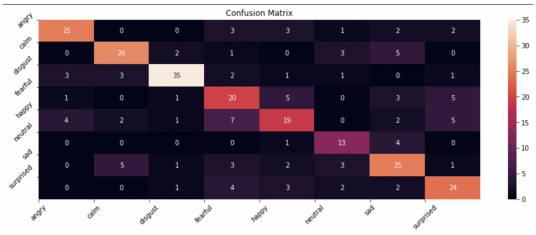

Fig. 12. Confusion Matrix for MLP

	precision	recall	f1-score	support
angry	0.76	0.69	0.72	36
calm	0.72	0.70	0.71	37
disgust	0.85	0.76	0.80	46
fearful	0.50	0.57	0.53	35
happy	0.56	0.47	0.51	40
neutral	0.57	0.72	0.63	18
sad	0.58	0.62	0.60	40
surprised	0.63	0.67	0.65	36
accuracy			0.65	288
macro avg	0.65	0.65	0.65	288
weighted avg	0.66	0.65	0.65	288

Fig. 13. Precision matrx for MLP

Convolutional Neural Networks

CNNs are good at automatically learning relevant features from high dimensional input images. CNNs use shared kernels (weights) to exploit the 2D correlated structure of image data. Max-pooling is added to CNNs to introduce invariance wherein only the relevant high dimensional features are learned for various tasks like classification, segmentation etc. Surprisingly, CNNs are good at audio recognition tasks. This is because, when Log mel filter bank is applied to FFT representation of raw audio , a ”linearity” is introduced in the log-frequency axis allowing us to perform convolution operation along the frequency axis. Else we would have to use different filters(or kernels) for different frequency ranges. This property along with CNN's superior representation power helps the model learn the underlying patterns effectively from short time-frames resulting in state of the art performance in Speech based Emotion Recognition systems.
We use a 1D CNN with the activation function ReLU, dropout of 0.4 and max pooling window size of 8. Next we give this to the flatten layer and then we give it to the dense layer with activation function - sigmoid. Here we use the Adam optimizer with a learning rate of 0.0001. We use categorical crossentropy for calculating the loss.

Long Short Team Memory - LSTMs

The RNN (Recurrent Neural Network) is a neural network in which the output of previous steps is supplied into the current phase as input. However, these recurrent neural networks have a practical disadvantage in that they can only look back a few steps. A system must be able to retain and utilise context information in order to understand difficulties like speech recognition. LSTM is a type of RNN. LSTM networks are neural networks that learn order relationships in sequence prediction tasks. Endless efforts have been made to correct the vanishing gradient problem, and LSTM is one such approach. In the time domain, the speech signal is continuous, so each frame function only provides the emotional features of a single frame. The information across neighboring frames is increased using LSTM, which helps to reflect the features' temporal continuity. As a result, LSTM clearly aids speech recognition.

TRANSFER LEARNING

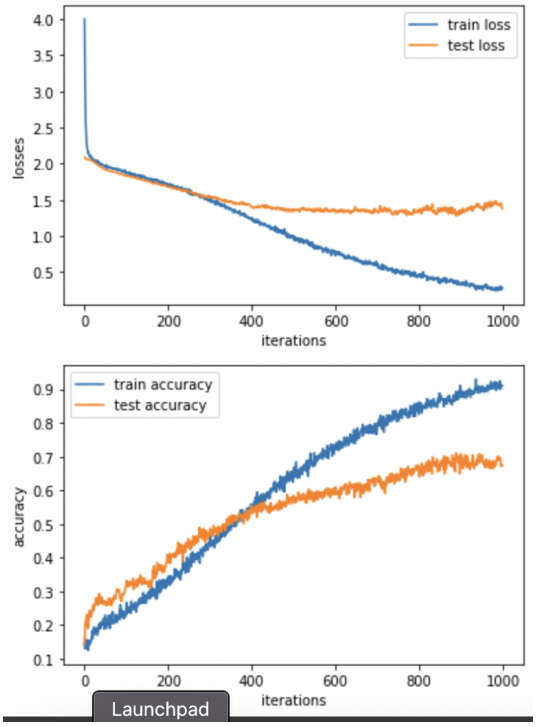

Fig. 14. 1. Losses over iterations, 2. Accuracy over iterations

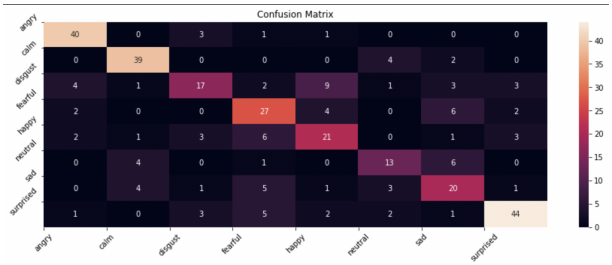

Fig. 15. Confusion matrix for CNN

	precision	recall	f1-score	support
angry	0.82	0.89	0.85	45
calm	0.80	0.87	0.83	45
disgust	0.63	0.42	0.51	40
fearful	0.57	0.66	0.61	41
happy	0.55	0.57	0.56	37
neutral	0.57	0.54	0.55	24
sad	0.51	0.57	0.54	35
surprised	0.83	0.76	0.79	58
accuracy			0.68	325
macro avg	0.66	0.66	0.66	325
weighted avg	0.68	0.68	0.68	325

Fig. 16. Precision matrix

The reuse of a pre-trained model on a new problem is known as transfer learning in machine learning. A machine uses the knowledge learned from a prior assignment to increase prediction about a new task in transfer learning. In this study, we re-purpose a model initially developed for speaker recognition to serve as a feature descriptor for SER. More specifically, first augment the signal data and perform an image augmentation we first train a VGG16 model on large amounts of speaker-labeled audio data. Then,

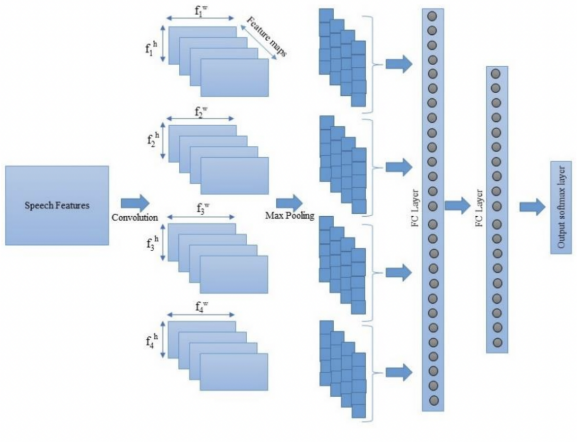

Fig. 17. CNN Architecture

we replace the FC layers of the pre-trained model with new randomly initialized FC layers. Finally, we re-train the new FC layers for an SER task on the RAVEDESS/SAVEE dataset.

VGG16-CNN Transfer Learning Model

VGG-16 network architecture is a sixteen (16) layer network used by the VGG group at the University of Oxford to obtain outstanding results in the ILSVRC competition held in 2014. The main feature of the vgg-16 network architecture was the increased depth of the network. It is known for its outstanding performance in several classification tasks like image, object and so on. It was trained on ImageNet dataset. It is considered to be one of the excellent vision model architecture till date. Most unique thing about VGG16 is that instead of having a large number of hyper-parameter they focused on having convolution layers of 3x3 filter with a stride 1 and always used same padding and maxpool layer of 2x2 filter of stride 2.

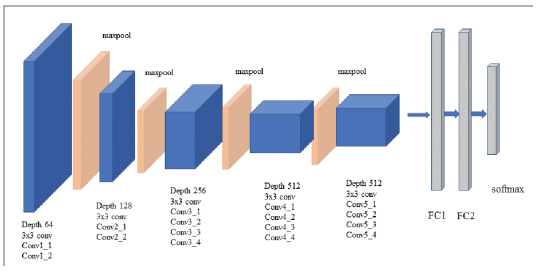

Fig. 18. VGG-16 Architecture

VGG16-Project Flow

To recognize speech emotion, we have proposed the deep convolutional transfer neural network using VGG16. The project flow involves in processing the audio input and performing EDA and data augmentation, then the emotion detection model is implemented by fine-tuning the CNN VGG16 model to obtain the necessary results. During EDA, two techniques are performed: time domain which shows how a speech signal changes over time and frequency domain extraction shows how much of the signal lies within each given frequency band over a range of frequencies. Further on, data augmentation is performed due to the lack of training data size, signal augmentation and image augmentation is performed to fine tune the VGG16 CNN model to identify emotions from speech.

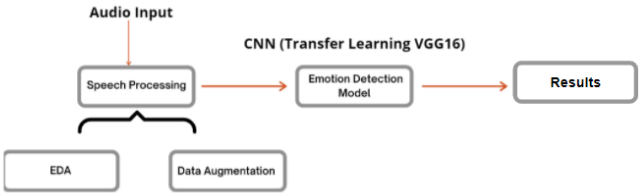

Fig. 19. VGG-16 Project Flow

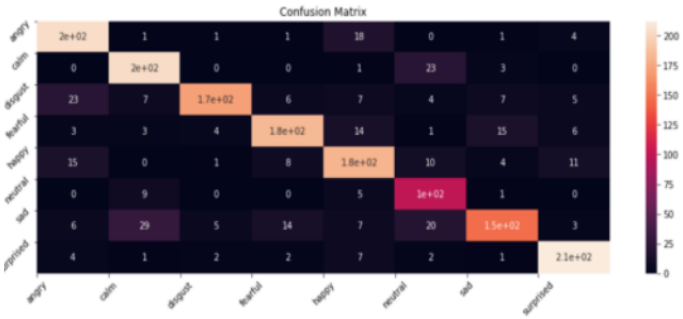

Fig. 20. Confusion Matrix VGG-16

	precision	recall	f1-score	support
angry	0.80	0.89	0.84	230
calm	0.80	0.88	0.84	230
disgust	0.93	0.74	0.83	230
fearful	0.86	0.80	0.83	230
happy	0.76	0.79	0.77	231
neutral	0.62	0.87	0.73	115
sad	0.82	0.64	0.72	231
surprised	0.88	0.92	0.90	231
accuracy			0.81	1728
macro avg	0.81	0.82	0.81	1728
weighted avg	0.82	0.81	0.81	1728

Fig. 21. Performance metrics VGG-16

IV. RESULTS

After experimenting with machine learning classifiers such as SVM, Decision Tree Classifiers, Random Forests and XGBClassifier. We obtained a lower performance score ranging from 46% to 58%. Neural networks have shown significant improvement in the speech recognition task. We implemented

```
Detected Text : what the hell
Predicted Emotion: angry
Classwise Scores
Class  angry  has a score of  0.91
Class  calm  has a score of  0.0
Class  disgust  has a score of  0.0
Class  fearful  has a score of  0.09
Class  happy  has a score of  0.0
Class  neutral  has a score of  0.0
Class  sad  has a score of  0.0
Class  surprised  has a score of  0.0
```

Fig. 22. Output of a test case using Fine-tuned VGG-16

Multilayer Perceptron, CNN, LSTM and CNN-Transfer Learning Model. After performing a comparison study it is clear that the transfer learning VGG16 model performs the better than the rest of the models providing an accuracy of 81%.

Confusion matrix, a visual representation of the performance of all the modela, has been incorporated. It depicts the number of predictions done by the model that have been classified correctly or incorrectly as true positives, false positives, true negatives and false negatives. The precision, recall and F1-scores have been calculated for each model from the confusion matrix obtained.

Precision and Recall: Precision is the measurement of the positive samples that are actually correct. It can be calculated by implementing the formula in Eq.(1).

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{4}$$

Recall is the proportion of the actual positives that have been correctly identified. Eq. (2) represents the formula used to calculate recall.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{5}$$

For a multi-class classification problem, macro averaging and micro averaging is performed for calculating precision and recall. In macro averaging, the average of all the various classes are calculated whereas for micro averaging the precision and recall is calculated separately for each class and later added to produce the final results of precision and recall.

F1-Score: F1-Score has also been calculated as a performance measure for all the models. It is the harmonic mean of Precision and Recall. The F1-score provides a realistic measure of the test data performance using both precision and recall. The formula for calculating F1-Score has been shown in Eq.(3) . According to Table 1, a classification report has been computed for all the language models that have been implemented.

$$F1 - Score = 2 * \frac{precision * recall}{precision + recall} \tag{6}$$

TABLE I
CLASSIFICATION REPORT

Models	Accuracy	Precision	Recall	F1 Score
Decison Tree	0.34	0.38	0.34	0.34
SVM Classifier	0.54	0.46	0.47	0.47
XGBClassifer	0.54	0.55	0.56	0.54
Random Forest	0.68	0.67	0.64	0.65
MLP	0.65	0.65	0.65	0.65
CNN	0.66	0.66	0.68	0.66
LSTM	0.70	0.71	0.70	0.70
Pretrained-VGG-16	0.81	0.82	0.81	0.81

V. CONCLUSION AND FUTURE WORK

One approach to Speech emotion recognition involved doing a comparison study using machine learning models and deep learning models and incorporating transfer learning technique using pre-trained VGG16 CNN model to obtain an accuracy of 81%. Disgust had a lower recognition rate as it is slightly complex in nature and difficult to be detected even by a human. The proposed work uses a compact feature set with a overall good recognition accuracy for seven emotions compared to the reference paper. The future scope of this work involves implementing a therapy chatbot with a suitable user interface. Application of the same in the area of Mental Health - to assess the tone of a person and accordingly provide audio responses catered to the mental health issue.

REFERENCES

[1] Enkhtogtokh Togootogtokh and Christian Klasen, "DeepEMO: Deep Learning for Speech Emotion Recognition" arXiv:2109.04081

[2] E. Lieskovska, M. Jakubec and R. Jarina, "Speech Emotion Recognition Overview and Experimental Results," 2020 18th International Conference on Emerging eLearning Technologies and Applications (ICETA), 2020, pp. 388-393, doi: 10.1109/ICETA51985.2020.9379218.

[3] Q. Zhang, N. An, K. Wang, F. Ren and L. Li, "Speech emotion recognition using combination of features," 2013 Fourth International Conference on Intelligent Control and Information Processing (ICICIP), 2013, pp. 523-528, doi: 10.1109/ICICIP.2013.6568131.

[4] Khalil, Ruhul Amin Jones, Edward Babar, Mohammad Jan, Tariqullah Zafar, Mohammad Alhussain, Thamer. (2019). Speech Emotion Recognition Using Deep Learning Techniques: A Review. IEEE Access. PP. 1-1. 10.1109/ACCESS.2019.2936124.

[5] Ando, A., Mori, T., Kobashikawa, S., Toda, T. (2021). Speech emotion recognition based on listener-dependent emotion perception models. APSIPA Transactions on Signal and Information Processing, 10, E6. doi:10.1017/ATSIP.2021.7

[6] Tang, D., Kuppens, P., Geurts, L. et al. End-to-end speech emotion recognition using a novel context-stacking dilated convolution neural network. J AUDIO SPEECH MUSIC PROC. 2021, 18 (2021). https://doi.org/10.1186/s13636-021-00208-5

[7] Yi-Lin Lin and Gang Wei, "Speech emotion recognition based on HMM and SVM," 2005 International Conference on Machine Learning and Cybernetics, 2005, pp. 4898-4901 Vol. 8, doi: 10.1109/ICMLC.2005.1527805.

[8] M. Xu, F. Zhang and S. U. Khan, "Improve Accuracy of Speech Emotion Recognition with Attention Head Fusion," 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), 2020, pp. 1058-1064, doi: 10.1109/CCWC47524.2020.9031207.

[9] Abbaschian, B.J.; Sierra-Sosa, D.; Elmaghraby, A. Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models. Sensors 2021, 21, 1249. https://doi.org/10.3390/s21041249

[10] https://github.com/musikalkemist/AudioSignalProcessingForML/tree/master/1-

[11] Motamed S, Setayeshi S, Rabiee A. Speech emotion recognition based on brain and mind emotional learning model. J Integr Neurosci. 2018;17(3-4):577-591. doi: 10.3233/JIN-180088. PMID: 30010138.