Affective Model Based Speech Emotion Recognition Using Deep Learning Techniques

** Karthika Renuka D.*
*** Akalya Devi C.*
**** Kiruba Tharani R.*
***** Pooventhiran G.*

Abstract

Human beings express emotions in multiple ways. Some common ways that emotions are expressed are through writing, speech, facial expression, body language or gesture. In general, it is believed that emotions are, first and foremost, internal feelings and experience. Speech is a powerful form of communication that is accompanied by the speaker's emotions. Specific prosodic signs, such as pitch variation, frequency, speech speed, rhythm, and voice quality, are accessible to speakers to express and listeners to interpret and decode the full spoken message. This paper aims to establish an affective model based speech emotion recognition system using deep learning techniques such as *RNN* with *LSTM* on German and English Language datasets.

Keywords : **Emotion recognition, RNN, Speech, Neural Network**

I. INTRODUCTION

Affective computing is the process of recognition of emotions in-order to enrich the Human Computer Interaction (HCI). Basic emotions types are happy, sad, disgust, surprise and neutral [1]. Some common ways that emotions are expressed are through writing, speech, facial expression, body language, gesture, etc. In real-time monitoring, the facial expressions, voice and biological signals help to identify the current emotions of users and know whether a person is under stress or not. This paper focuses on developing a deep learning based model for emotion recognition from speech signals. The application of speech emotion recognition plays a major role in medicine. The speech of a patient in Psychology treatment helps to analyze the patient's mentality and provide treatment. It also provides solutions for Autism Spectrum Disorder by social contact and communication together with limited and repetitive behaviors. Many healthcare systems use virtual reality to identify, assess, and manage patients with mental, phobic, anxiety, and stress disorders.

Applications like e-Learning, driver drowsiness video gaming, call center systems, and marketing would improve user experience by incorporation of emotion recognition. There are different ways to detect emotions conveyed across different modalities, and some of them are described next.

II. EMOTION RECOGNITION

A. Text Based Emotion Recognition

One of the intelligent machine's important capabilities is the affective ability, which allows it to understand

Manuscript Received : May 22, 2020 ; Revised : August 10, 2020 ; Accepted : August 16, 2020. Date of Publication : September 5, 2020.
* K. Renuka D. is Associate Professor with Department of IT, PSG College of Technology, Coimbatore 641 004, Tamil Nadu, India. (email : karthirenu@gmail.com)
** A. Devi C. is Assistant Professor with Department of IT, PSG College of Technology, Coimbatore 641 004, Tamil Nadu, India. (email : akalya.jk@gmail.com)
*** K. Tharani R. is PG Student with Department of IT, PSG College of Technology, Coimbatore 641 004, Tamil Nadu, India. (email : meetkirupa@gmail.com)
**** Pooventhiran G. is UG Student with Department of IT, PSG College of Technology, Coimbatore 641 004, Tamil Nadu, India. (email : pr.gksp@gmail.com)
Doi : 10.17010/ijcs/2020/v5/i4-5/154783

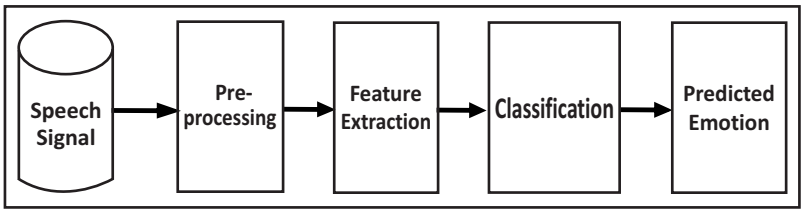

Fig. 1. General Speech Emotion Recognition Frameworks

and express emotions, and it has become an emerging research area within artificial intelligence. However, text messages are still the most popular means of communication. There are many applications for text messages and it is important to effectively recognize emotions from texts. For instance, an intelligent Twitter conversation can recognize emotions from a user's discussion and provides human-like reactions. For the implementation of the attractive human-computer communication, the detection of emotions from text messages is noteworthy. The two main methods for recognizing emotions from text include machine learning and knowledge-based approach. They are initiated by building an affective lexicon and then combining the syntactic and semantic rules to recognize the emotions of texts. In the specific domains, the above two methods work well, but their success is highly dependent on the quality of their syntactic rules and affective lexicon. The real challenge in text based emotion recognition is to create higher quality affective lexicons and syntactic rules.

B. Image & Video Based Emotion Recognition

In human communication, facial expressions are essential factors that help to understand other people's intentions. The facial recognition system through image and video is capable of identifying, analyzing and manipulating emotions. The video data can also be aggregated through a series of frames of variable length images from which emotions can be captured and recognized. Recently, deep learning approach based on Convolutional Neural Networks (CNN's) which focused on temporary averaging and pooling operations to facial analysis and video analysis have shown high performance for emotion recognition. [2].

C. Speech Based Emotion Recognition

Speech emotion recognition is the most recent research in speech handling. Other than human outward appearances, speech has been demonstrated as one of the most encouraging modalities for the programmed recognition of human emotions. Speech emotion recognition technology is one of the fastest-growing fields of engineering. It has several applications in different areas and provides potential benefits. About 20% of people in the world are suffering from various disabilities; a majority of them are blind. In these cases, the speech emotion recognition systems offer a significant advantage. In general speech emotion recognition model as shown in Fig. 1. The speech signal is first preprocessed and the feature extraction is done. Then classification is used to predict the various emotions.

The challenges in recognizing emotions from the speech of the speaker are due to the following reasons :

(1) There may be different emotions in the same utterance.

(2) If the sound level of the voice is low then the conclusion cannot be drawn.

III. SPEECH RECOGNITION

Speech recognition is a method of changing over an acoustic sign which contains data of thought that is shaped in the speaker's mind. Automatic Speech Recognition (ASR) considers acoustic data contained in the speech signal. Audio-Visual Speech Recognition (AVSR) loads ASR as it utilizes acoustic data contained in the speech.

The central difficulty of speech recognition is that the speech sign is a profound factor because of various speakers' substance and acoustic conditions. The element examination segment of an ASR framework assumes a critical job in the general execution of the framework.

Types of Speech

Speech recognition frameworks can be isolated in a few distinct classes by depicting what kinds of expressions they can perceive. These classes are differentiated as follows:

A. Isolated Words

Isolated word recognizers as a rule require every expression to have quiet on the two sides of the example window. It acknowledges single words or single articulation at once. These frameworks have "Listen/Not-Listen in" states, where they require the speaker to hold up between articulations. Isolated expression may be a superior name for this class.

B. Connected Words

Connected word frameworks are like isolated words. However, it enables separate articulations to be run-together with an insignificant delay between them.

C. Continuous words

Continuous speech recognizers enable users to talk normally, while the PC decides the substance. Recognizers with continuous speech capacities are the most hard to make since they use unique strategies to decide articulation limits.

D. Spontaneous words

At a fundamental level, it tends to be thought of as speech that is regular sounding and not practiced. An ASR framework with unconstrained speech capacity ought to have the option to deal with an assortment of common speech highlights.

This paper aims to develop an affective model based speech emotion recognition using LSTM-RNN to identify various emotion states such as happy, sad, neutral, angry, surprise on RAVDESS and EMO-DB datasets. First by sample framing with a hamming window, then pre-process the wav files in the database. Each frame's data are known as a test. Divide the data set for training and testing. Next, the training data set is inserted into the proposed affective model for automated extraction of features and identification of emotion states. Eventually, experimental design demonstrates model analysis.

IV. RELATED WORKS

(1) KunHan et al. proposed a paper on Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine to use deep neural networks (DNNs) [3] to extract high-level features from raw data and demonstrate that they are effective in recognizing speech emotions. First, they generate a distribution of probability of emotional state using DNNs for each segment of speech. First, generate a distribution of probability of emotional state for each segment of speech using DNNs. Then construct utterance-level features from probability distributions at the segment-level. Their experimental results suggest that this approach significantly improves the quality of recognition of emotions from speech signals and it is very exciting to use neural networks to learn emotional information from low-level acoustic characteristics.

(2) Abdul Malik Badshah et al. introduced an emotion speech recognition system for smart affective services based on deep functionality [4]. They presented a study of speech emotion recognition based on features with rectangular kernels derived from spectrograms using a deep convolutional neural network (CNN). To analyse speech through spectrograms, rectangular kernels of varying shapes and sizes, along with max pooling in rectangular neighbourhoods, to extract discriminative features was developed. The performance is evaluated on Emo-db and Korean speech dataset.

(3) Yuki Saito et al. [5] developed a paper on Statistical Parametric Speech Synthesis Incorporating Generative Adversarial Networks. They proposed a framework including the GANs. The discriminator is prepared to separate among normal and created discourse parameters, while the acoustic models are prepared to limit the weighted aggregate of the standard insignificant loss of age and the ill-disposed loss of the discriminator.

(4) Xi Zhou et al. represented a paper on deep learning based Affective Model for Speech Emotion Recognition [6]. They established two affective models based on two methods of deep learning of a stacked auto encoder network and a deep belief network for automatic emotion feature extraction and emotion state classification. The results are based on a well-known German Berlin Emotional Speech Database and, in the best case; the accuracy of recognition is 65%.

(5) Teng Zhang et al. [7] Have proposed speech emotion recognition with i-vector feature and rnn model to recognize the real-world communication function. They developed conventional prosodic acoustic features and compared the novel features to reflect the signal of speech.

Here the method of the Recurrent Neural Network is used to map the features of emotion tags. The features of the i-vector resulted in a performance improvement from 38.3% to 42.5%, and a simple combination of them achieved 43.3% better performance.

(6) Kun-Yi Huang et al. presented a paper on Speech Emotion Recognition Using Autoencoder Bottleneck Features and LSTM [8]. This work combines the characteristics of LLDs and DSS. Deep Scattering Spectrum (DSS) can achieve more accurate distributions of energy in the frequency domain than the Low Level Descriptors (LLDs), and then the Autoencoder neural network is used to extract bottleneck features to minimize dimensionality. The emotional MHMC database was compiled and used to assess performance. The experimental results show that the proposed method using bottleneck features from the combination of LLDs and DSS has achieved a 98.1% accuracy in emotion recognition.

(7) Shumin An et al.,published a paper using LSTM-RNNs on Emotional Statistical Parametric Speech Synthesis [9]. This study uses recurrent neural networks (RNN) with long-term memory units (LSTM) to test methods for parametric speech synthesis (SPSS) in emotional computation. Experiments integrate and compare emotion-dependent modeling and integrated modeling with LSTM-RNN-based emotion codes. Acoustic models are designed separately for each emotion type. An emotion code vector is introduced into all layers of the process in the integrated LSTM-RNN model to display the emotional characteristics of the current utterance. Experimental results on an emotional speech synthesis database with four emotional. types (neutral, happiness, anger and sadness). Using HMM, in consideration to the identification of subjective emotions for artificial expression, the emotion-dependent modeling approach outperforms the integrated modeling approach and emotion-dependent modeling.

V. DATASET DESCRIPTION

The proposed model classifies the speech emotion recognition using deep learning techniques for Ravdess from Kaggle and Emodb German database as follows:

EMO Database

Berlin Emotional Speech Database : The EMODB Dataset comprises 494 utterances documented in an isolation chamber by 10 professional actors (5 male/5 female). The actors performed ten sentences in seven different emotions neutral, angry fear, joy, sorrow, disgust, and boredom. A total of 800 phrases were registered. Through performing a perception experiment on the recognisability of the emotions and their naturalness, all utterances with a recognisability of more than 80% and naturalness of more than 60% were selected as final data samples.

RAVDESS Database

The RAVDESS is a multimodal repository evaluated by speech and music in terms of emotion. The dataset has about 24 professional actors who express lexically appropriate statements in a neutral North American accent. Speech contains expressions of sorrow for peaceful joy, shock, and indignation for rage anxiety, and song combines emotions of peace, satisfaction, disappointment, annoyance, and terror. The expression is produced at two levels of emotional frequency with an additional neutral expression. For all circumstances, face-to-face, face-only, or voice-only versions are available.

The collections of 7356 recordings were analyzed on the level of emotional validity and 10 times per quality. 247 individuals representing untrained study participants in North America earned ratings. Next, a set of 72 participants received test-retest data. It has archived significant levels of mental unwavering quality and exactness of test-retest translators. The exactness and composite proportions of "goodness" were given to help analysts in choosing improvements.

TABLE I.
DATASETS DESCRIPTION

Database	RAVDESS	EMO Database
No of samples	1440	500
No of Class	5	5
Types of class	Happy, Sad, Angry, Surprise, Neutral	Happy, Sad, Angry, Surprise, Neutral

VI. PROPOSED METHODOLOGY

The proposed work using hybrid deep learning

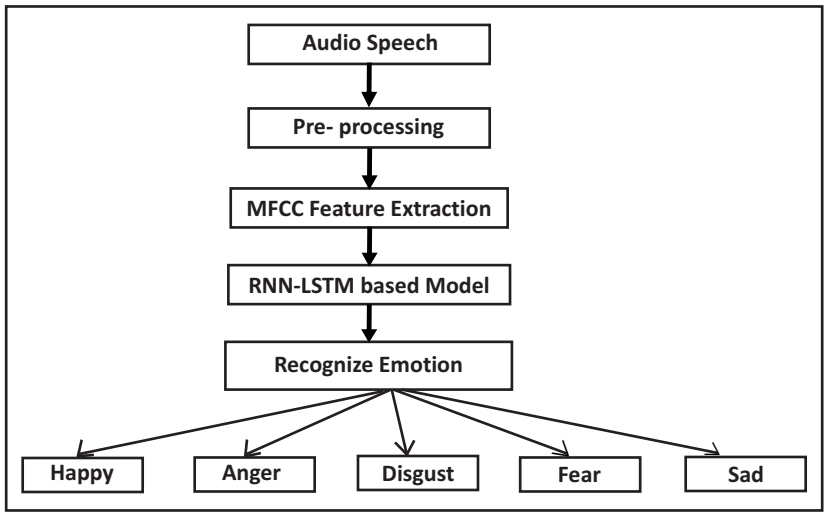

Fig. 2. Proposed Framework

methods namely RNN and LSTM to evaluate human emotion through speech signals is shown in Fig. 2. Table I shows the summary of datasets used for evaluation of the proposed model to identify the type of emotion taken from the datasets, such as happy, sad, anger, surprise disgust, fear for the specific audio expression.

Feature Extraction

Specific information found in the speech signal is characterized and understood by the speaker. The feature extraction process incorporates the raw signal into feature vectors illustrating speaker-specific properties and removing statistical redundancies.

MFCC

Mel-Frequency Cepstral Coefficients (MFCC) properties are the most widely used in feature extraction process of speech recognition. This blends the advantages of cepstrum research with a critical band-based perceptual frequency scale. MFCC is based on perceptions of human hearing that cannot detect frequencies above 1 KHz. In other words, in MFCC, the critical range of the human ear with amplitude is based on established variability. MFCC has two filter types that are linearly spaced at low frequencies below 1000 Hz and logarithmic spacing above 1000 Hz.

MEL SCALE

The Mel scale refers to its actual measured frequency perceived frequency or pitch of a pure tone. Human beings are much easier at low frequencies to detect small changes in pitch than at high frequencies. Through implementing this scale, our features are xx

The formula for converting from frequency to Mel scale is:

$$M(f) = 1125 \ln (1 + f/700) \quad (1)$$

To go from Mels back to frequency:

$$M^{-1}(m) = 700(\exp (m/1125) - 1) \quad (2)$$

RNN- Recurrent Neural Network

The Recurrent Neural Network (RNN) is the network that has at least one feedback relation, which enables the activations to flow in a loop as shown in Fig 3. This helps the networks temporarily process and understand sequences such as sequence recognition / reproduction and temporary association / prediction. The recurrent neural network architectures is in various types. The common type of RNN is a generic Multi-Layer Perceptron (MLP) plus an additional loop. The loop s of MLP exploit efficient nonlinear mapping abilities and have some form of memory. Others have more hierarchical structures, potentially connected with each other's neurons and may have stochastic activation functions as well as learning can be done with similar gradient descent procedures for simple architectures and deterministic activation functions to those leading to the back-propagation algorithm for feed-forward network.

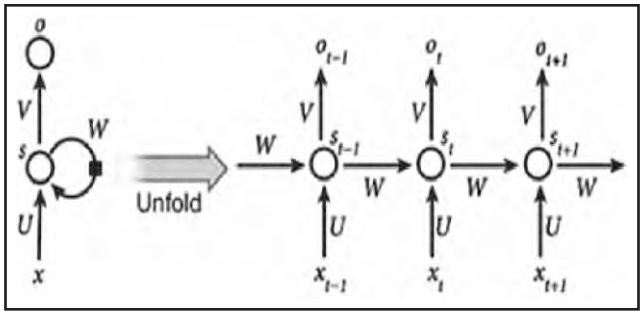

Fig. 3. Architecture of RNN

$$h_t = \tanh(W_x h x_t + W_{hh} h_{t-1} + b_h) \quad (3)$$

$$y_t = W_{hy} h_t + b_y \quad (4)$$

Where, $h(t)$ states the function f of the previous hidden state $h(t\text{-}1)$ the current input $x(t)$.

LSTM

Long Short Term Memory Networks commonly referred to as "LSTMs" are a special type of RNN that can learn long-term dependencies. LSTMs are designed mainly to avoid the problem of long-term dependence and it is completely based on the sequence-to-sequence labeling with context. RNN architecture in Fig. 3 addresses the vanishing / exploding gradient problem and it allows learning of long-term dependencies.

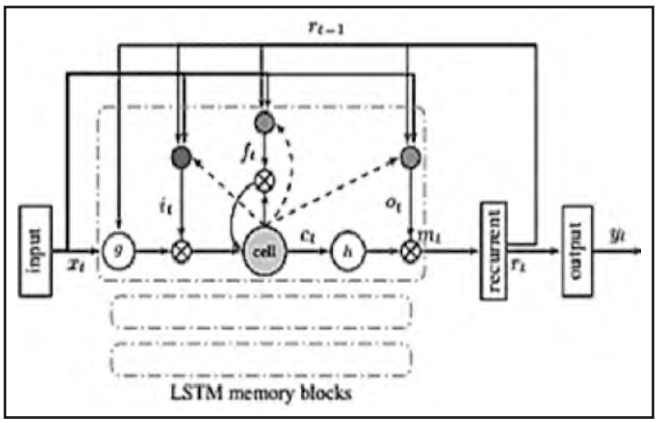

Fig. 4. LSTM Memory Cell Architecture

The LSTM memory cell has three gates as shown in Fig. 4. The first gate is a forget gate to decide what information to throw away from the cell state, this decision is made by a sigmoid layer.

$$f^{(t)} = \sigma(W^{(f)} x^{(t)} + U^{(f)} h^{(t-1)} + b^f) \quad (5)$$

The second gate is an input gate which consists of sigmoid layer to decide which values will be updated, and tanh layer which creates a vector of new updated values as described in (6) and (7).

$$i^{(t)} = \sigma(W^{(i)} x^{(t)} + U^{(i)} h^{(t-1)} + b^i) \quad (6)$$

$$\tilde{c}^{(t)} = \tanh(W^{(c)} x^{(t)} + U^{(c)} h^{(t-1)}) \quad (7)$$

Finally, the output of the current state will be calculated based on the updated cell state and a sigmoid layer which decides what parts of the cell state will be the final output

$$o^{(t)} = \sigma(W^{(o)} x^{(t)} + U^{(o)} h^{(t-1)} + b^o) \quad (8)$$

where σ is sigmoid activation function which squashes numbers into the range (0,1), tanh is hyperbolic tangent activation function which squashes numbers into the range (-1,1), W^f, W^i, W^c, W^o are the weight matrices, x^t is the input vector, h^{t-1} denotes the past hidden state and bf, b^i, b^c, b^o are bias vectors.

$$c^{(t)} = f^{(t)} \circ \tilde{c}^{(t-1)} + i^{(t)} \circ \tilde{c}^{(t)} \quad (9)$$

$$h^{(t)} = o^{(t)} \circ \tanh(c^{(t)}) \quad (10)$$

(9) and (10) indicates the final memory cell where first it takes the advice of the forget gate $f^{(t)}$ and accordingly forgets the past memory $c^{(t-1)}$. Then, it takes the advice of the input gate $i(t)$ and the new memory $\tilde{c}^{(t)}$. Finally, it sums these two results to produce the final memory $c^{(t)}$.

VII. IMPLEMENTATION

The proposed model uses LSTM-RNN as a model converts raw speech data into frame-level emotion features. While these recurrent connections are capable of memorizing previous information due to the vanishing gradient issue, the RNN still has a limitation on covering long background information such as emotions. To overcome this problem, a long-term memory (LSTM) network was introduced. Each block of the LSTM network contains three multiplicative self-connected memory cell gates such as (input, output, and forget). MFCC is used to determine the characteristics. A speech signal is first overtaken by a filter to amplify the energy of a speech signal at a higher frequency to measure the

MFCC function vector. It is then divided into images. The MFCC vector is used to extract the characteristics. Hence, various deep learning techniques are used to develop this model. Implementation of this model is done in python [10], where various python libraries are installed for converting the audio waves and sound files libraries are used to read and write the audio files.

LSTM – RNNs Speech based Emotion Recognition

RNNs have the advantage of studying complex temporal dynamics in sequential data compared with feed forward neural networks. Training RNNs to learn long-short term temporal dependency, however, can be difficult in practice due to the problem of gradient vanishing and exploding. The LSTM architecture provides a solution that partially overcomes simple RNNs weakness and achieves state-of-the-art speech recognition efficiency. More about the standard form of LSTM in a layer of projection to project the outputs of the memory cells to a lower-dimensional vector that is particularly effective for speech recognition where the amount of training is required. In this work, the same LSTM-RNN architecture is used as shown in Fig. 5.

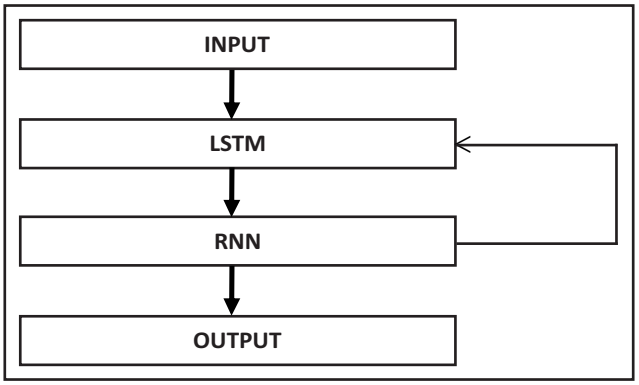

Fig. 5. LSTM-RNN Architecture

First, a raw audio file in sequences of characters at pre-processing stage transforms a raw audio file into multi-frame element vectors. First, divide each audio file with a 10ms overlap into 20ms Hamming windows and then measure the ceptral coefficients of 12 mel frequency, adding an energy factor to each frame. Speech signals involve several silent segments with less emotional information, which means that the emotional saturation between times is different. Weights can therefore be applied to the time dimension of the output of the LSTMs to distinguish the difference between layer outputs. Frames from adjacent windows are concatenated to form a spectrogram after computing all frames equally. This spectrogram acts as input features for the architecture of the RNN model. Assume that

$$X = \{(x(1), y(1)), (x(2), y(2)), ...\}$$

is a training set and a single utterance x with label y is sampling from this training set X. In this training set, every utterance $x(i)$ is a time series of $T(i)$ length. Here, the $T(i)$ time slice is a vector representation of audio features $xt(i)$ where t=1 to $T(i)$. In addition to propagation through time and LSTM layers, the input to the network at a given time stage passes through multiple LSTM layers. The layers in RNNs have allowed the network to learn about the input at different time scales. A softmax activation function is applied in hidden and output layers and training and testing is done for every utterance of audios.

VIII. EXPERIMENTAL RESULTS

The experimental results for EMODB and RAVDESS databases are trained and tested and the outputs for the sample audios are generated.

Results of EMO Database - German Language

In this model, a sample of 339 utterances is taken for processing. The training samples 271 out of 339 are trained. The testing samples 67 out of 339 are tested to obtain the output. In the developed model using EMODB, the outputs are obtained for various emotions of the given sample audio.

Results of RAVDESS Database- English Language

The training and testing audio files are seperated

TABLE II.
CLASSIFICATION ACCURACY FOR EMODB AND RAVDESS

Emotion	Accuracy for EMODB	Accuracy for RAVDESS
Sad	82.82	83.84
Happy	81.15	83.33
Angry	85.12	82.56
Surprise	82.05	83.07
Neutral	83.45	84.87

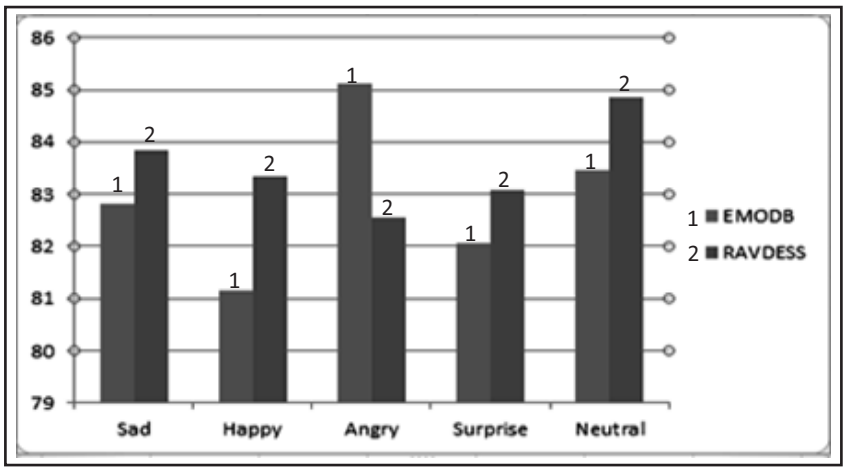

Fig. 6. The Accuracy of Different Emotion States

according to different emotions. There are about 808 training audio files present in the cateogry for angry emotion and the testing audio samples include 147. The sad emotion has 812 training audio files categorized and for testing 146 samples are used. The neutral emotion has about 585 training audio samples and for testing 93 audio samples are used. The accuracy results obtained for the various emotions like sad, happy, angry, surprise and neutral using RAVDESS and EMODB datasets are shown in Table II.

The happy emotion has about 805 training audio samples and for testing 146 audio samples are used in the model. Surprise emotion has 513 audio samples for the training data and about 77 audio samples are used for the training audio samples. Thus, various emotions like happy, sad, angry, neutral, surprise have been implemented in this model for English and German languages and the testing of the audio samples obtained from RAVDESS and EMODB database are categorized in the output for emotion prediction using audio signal as shown in Fig. 6.

IX. CONCLUSION

The accuracy of the developed model using RNN along with LSTM is listed in Table II. The model includes MFCC for feature extraction of speech. This model can be used in the field of healthcare to get the affective state of the patients with Autism and other voice-related fields.

X. FUTURE WORKS

This work can be extended considering the text in the speech and recognize emotion through Natural Language Processing (NLP). Further, a multi-model fusion system which integrates text, audio, video, and image based emotion recognition would improve the HCI.

ACKNOWLEDGMENT

The authors sincerely thank Department of Science and Technology (DST), Govt. of India Indo-U.S. Science & Technology Forum (IUSSTF) for granting the funds to carry out this collaborative research work under Indo-U.S. Fellowship for Women in STEMM (WISTEMM). They also thank Prof. Krishnaprasad Thirunarayan, Professor, Department of Computer Science and Engineering, Wright State University for his guidance during the fellowship period.

REFERENCES

[1] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, N. A. Mohamed, and H. Arshad, "State-of-the-art in artificial neural network applications: A survey," *Heliyon*, vol. *4, no.* 11, 2018.

[2] D. Kollias, M. Yu, A. Tagaris, G. Leontidis, A. Stafylopatis, and S. Kollias, "Adaptation and contextualization of deep neural network models," In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1–8). IEEE. doi: https://doi.org/10.1109/SSCI.2017.8280975

[3] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," In *Fifteenth annual conference of the international speech communication association*, pp. 223–227, 2014. Retrieved from

https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/IS140441 . pdf

[4] A. M. Badshah, N. Rahim, N. Ullah, J. Ahmad, K. Muhammad, M. Y. Lee, and S. W. Baik, "Deep features-based speech emotion recognition for smart affective services," *Multimedia Tools and Applications*, vol. *78,* pp. 5571–5589, 2017. Doi: https://doi.org/10.1007/s11042-017-5292-7

[5] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. *26, no.* 1, pp. 84 – 96, 2018. Doi: 10.1109/TASLP.2017.2761547

[6] X. Zhou, G. Junqi, and R. Bie, "Deep learning based affective model for speech emotion recognition," In *2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC*/ATC/ScalCom/CBDCom/IoP/SmartWorld), pp. 841 – 846. IEEE.

[7] T. Zhang, and j. Wu, "Speech emotion recognition with i-vector feature and RNN model," In 2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP), pp. 524 – 528, 2015. IEEE. Doi: https://doi.org/10.1109/ChinaSIP.2015.7230458

[8] K. Y. Huang, C. H. Wu, T. H. Yang, M. H. Su, and J. H. Chou, (2016, December). Speech emotion recognition u Chou, "Speech emotion recognition using autoencoder bottleneck features and LSTM," In *2016 International Conference on Orange Technologies (ICOT)*, pp. 1 – 4, IEEE. Doi: https://doi.org/10.1109/ICOT.2016.8278965

[9] S. An, Z. Ling, and L. Dai, "Emotional statistical parametric speech synthesis using LSTM-RNNs," In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1613–1616. IEEE. Doi:https://doi.org/10.1109/APSIPA.2017.8282282

[10] S. L. Rose, L. A. Kumar, and D. K. Renuka, *Deep learning using Python*, 2019.

About the Authors

Dr. Karthika Renuka D. is Associate Professor with Department of Information Technology, PSG College of Technology since 2004. Her areas of specialization include Soft Computing, Machine Learning and Deep Learning, Affective Computing, and Computer Vision. She has received research funding from UGC, AICTE & DST. She has published several papers in reputed national and international journals and conferences.

Akalya Devi C. is working as Associate Professor in Department of Information Technology, PSG College of Technology. She ha 1.5 years industry experience and 10 years of teaching experience. She is currently perusing part-time Ph.D in Anna University, Chennai.

Kiruba Tharani R. is pursuing M. Tech (final year) in Information Technology at PSG College of Technology. Her areas of interest include deep learning, artificial intelligence, and network security.

Pooventhiran G. is an Undergrad in Information Technology at PSG College of Technology, Coimbatore, India. He is a curious learner who craves learning and highly interested in research. His areas of interest include Machine learning and Graph Algorithms.