Deploying a Neural Network Model on Raspberry Pi to Identify the Emotion from Speech

[1]*Rupali Kawade*
dept. of E and TC
Pimpri Chinchwad College of Engineering and Research
Pune, India
rupali.kawade@pccoer.in

[2]*Sayli Kanchan*
dept. of E and TC
Pimpri Chinchwad College of Engineering and Research
Pune, India
sayli.kanchan_etc2019@pccoer.in

[3]*Niranjan Kangane*
dept. of E and TC
Pimpri Chinchwad College of Engineering and Research
Pune, India
niranjan.kangane_etc2019@pccoer.in

[4]*Tejas Kulkarni*
dept. of E and TC
Pimpri Chinchwad College of Engineering and Research
Pune, India
tejas.Kulkarni_etc2019@pccoer.in

***Abstract*—The recognition of emotions is a vast significance and a high developing field of research in the recent years. The applications of emotion recognition have left an exceptional mark in various fields including education and research. Traditionally, in Speech Emotion Recognition, models require a large number of manually engineered features and intermediate representations such as spectrograms for training. For this research paper we've studied many research paper which suggested techniques like Multi Perceptron, Two Stream convolutional network, Multi-task learning model, MFCC, SVM, HMM. The perfection of speech emotion recognition greatly depends upon the types of feature used and also on the classifier employed for recognition. The classification performance is based on extracted characteristics. This paper proposes an implementation of deep learning model on raspberry pi to identify emotion from speech using MFCC (Mel-frequency cepstral coefficients) as an extraction feature and LSTM (Long Short Term Memory) as a classifier, as they proposed higher accuracy compared to other techniques. The proposed deep learning model will be implemented on Raspberry Pi to create a standalone Speech Emotion Recognition system.**

***Index Terms*—Speech Emotion Recognition, Deep Learning, CNN, LSTM, RAVDESS**

I. INTRODUCTION

The speech signal is the quickest and most natural way for people to communicate with one another. This characteristic has encouraged academics to consider voice as a quick and effective means of human-machine communication.Speech emotion recognition is thought to enhance the effectiveness of speech recognition systems by allowing for the extraction of valuable semantics from speech[1].

Human beings have emotions for every item related to them. New designs are the result and technology continue to often enter the market . Since Human thought, feelings, decision-making, communication, and interacting with technology all depend on emotion[1].

Speech emotion recognition is very helpful for applications that call for genuine human-machine interaction, such vehicle board systems that can use driver mental state information to start safety mechanisms. It can also be used as a diagnostic tool by therapists. Speech recognition systems that are trained to recognise stressed speech perform better in aeroplane cockpits than those that are trained to recognise normal speech. Call centre applications and mobile communication have both used speech emotion recognition. Speech emotion detection is mostly used to modify the system's reaction when it detects anger or frustration in the speaker's voice[2].

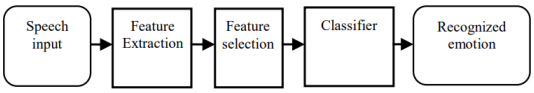

Fig. 1. Structure of Speech Emotion Recognition System

This paper makes a contribution by analysing a distinct Long Short Term Memory network model works as classifier based on the Mel-Frequency Cepstrum Coefficients feature.

The RAVDESS dataset was used to determine the performance parameters using important features such as recall, F-1 score, precision, and accuracy for four emotions, including happy, neutral, sad, and angry.[5]

Many common characteristics are retrieved in recent work, including energy, pitch, formant, as well as certain spectrum characteristics such as linear prediction coefficients and mel-frequency spectral coefficients [5]. In this paper we mainly focused on the MFCC feature for deployment of the model on Raspberry-Pi.[5]

Speech emotion detection is predicted to be useful for a wide range of applications, including other examples include contact centers, video games, audio surveillance, computer lectures, in-car board systems, and robot interfaces . Researchers have developed numerous databases of audio recordings of different languages, such as English, German, etc., recorded by actors or professionals, for the purpose of emotion recognition through speech.

Many of these databases are accessible to the general public. There are numerous languages in India, including Bengali, Tamil, Marathi, Gujarati, Hindi, and more. For the purpose of analyzing emotions, however, there is no database accessible for any of these languages. Some scholars have developed theirown databases, but because they are not standardized, others cannot access them. Consequently, our goal is to develop a database that contains audio files of many states of feeling, including happy, angry, and sad. Additionally, we can make databases for different languages[6].

We encountered a number of issues when putting the Emotion Recognition System into practice, including issues with model accuracy, speaker gender and language use, background noise from the audio recordings, etc.

As a result, we are attempting to create a Speech Emotion Recognition model that will be resilient to all the aforementioned issues and provide the best results possible utilizing the CNN and LSTM model.

II. Related Work

The field of emotion recognition has become more important in an effort to boost the effectiveness of human computer interference. Speech emotion recognition seeks to recognize a person's fundamental emotional state from their voice. Recent years have seen an increase in study interest in this area. There are numerous other examples include contact centers, video games, surveillance cars, online lectures, and robot interfaces [1]. With a goal of increasing efficiency of machine intelligence, the study of emotion identification has become more important. EEG, facial expressions, gestures, voice, and other bodily aspects of emotion have all been studied by researchers. The effectiveness of each LSTM model's Re- call,F1 score, Precision, and Accuracy are evaluated for the four emotions of Happy, Sad, Neutral, and Angry.

Deep neural networks (DNNs) outperform naive networks and Gaussian mixture models (GMMs) on big vocabulary audio recognition tasks, according to recent studies. Through studies on speech recognition, they showed that DNNs can extract more discriminative and invariant characteristics at higher levels. This means that DNNs' learned features are less sensitive to small changes in the input features. This trait makes DNNs more generalizable than shallow networks and makes Context-Dependent-Deep-Neural-Network-Hidden-Markov-Model

(CD-DNN-HMM) more robust while performing voice recognition due to speaker, environment, or bandwidth mismatches.[2] Deep learning techniques such as DBM, RNN, DBN, CNN, and AE have been the subject of much research in recent years. These deep learning methods and their layer-wise architectures are briefly elaborated based on the classification of various natural emotion such as happiness, joy, sadness, neutral, surprise, boredom, disgust, fear, and anger. These methods offer easy model training as well as the efficiency of shared weights. Limitations of deep learning techniques include their large layer-wise internal architecture, less efficiency for temporallyvarying input data and over-learning during memorization of layer-wise information.

This research work forms a base to evaluate the performance and limitations of current deep learning techniques.[4]

Systems for recognising speech emotions based on several classifiers are shown. The signal processing unit, which extracts pertinent features from the available speech signal, and a classifier, which discerns emotions from the speech signal, are the two key components of a speech emotion recognition system.

Most classifiers' average accuracy for speaker independent systems is lower than that for speaker dependent systems.[7] Further, it highlights some promising directions for better SER systems. I

III. PROBLEM STATEMENT

"To identify emotions from speech by implementing anddeploying a deep learning model based on the MFCC feature."

References	Database used	Methodology	Features used	Advantages
[1]	RAVDESS	Based on A Semantic Atlas of Emotional Concepts and Constructivist View of Emotion.	For recognition literature on various databases, distinct characteristics, and classifiers have taken into consideration.	This paper states connections between scientific and traditional conceptions of emotion.
[2]	RAVDESS, SAVEE	In this paper, various deep learning algorithms such as DBMS, DBNS, CNNS, RNNS, RVNNS, are discussed.	Deep neural network, deep Boltzmann machine, recurrent neural network, deep belief network.	This paper provides both simple model training and the effectiveness of shared weights.
[3]	RAVDESS	ANN are used as classifier as they identify nonlinear boundaries separating different emotions. The feed forward neural network is used for speech emotion recognition.	Mel frequency cepstrum coefficient (MFCC), and linear prediction cepstrum coefficient (LPCC)	SER using multilayer perceptron layer is common as it is simple to implement and has a well-defined training algorithm.
[4]	RAVDESS	Deep learning methods and their layer-wise architectures are briefly elaborated based on the classification of various natural emotion	DBM, RNN, DBN, CNN, and AE	Offer easy model training as well as the efficiency of shared weights
[5]	RAVDESS	Evaluate twelve different LSTM models as classifier based on MFCC to identify speech from emotion.	LSTM, MFCC	Accuracy obtained is 89% which is 9.5% more than reported in recent literature.
[6]	CREMA-D, SAVEE	In this paper methods like HMM, GMM, and SVM are used.	SER, LSTM, MFCC	SER has started to gain improved efficiency from the tools offered by deep learning
[7]	RAVDESS	Proposed a two-stream DCNN with an iterative MLP.	MFCC, LSTM	Spectral features including formants, band energies, centers of gravity, roll-off points, cepstral features are used.
[8]	RAVDESS	The speech signal is processed to create a spectrogram, which is used as an image. The CNN is supplied with a sample of frames from a video segment.A SVM receives the to determine the emotions.	CNN, SVM, ELM	The proposed system is evaluated using two audio–visual emotional databases, one of which is Big Data.
[9]	FAU Aibo	A group of hybrid classifiers based on DBN & HMM are proposed and evaluated. The dataset for emotion recognition known as FAU Aibo, we achieve cutting-edge results.	HMM, DBN	Deep Belief Networks are capable of simulating intricate, non-linear high-level interactions between low-level features.
[10]	SAVEE	In this six different types of classifiers were developed to forecast emotions. Six audio files that were taken from the interface audio visual emotion database and used in the classification methods.	Deep Belief Network	In this paper the decision tree was used to evaluate the classifiers. With this feature selection, it was possible to see that each of the compared classifiers improved recall and overall accuracy.
[11]	RAVDESS	Their methodology divides work into several phases where each phase includes several stages, which are: the planning phase, conducting phase, and finally reporting phase.	DNN, HMM, GMM, LSTM.	Explained how to divide task into phases and to achieve maximum accuracy.
[12]	RAVDESS	Using DNN audio signals are evaluated. The approach is analysis of perceptual features to find out most important emotional cues.	PLPC, MFCC, BFCC, DNN	Using Berlin database, the algorithm's validity is examined using seven emotions in 1-D .
[13]	RAVDESS	It shows training on the combination of the ASR plus SER tasks, performs better than training on a single task	ASR,MTL	Proposed multi-task learning (MTL) approach.
[14]	EMO-DB, SAVEE, and RAVDESS	We proposed a two-stream deep convolutional neural network with an iterative neighborhood component analysis.	Spatial-spectral features	In this proposed system using three benchmarks, which included the EMO-DB, SAVEE, and RAVDESS emotional speech.
[15]	EMO-DB and RAVDESS.	Reviewed various emotional SER methodologies and the associated speech databases and compared them from different aspects. Among the databases two early and	GAN, LSTM, Machine learning;	Incorporation of LSTM networks and the introduction of DCN LSTM structures has helped to take the solution to a new level and to give the

IV. METHODOLOGY

Multiple learning algorithms are available for classification. The proposed methodology uses LSTM as the prior value is stored in the memory of the LSTM, and the MFCC coefficients are computed using the previous and current frames of speech [6].

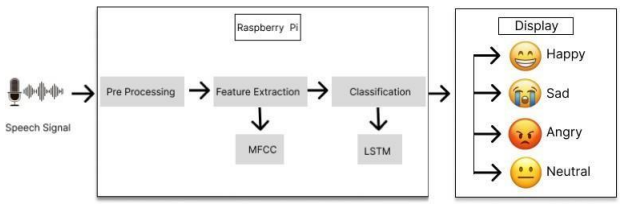

Fig. 2. Speech Emotion Recognition Model

When the speech signal is passed to the Raspberry pi through mic then the speech signal pocesses through the layersof MFCC to extract out feature. MFCC extraction process consist of Pre-emphasis, framing, windowing, the Fourier transform, and wrapping. Pre-emphasis increases a signal's high-frequency portion. In return, the frequency spectrum is balanced, and the signal-to-noise ratio is enhanced. The input signal is imported into frames with a time slots of between 20 and 30 milliseconds and an optional overlap of between 1/3 and 1/2 of the frame size during the framing stage.[5][6]

$$Framesize = FrameDuration * SampleRate \quad (1)$$

Next step is windowing to maintain the continuity of the first and last points in the frame, each frame in this needs to be multiplied by a hamming window. Varied tones in speech signals correspond to different energy distribution over frequencies as more frames are obtained. Hence, a fast Fourier transform is performed to obtain the frequency magnitude response of frame. Next is Mel-frequency,which is kept proportional to the logarithm of the linear frequency. Along the Mel frequency,which is connected to the common linear frequency, these filters are evenly spaced apart.

According to a specific relevant assessment criterion, feature selection seeks to pick a selection of the most important features from the original ones to ensure maximum accuracy. It can considerably shorten the duration of learning algorithms. Long-term dependencies in data are handled by LSTM. The data is kept in LSTM networks for a long time.[6][7]

Python is used to implement the LSTM network. Utilizing the Keras library for implementation. Python-based Keras is an open-source neural network library [8].

Using the method described above, MFCC values for the different emotions are obtained. It graphs the values with respect to coefficients for Neutral, Angry, Happy and Sad emotions. For the 4 samples in the data set, the MFCC is calculated and a graph is drawn. Coefficient values change according to emotions.

A. Steps of implementing trained model on Raspberry PI:

Step 1: Installing a Trained Model on a Raspberry Pi Deployment has a wide range of choices. However, we've decided to use the Raspberry Pi as a deployment platform. Installing the Raspian operating system is the initial step in deployment.

Step 2: After an operating system has been successfully installed, a virtual environment must be created using the Pycharm IDE.

Step 3: Following deployment, libraries will be installed.

Step 4: We require several libraries, including librosa, keras, matplot, panda, and numpy, depending on the needs of our models.

Step 5: The trained model for speech emotion detection must be saved on the Raspberry Pi when this installation is complete.

Step 6: The model must be tested after being saved.

The 1D CNN LSTM model, which consists of three convolutional layers followed by two LSTM layers, is included in the suggested methodology. This model was developed via Tensor flow's sequential technique, keras. The model is also trained, evaluated on a test dataset, and then using Matplotlib, loss and accuracy curves are displayed for both datasets.

Below is a description of these layers and the rest of the deep learning pipeline.

B. CNN - Convolutional Neural Network

A convolutional neural network can have tens or hundreds of layers that each learn to detect different features of animage.Filters are applied to each training image at different resolutions, and the output of each convolved image is usedas the input to the next layer. The filters can start as very simple features, such as brightness and edges, and increase in complexity to features that uniquely define the object. Convolutional neural networks are highly suited for pattern recognition and classification applications. Because CNNs are adept at learning characteristics on their own from the input data provided, many scholars have recently started using them for emotion analysis from voice signals.

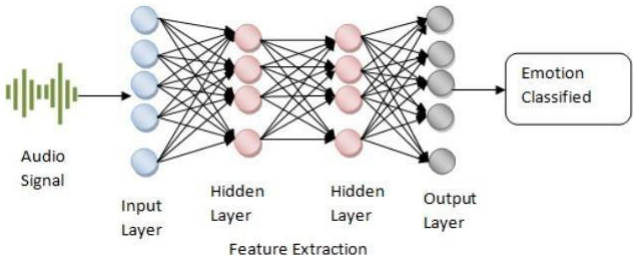

Fig. 3. CNN Architecture

One-dimensional, two-dimensional, and three-dimensional CNN models are the three different model types [9]. While 3D CNN models are frequently used for tasks involving video understanding, 2D CNN models work better for image processing.

C. LSTM - Long Short term memory

Long-short term memory (LSTM), a special type of recurrent neural network, may learn long-term dependencies. This method is now used to manage an array of deep learning difficulties. Many recurrent neural networks (RNNs) are able to learn long-term dependencies, particularly in tasks involving sequence prediction. Besides from singular data points like photos, LSTM has feedback connections, making it capable of processing the complete sequence of data.

This has uses in machine translation and speech recognition, among others. A unique version of RNN called LSTM exhibits exceptional performance on a wide range of issues.
Several tasks that are intractable by earlier learning algorithms like RNNs can be solved by LSTM neural networks. By using LSTM, long-term temporal dependencies may be efficiently captured without facing many optimization challenges. This is applied to solve complex issues.

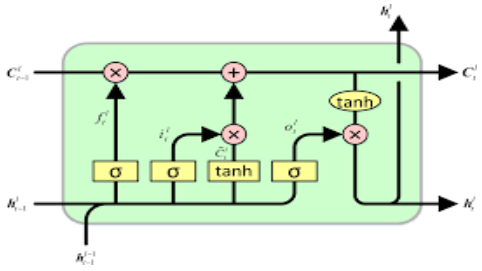

Fig.4. LSTM Architecture

D. MFCC - Mel Frequency Cepstral Coefficient

The windowing of the signal, application of the DFT, calculation of the magnitude's log, warping of the frequencies on a Mel scale, and application of the inverse DCT are the main steps in the MFCC feature extraction technique. A discrete cosine transform (DCT) of a real logarithm of the short-term energy represented by the Mel frequency scale is provided by the MFCC.

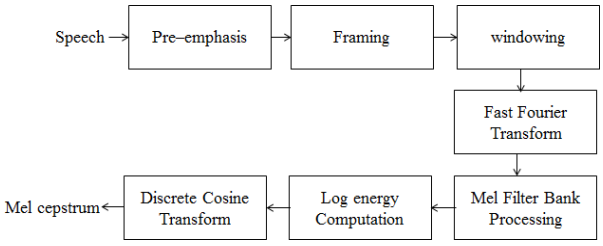

Fig. 5. MFCC Architecture

Pre-emphasis, Framing, Windowing,FFT, Mel filter bank, and computing DCT are steps in MFCC.MFCC extraction process consist of Pre-emphasis, framing, windowing, the Fourier transform, and wrapping. Pre-emphasis increases a signal's high-frequency portion.

In re- turn, the frequency spectrum is balanced, and the signal-to- noise ratio is enhanced. The input signal is imported into frames with a time slots of between 20 and 30 milliseconds and an optional overlap of between 1/3 and 1/2 of the framesize during the framing stage.

Next step is windowing to maintain the continuity of the first and last points in theframe, each frame in this needs to be multiplied by a ham- ming window. Varied tones in speech signals correspond todifferent energy distribution over frequencies as more framesare obtained. Hence, a fast Fourier transform is performed to obtain the frequency magnitude response of frame. Next is Mel-frequency,which is kept proportional to the logarithm of
the linear frequency[5][6].

E. PERFORMANCE MATRICES

Formulae:

Accuracy: Accuracy is how close the value goes to the predicted output.

Accuracy = No. of correct predictions / Total No. of predictions (2)

Precision :It reveals the percentage of predictions in the positive class that were true positive predictions. The following formula should be used to determine precision:

Precision: TP / (TP+FP) (3)

Recall: The number of relevant documents found by a search divided by the total number of existing relevant documents.

Recall = Number of documents searched / Total Number of documents retrieved (4)

F1 score: Combination precision and recall is the harmonic mean of precision and recall.

*F1 Score = 2 * (Precision*Recall) / (Precision + Recall)* (5)

Now the performance parameters for each model for all emotions are calculated using above formulas.[5][6] Based on the logarithmic values from the graph a matrix for each emotion is created, then above-mentioned formulas are implemented on that matrix in LSTM to obtain the accuracy of each matrix to predict the emotion as output.

V. EXPERIMENTAL EVALUATION

Our goal is to create a powerful voice emotion recognition system that predicts emotions without regard to the speaker. Using CNN models, some researchers have attained an average accuracy of about 80% [11]. Compared to the current models, oursoffers 80% accuracy.

Therefore, in this case, we would be classifying and predicting emotions from voice signals using the RAVDESS dataset and a 1D CNN-LSTM model.

A total of 2496 audio files make up the RAVDESS dataset, which is divided into training and testing datasets. A total of 2304 speech files, or 80% of the training dataset, are audio files. A total of 576 speech files, or 20% of the testing dataset, are audio files. Our accuracy with this model and database was 87%.

	Accuracy	Recall	Precision	F1-Score
Angrt	AF	AFG	004	
Happy	AX	ALA	248	
Sad	AL	ALB	008	
Neutral	DZ	DZA	012	

VI. CONCLUSION

This research gathered specific data from 15-17 papers published between the years 2005 and 2021 to give a complete statistical analysis on the use of deep learning in speech-related applications. About 50% of the papers that were found were published in IEEE, and conference papers made up the majority of the papers (40%) that were found.

In the present study, we introduced a speech emotion detection system that classifies emotions using machine learning methods. In order to show a mixture of MFCC traits, they were collected from various acted databases (English). In fact, we research the effects of classifiers and features on the precision of speech emotion recognition. The use of feature selection approaches demonstrates that in machine learning applications, more information isn't necessarily better. In order to identify emotional states from these features,the machine learning model was trained and validated. On the English database, SER claimed the highest recognition rate at "80%".

This result demonstrates that LSTM performs better with less data. As a result, we came to the conclusion that the LSTM model has a better chance of being used practically with little amounts of data.

VII. REFRENCES

[1] Swain, Monorama, Aurobinda Routray, and Prithviraj Kabisatpathy. "Databases, features and classifiers for speech emotion recognition: a review." International Journal of Speech Technology 21 (2018): 93-120.

[2] Khalil, Ruhul Amin, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. "Speech emotion recognition using deep learning techniques: A review." IEEE Access 7 (2019): 117327-117345.

[3] Ingale, Ashish B., and D. S. Chaudhari. "Speech emotion recognition." International Journal of Soft Computing and Engineering (IJSCE) 2, no. 1 (2012): 235-238.

[4] Pagidirayi, Anil Kumar, and B. Anuradha. "An efficient Speech Emotion Recognition using LSTM model." NeuroQuantology 21, no. 1 (2023): 117.

[5] Yu, Dong, Michael L. Seltzer, Jinyu Li, Jui-Ting Huang, and Frank Seide. ”Feature learning in deep neural networks-studies on speech recognition tasks.” arXiv preprint arXiv:1301.3605 (2013).

[6] Bhandari, Sheetal U., Harshawardhan S. Kumbhar, Varsha K. Harpale, and Triveni D. Dhamale. ”On the Evaluation and Implementation of LSTM Model for Speech Emotion Recognition Using MFCC.” In Proceedings of International Conference on Computational Intelligence and Data Engineering: ICCIDE 2021, pp. 421-434. Singapore: Springer Nature Singapore, 2022.

[7] Schuller, Bjo¨rn W. ”Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends.” Communications of the ACM 61, no. 5 (2018): 90-99.

[8] Hossain, M. Shamim, and Ghulam Muhammad. ”Emotion recognition using deep learning approach from audio–visual emotional big data.” Information Fusion 49 (2019): 69-78.

[9] Ra´zuri, Javier G., David Sundgren, Rahim Rahmani, Antonio Moran, Isis Bonet, and Aron Larsson. ”Speech emotion recognition in emotionalfeedbackfor human-robot interaction.” International Journal of AdvancedResearch inArtificial Intelligence (IJARAI) 4, no. 2 (2015): 20-27.

[10] Le, Duc, and Emily Mower Provost. ”Emotion recognition from sponta- neous speech using hidden markov models with deep belief networks.” In 2013 IEEE Workshop on Automatic Speech Recognition and Under- standing, pp. 216-221. IEEE, 2013.

[11] Nassif, Ali Bou, Ismail Shahin, Imtinan Attili, Mohammad Azzeh, and Khaled Shaalan. ”Speech recognition using deep neural networks: A systematic review.” IEEE access 7 (2019): 19143-19165.

[12] Lalitha, S., Shikha Tripathi, and Deepa Gupta. ”Enhanced speech emotion detection using deep neural networks.” International Journal of Speech Technology 22 (2019): 497-510

[13] Cai, Xingyu, Jiahong Yuan, Renjie Zheng, Liang Huang, and Kenneth Church. ”Speech Emotion Recognition with Multi-Task Learning.” In Interspeech, vol. 2021, pp. 4508-4512. 2021.

[14] Kwon, Soonil. ”Optimal feature selection based speech emotion recogni-tion using two-stream deep convolutional neural network.” International Journal of Intelligent Systems 36, no. 9 (2021): 5116-5135.

[15] Abbaschian, Babak Joze, Daniel Sierra-Sosa, and Adel Elmaghraby. ”Deep learning techniques for speech emotion recognition, from databases to models.” Sensors 21, no. 4 (2021): 1249.