SPEECH EMOTION DETECTION USING DEEP LEARNING TECHNIQUE

PokanatiJyothsna PG Scholar, Dept. of Electronics and Communication Engineering, Andhra University College of Engineering, Andhra University, Visakhapatnam.
P. V. Sridevi Professor, Dept. of Electronics and Communication Engineering, Andhra University College of Engineering, Andhra University, Visakhapatnam

ABSTRACT
The key issue of emotion detection is choosing the speech database, identification of various variables connected to speech, and model selection. Emotional speech recognition has advanced from a routine activity to a crucial part of Human-Computer Interaction (HCI). Mel Frequency Central Coefficient, or MFCC, is employed in this article to extract features. The approach is based on recurrent neural networks (RNN) and long short-term memories (LSTM). The database is TESS (Toronto Emotional Speech Set). There are 7 emotions in the TESS dataset. They are indifferent, fearful, happy, surprised pleasantly, sad, and angry. This essay makes use of these 7 emotions. By utilizing this model, an accuracy of about 83% is obtained.

KEYWORDS:RNN, LSTM,MFCC,TESS,HCI

I.INTRODUCTION

Humans communicate their feelings most naturally through words. The detection and analysis of emotions are crucial in this digital age of distant communication since they play a significant role in communication. Because emotions are so subjective, recognizing them can be difficult. Emotions cannot be quantified or categorized using a standard technique. Because speech is subjective, it has become increasingly difficult to discern emotions through it.

This task is complex and demanding for people, but simple for machines. Instead of using conventional devices as input, these systems primarily strive to give natural engagement with machines by direct speech interaction, making it simple for human listeners to react[1]–[3]. Speech emotional analysis is a machine task that can recognize and foretell a person's emotions. Speech recognition has advanced from a specialized field to become the key element of the human-computer interface [4]– [7]. Providing natural interaction with the device's interface is the primary goal of this article.

Emotion recognition is a lot needed in fields like medicine, autopilot machines, division, artificial sound help, etc. Many real problems can be simplified with this technology. The unusual problem of figuring out that person's emotional state can serve as a benchmark for each emotion detection software. Identifying emotions like anger, sadness, disgust, fear of surprise, fun, and joy must be the model's primary goal. The use of automatic emotion recognition has great potential in numerous intelligent systems, including digital advertising, online games, evaluating customer health feedback, and many others. Players can have more options, for instance, in an online game system that includes emotion recognition features. The game's aesthetic and entertainment factor can be adjusted when there is an immediate emotion. Deep Learning is a research field like machine learning which comes under artificial intelligence and has achieved more attention in recent years [8]. Recently SER Technology has improved dramatically. Before a decade there is a big difference between old generation technology and present technology. Now different Deep learning and machine learning technology is available and made the analysis easy.

II. SPEECH EMOTION RECOGNITION

Emotion recognition is an area of research that tries to infer feelings from audio signals. SER has its own application designed to detect emotions. Recognizing emotions is difficult in terms of manners, for example, feelings can be contrasted in the face of climate, culture, and individual facial reactions. Learning different speeches is very helpful in learning emotion recognition through language.

Basic Steps Involved in SER

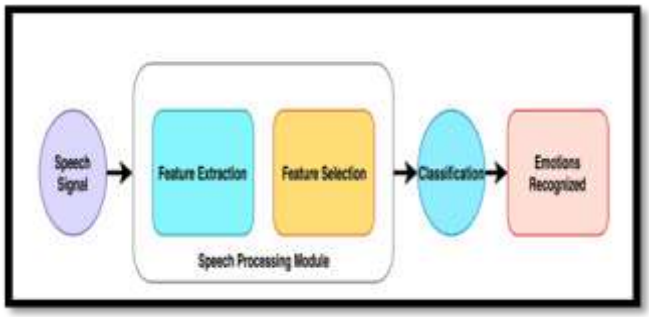

Figure 1: Block Diagram of Emotion Recognition of Speech Signals

Figure 1 shows the most important steps of Emotion Speech Recognition. The most important step involved in recognizing emotions from the speech is audio signal processing, feature extraction, and classification. Pre-processing includes Noise reduction and extracting quieter and more economical sound which means complete information. Feature extraction involves extracting relevant data such as MEL coefficients available in the audio input. Finally, the classification block says what is certain emotions from certain human languages.

Feature extraction is the most important aspect of speech recognition. Feature extraction reduced the magnitude of the speech signal, which decreases the interference with other signals. Speech Classification refers to a set of tasks or problems of getting a program to automatically classify input utterances or audio segments into categories, such as Speech Command Recognition(multi-class), Voice Activity Detection (binary or multi-class), and Audio Sentiment Classification (typically multiclass) etc. The out put will be the respective emotion recognized.

Feature extraction with MFCC includes the following steps:

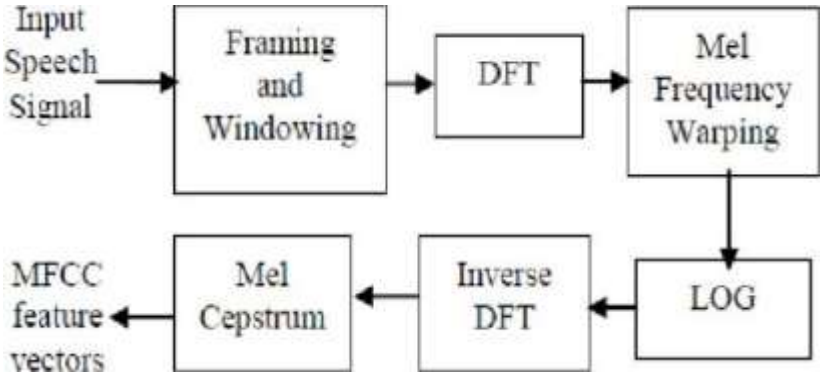

Figure2: Feature Extraction Using MFCC

Step1:Frame blocking

The input audio signal is divided into several frames. Each frame is about 20 to 40 milliseconds long and each frame has an interval. The reason for keeping the distance is to get a smooth transition between frames with pre-set sizes.

Step2:windowing

Windowing is the process that contains the signal multiple time tracking by smoothing window with a finite length of varying amplitude gentle and gradual at the end to zero. The smoothing time or interval of the window is determined by the number of samples. Multiplication in the time domain is equivalent to convolution in the frequency domain.

The signal from each frame passes through the window function. Windows are used to reduce cracking and distortion at each frame's boundaries.

The Hamming window is used here.

$$Y(n) = x(n)w(n), \quad 0 \leq n \leq N-1$$

$$W(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N-1.$$

Step3: Fast Fourier Transform

The specified window signal is transformed to the frequency domain using the Fast Fourier Transform. It converts N samples in the time frame to the time-frequency domain.

Step4: Mel Frequency Wrapping

The Mel frequency scale is linearly distributed for frequencies below 1000 Hz and logarithmically distributed above 1000 Hz. The image below explains what the MEL scale looks like. Here is the formula to calculate MEL frequencies.

$$mel(f) = 2595 * \log 10\left(1 + \frac{f}{700}\right)$$

In general, the MEL filter bank converts to MEL frequency. The structure on the back of the MEL filter has overlapping triangular filters. By taking the reference cut off frequency from the center frequency for each calculated filter.

Step5: Computing the Mel filter bank

To view the bottom and top filter banks frequency must be selected. The first filter bank starts at the first point, peaks at the second point, and then returns to zero at the third point. The second filter bank starts at the 2^{nd} point, peaks at the 3^{rd} point, then becomes zero at the 4^{th} point, and so on.

The formula for their calculation is as follows.

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \le k \le f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)} & f(m) \le k \le f(m+1) \\ 0 & k > f(m+1) \end{cases}$$

Here M is the required number of filters

Table 1: Summary of Some of the Acoustic Variations Observed Due to Emotion

Emotions	Pitch	Intensity	Speaking rate	Voice quality
Anger	abrupt on stress	much higher	marginally faster	breathy, chest
Disgust	wide, downward inflections	lower	very much faster	grumble chest tone
Fear	wide, normal	lower	much faster	irregular voicing
Happiness	much wider, upward inflections	higher	faster/slower	breathy, blaring tone
Joy	high mean, wide range	higher	faster	breathy; blaring timbre
Sadness	slightly narrower	downward inflections	lower	resonant

III. DEEP LEARNING

Basically, deep learning is a subset of machine learning, which is summarized under the general term AI, artificial intelligence. Inspired by the neurons of the human brain and their functions, deep learning comes in the form of artificial neurons. After that, deep learning became an interesting topic in AI and caused a resurgence in neural network research. It is an artificial neural network (ANN) with a large number of hidden layers between the input (information) and output (revenue) layers. This is what an artificial neuron or neural network with a single-layer perceptron looks like. When more neurons are connected in multiple layers, it is called DNN (Deep Neural Network).Figure 3 shows the single-layer neural network of the perceptron. Figure 4 shows the inner layer of the neural network, which consists of an input layer, two hidden layers, and an output layer.

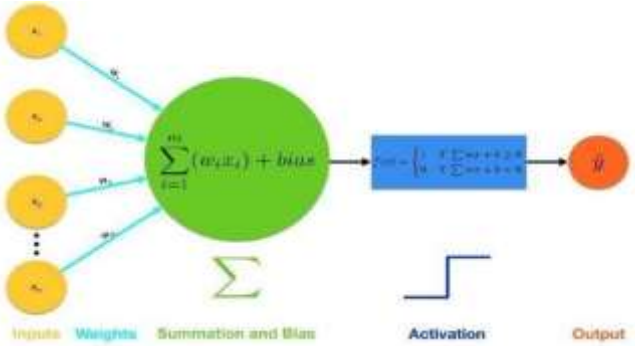

Figure 3: Single Layer Perceptron Neural Network

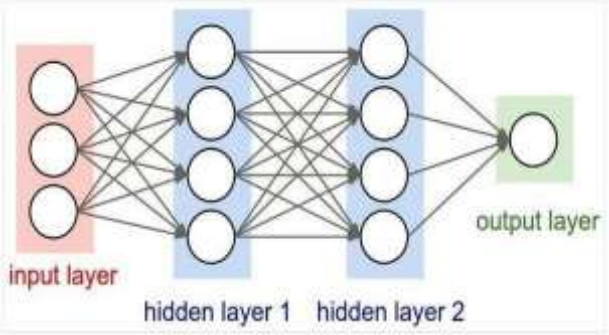

Figure4: Two or More Hidden Layers Contain Deep Neural Networks

IV. Recurrent Neural Network

RNN works on the principle of storing the output of a given layer and feeding the output layer back to the input to predict the output of that layer. An RNN can process sequential data by taking the current input and the previously received input. The RNN can remember previous inputs thanks to its internal memory. In this work, many-to-many RNNs are used, i.e., this RNN takes an input sequence and produces an output sequence. An RNN has a "memory" that remembers all the information about a computation. It uses the same parameters for each input as it performs the same task for all inputs or hidden layers to produce output. Unlike other neural networks, it reduces parameter complexity. As can be seen in Figure 5, nodes in different layers of the neural network are compressed to form an iterative neural network layer. Network parameters are A, B and C.

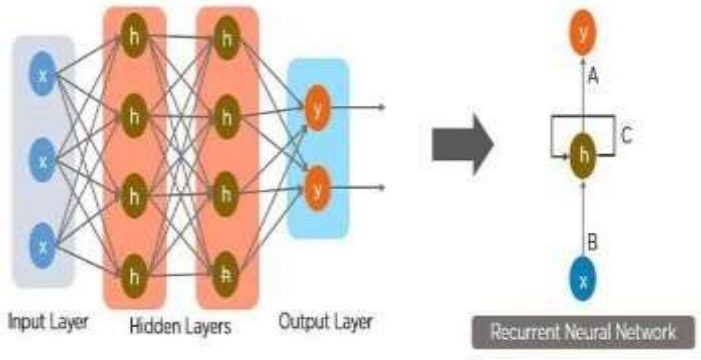

Figure 5: Recurrent Neural Network

As can be seen in Figure 5, nodes in different layers of the neural network are compressed to form an iterative neural network layer. Network parameters are A, B, and C.

V. LITERATURE SURVEY

Extracting key features is the key to this research recognizing emotional language. F. Wang et al used the Deep Auto Encoder (DAE) method which contains five hidden layers for input extraction characteristics of the sound signal [9]. They also extract traditional features of speech signals for emotions including MFCC (Mel frequency cepstral coefficient), and linear prediction of perception. The cepstral coefficient (PLP), and LPCC verbal signal. Finally, the model they use SVM models. The result of the confession is obtained by using all functions. The results of this study indicate that the emotional speech features extracted from DAE showed a marked improvement compared to other documents. ZT Liu et al. have proposed the SER framework.

First, they take the initial set of functions consisting of speaker-independent functions and speaker dependent functions by extracting functions from the audio input signal. In the second step, they selected features using correlation analysis, which consisted of distance analysis, partial correlation analysis, bivariate correlation analysis, and Fisher's exact test. Then the emotional features of the speech, which are repeated, are discarded. Then the optimal feature subset is obtained. Finally, a decision tree for extreme machine learning (ELM) for emotion recognition was constructed. The experimental results show that the ELM for the decision tree algorithm works more efficiently and the work efficiency of feature selection based on correlation analysis and Fisher's criteria is fully verified. A complete analysis of the emotional speech recognition system is described in, in which various properties of the data set and the choice of classifiers for emotional speech recognition studies are considered.

In [10] different acoustic features of sound and voice signals are examined and different classification methods are analyzed, which is useful for further application of modern emotion recognition methods. This work analyzed the prediction of subsequent responses of emotional audio signals based on emotion recognition using various types of classifiers. Some of the most famous algorithms such as KNN and Random Forest are used in [11] to classify different emotions. The Recurrent Neural Network is growing very strong, which has helped solve many problems in the field of artificial intelligence (AI). In [12], a deep RNN like LSTM, a two-way LSTM trained on salient features, was used. In [13] different CNN ranges and models were analyzed and implemented and trained for speech and emotion recognition.

VI. PROPOSED METHODOLOGY

This section describes the proposed methodology, the audio database used for the study, and the LSTM model.

A. Audio Database

In order to train and test the model, any audio signal is given as input. The audio data can be of real time audio which generally people speak or can consider any audio databases which are available on different online platforms like Kaggle. Kaggle offers different varieties of audio databases. The audio dataset used in this article is the TESS (Toronto Emotional Speech Set) dataset, which stands for Toronto Emotional Speech Set. As this data is rarely used, this project examines this data set further. It contains 2800 language files in .wav format. It contains a set of 200 target words spoken by two female actresses (26 and 64 years). The set shows each of the seven emotions. They are anger, disgust, fear, happiness, pleasant surprise, sadness, and neutrality. During the data collection session, the actor's emotions were assessed by different commentators in seven emotion domains. All data is provided along with the database.

B. LSTM Model

LSTM stands for Long Short-Term Memory Networks which is used in deep learning. These are various repetitive neural networks (RNNs) capable of long-term dependency learning, especially in sequence prediction problems. LSTM networks are well suited for classifying, processing, and making predictions based on time series data. In this paper, the LSTM model is used as the modeling basis. This item has an input layer, two hidden layers, and an output layer. The two hidden layers are the solid layer and the falling layer. There are 128.64 neurons in the dense and dropout layers. The output layer consists of 7 neurons because this model recognizes 7 emotions. In this model, the RELU and SOFTMAX functions are used as activation functions and the ADAM optimizer is used to optimize the model.

C. Preparation of Training Dataset

All audio clips in the TESS data set were extracted from different sessions. It contains a set of 200 target words spoken by two female actresses (26 and 64 years). Using the emotion rating report provided with the database, various audio files in .wav format are labeled and grouped into seven categories. The range of emotions, as mentioned earlier. The speech signal in .wav format is converted into a spectrogram image in the emotion class.

D. Training Method

All audio recordings tagged with appropriate emotions were prepared for model training. The proposed LSTM model is implemented using an iterative neural network. The spectrogram images were generated from the audio recordings of the TESS data set. The learning process lasts more than 50 epochs. The batch size is set as 64. The training data model is performed on a Kaggle online notebook. The practice lasted about 35 minutes and the best accuracy was achieved in the 49th epoch. A loss of 0.0068 was achieved in the training set, while a loss of 1.8283 was recorded in the test set. 83% accuracy achieved. It should be noted that the overall accuracy is somewhat lower. This can be caused by missing files in the respective dataset emotion class.

Fig.6: Training and Testing the Model

E. Results &Analysis

The emotion prediction data model achieves an accuracy rate of about 83%. It is evident that 83 % is the highest level of accuracy achieved during input validation. The loss rate achieved for data validation is 1.8283.

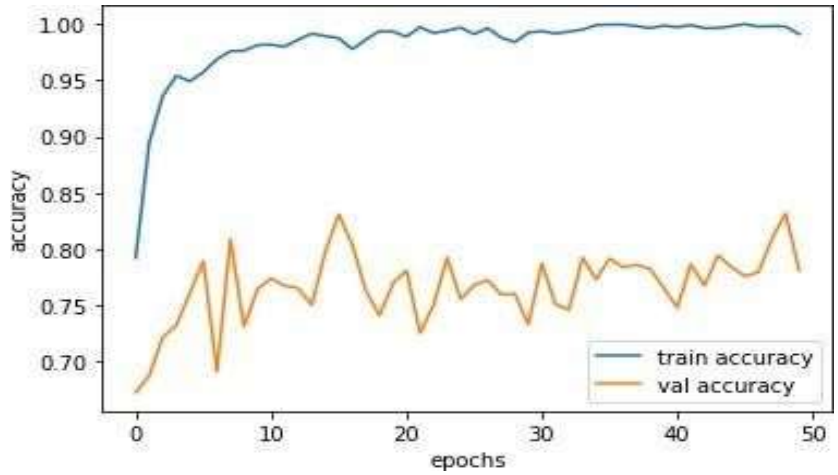

Figure 7: Plot of Train Accuracy and Validation Accuracy

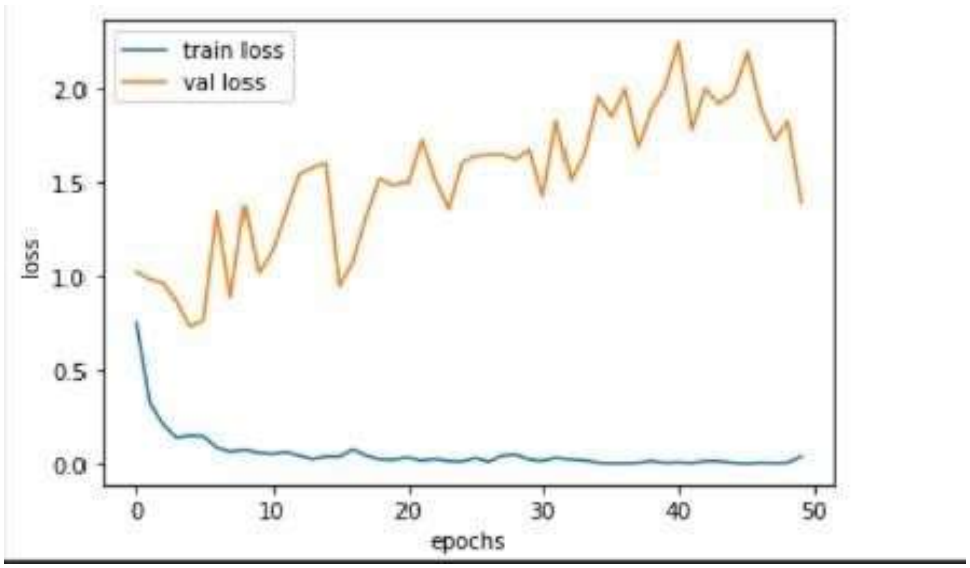

Figure 8: Plot of Train Loss and Validation Loss

In figure7 the blue color graph represents accuracy of training data and the orange color graph represents the accuracy of testing data. As it can be seen training accuracy is more than validation accuracy. The validation accuracy is fluctuating in different epochs. Among the input data, 80% of the data went into training the model. Validation accuracy represents the accuracy of remaining data which is not included in testing. Hence the validation accuracy is the accuracy of testing data. In figure8 the blue color graph represents the accuracy of training data loss and the orange color graph represents the accuracy of testing data loss. As it can be seen training data loss accuracy is less than validation data loss. The validation data loss is fluctuating in different epochs. Among the input data, 20% of data is went into testing the model. If the validation loss is reduced then this model can achieve even better accuracy.

VII. FUTURE WORK

In this paper the LSTM model has been built and it is working with 83% accuracy. This model is able to depict with how much accuracy it can predict the emotion. But it doesn't display which emotion is predicted. Hence in the future, this model can be modeled in such a way that it can give concrete emotions as output. Many techniques are available to help further improve network consistency and generalizability. Additional datasets such as the IEMOCAP, SAVEE, and RAVDESS datasets can be added to the model to improve prediction accuracy. Different models such as CNN + RNN + LSTM can be combined to get a more accurate and efficient model. These different approaches will lead to different

applications of emotion detection from speech, trying to detect emotions using sound waves and using different techniques such as: B. Recognition of emotions through text in conversation, through speech.

VIII. REFERENCES

[1] J. G. Rázuri, D. Sundgren, R. Rahmani, A. Moran, I. Bonet, and A. Larsson, "Speech emotion recognition in emotional feedback for human-robot interaction," *International Journal of Advanced Research in Artificial Intelligence(IJARAI)*, vol. 4, no. 2, pp. 20–27, 2015.

[2] A. B. Nassif, I. Shahin, I. Attili, and M. Azzeh, "Speech recognition using deep neural networks: A systematic review," *IEEE Access*, vol. 7, pp. 19–143, 2019.

[3] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[4] M. S. Hossainandg, "Emotion recognition using deep learning approach from audio-visual emotional bigdata," *Information Fusion*, vol. 49, pp. 69–78, 2019.

[5] M. Chen and P. Zhou, "Emotion communication system," *IEEE Access*, vol. 5, pp. 326–337, 2017.

[6] N. D. Lane and P. Georgiev, "Can deep learning revolutionize mobile sensing," in *in Proceeding softhe16thInternationalWorkshop on Mobile Computing Systems and Applications*, ACM, 2015, pp. 117–122.

[7] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, *Emotion recognition by speech signals*. 2003.

[8] W. Fei, X. Ye, and Z. Sun, "Research on speech emotion recognition based on deep autoencoder," in *Cyber Technology in Automation, Control, and Intelligent Systems(CYBER)*, IEEE, 2016, pp. 308–312.

[9] S. G. Koolagudi and S. R. Krothapalli, "Emotion recognition from speech using sub-syllabic and pitch synchronous spectral features," Int," *Int.J.SpeechTechnol*, vol. 15, no. 4, pp. 495–511, 2012.

[10] F. Noroozi, N. Akrami, and G. Anbarjafari, "Speech-based emotion recognition and extraction prediction," 2017 25th Signal Process," *Commun. Appl. Conf.SIU2017*, no. 1, 2017.

[11] A. Graves, A. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," *IEEE International Conference on Acoustics*, pp. 6645–6649, 2013.

[12] C.-W. S. Huangands, "Characterizing Types of Convolutions in Deep Convolutional Recurrent Neural Networks for Robust Speech Emotion Recognition," pp. 1–19, 2017.

[13] A. Batliner *et al.*, "The automatic recognition of emotions in speech," in Emotion-Oriented Systems," pp. 71–99, 2011.