Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7766109

HUMAN HINDI SPEECH EMOTIONS IDENTIFICATION USING DEEP LEARNING ALGORITHMS

Mehul Patel[1,a],Dr. Amit Barve [1,b], Dr. Daxa Vekariya [1,c], Pro. Ankit Chauhan [1,d]

[a]Student of Computer Science Department, Parul Institute Of Engineering and Technology, Limda, Vadodara, Gujarat, India
[b] Head Of Computer Science Department, Parul Institute Of Engineering and Technology, Limda, Vadodara, Gujarat, India
[c] Associate Professor Of Computer Science Department, Parul Institute Of Engineering and Technology, Limda, Vadodara, Gujarat, India
[d] Assistant Professor Of Computer Science Department, Parul Institute Of Engineering and Technology, Limda, Vadodara, Gujarat, India
[a)]mehulrpatel123@gmail.com
[b)] barve.amit@gmail.com
[c)]daxa.vekariya18436@paruluniversity.ac.in
[d)]ankit.chauhan@paruluniversity.ac.in

ABSTRACT--Speech emotion recognition is a technique used to identify and understand the emotions conveyed in spoken language. Deep learning, a subset of machine learning, has been proven to be an effective method for speech-emotion recognition. This is due to its ability to learn and improve upon a wide range of features in speech, such as MFCC, pitch, intonation, and rhythm. In this abstract, we present a deep learning approach to speech emotion recognition, which utilizes a combination of recurrent neural networks (RNNs) and convolutional neural networks (CNNs) to accurately classify emotions in speech. The Speech Corpora dataset is used in speech emotion Identification and gets a Taining Accuracy of 86.25% For the Recurrent Neural Networks (RNN) Model. This paper aims to explore the use of deep learning techniques for speech emotion recognition. We begin by reviewing the current state of the field, highlighting recent advances and key challenges. We then describe the deep learning models that have been most commonly used for speech emotion recognition, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). We discuss the advantages and limitations of these models and the various pre-processing and feature extraction methods that have been proposed. Finally, we present recent results and evaluate the performance of deep learning-based speech emotion recognition methods, and conclude with some future directions for the field.

Keywords: Emotion Recognition, Deep Learning, Recurrent Neural Network (RNN), LSTM.

I. INTRODUCTION

Speech emotion recognition is the process of recognising and detecting emotions in spoken language. It is a quickly growing field of study with numerous potentials uses in industries like virtual assistants, human-computer interaction, and healthcare. Speech contact is heavily

influenced by emotion [1]. One of the most effective methods for recognising speaking emotions is the use of deep learning techniques. These methods, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have been shown to be very effective at capturing the complex patterns and subtle features of human speech. By using large datasets of labelled speech samples to train these models, researchers are able to recognise emotions like happiness, anger, and sadness with high levels of accuracy.

Speech is the most efficient way for people to communicate. This encourages researchers to closely examine speech data and pinpoint distinctive speech characteristics. Both content and emotions are evident when a person speaks. Emotion has a significant impact on speech transmission [1]. Such an emotion identification system can be used in call centres, to anticipate tiredness, in smart homes, and in smart workplaces. A increasing application for such a system has been noted in the counselling of patients with mental health issues [2]. If he has no feelings, he might have mental health problems. As a result, emotions play a bigger role in a person's life. Everybody feels the six fundamental feelings of anger, disgust, fear, happiness, sadness, and surprise. There are some feelings that are constant among all of these different emotions, such as two- or three-dimensional spaces with valence, arousal, and dominance dimensions. Emotions are necessary for individuals to be in good physical and psychological health. Grief and anger are examples of negative emotions that are connected to mental health problems that, if left untreated, can have disastrous outcomes like self-harm or suicide. Utilizing speech emotion detection is one choice (SER). Mental health and healthcare are analysed with a view to prescribing and managing people who are not depressed [4]. Emotion Subjective and objective assessment techniques fall into two groups. The subjective can gauge feelings through a questionnaire and interpersonal interaction. It is possible to determine how individuals are feeling by using physiological cues.

The methods for emotion recognition using a recurrent neural network are mostly described in this article using machine learning and deep learning techniques (RNN).

Men and women between the ages of 18 and 40 are chosen to participate in the data collection process. Participants must be psychologically stable because only then can emotions be accurately recorded. More than 50 people are needed to gather the data. Individuals who watch emotionally charged movies experience emotional responses. Obtain the physiological signals using wearable devices. The EEG explains the current state of the cerebral activity. To detect the signals, the device is therefore placed on the scalp. To record emotional activity like happy and unhappy emotions, ECG places a wearable on the left ankle, right wrist, and both wrists. Placing a wearable on your middle and index fingers will allow you to effectively record your emotional states. GSR provides an explanation for the conductivity of electricity through epidermis. To obtain RSP, wear the device on your chest while you record your breathing activity [29].

The feature extraction primarily behaves differently for distinct signals. Signals are processed to derive properties such as Pitch Value, Energy, Zero Crossing Rate (ZCR), Mel frequency Cepstral Coefficients (MFCC), and others [1].

Different dataset types are used for the classification of feelings. The datasets for emotion analysis using physiological signals (DEAP), the SJTU emotion EEG dataset (SEED), the MANHOB dataset, and the TYUT 2.0 EEG-based dataset are the main datasets used for

emotion detection. choice graphs,When used to describe the emotions in this SVM, the K-nearest neighbours (K-NN), random forests, multi-dimensional dynamic weighted temporal warping (WMD-DTW), and decision trees algorithms. Excellent classification accuracy across a range of methods.

The accuracy of each strategy depends on the techniques used for feature extraction, classification,and the different dataset types. Here, a variety of methods are used to increase classification accuracy by up to 94 percentage points. These various emotions can be recognised, which makes them helpful for both medical and human-computer interaction applications.

II. RELATED WORK

Speech emotion recognition using deep learning has been an active research topic in the field of artificial intelligence and signal processing.Emotional analysis from speech has been attracted various researchers to work in this field. Initially study were happened using statistical properties from acoustic features [1]. neural networks and deep learning have also been employed due to their ability to learn hierarchical high-level representations from raw audio features [2]. There are numerous related works that have been published in recent yearsWhen compared to conventional machine learning techniques, deep learning techniques have significantly improved the efficacy of speech emotion recognition. However, in order for speech emotion detection systems to be used in real-world scenarios, issues like dataset variability, model complexity, and real-time processing must be resolved.

Studies on Speech Emotion Recognition (SER), incorporating various features and methodologies, have been conducted over the past ten years. Historically, based on how the target label was assigned, either short segments or individual utterances were used. The Low-Level Descriptors (LLD) of speech are extracted from each individual segment of an utterance and fed to a sequential classifier using segment level methods. This emotional condition of a speech has been modelled using classifiers like the Hidden Markov Model and the Gaussian Mixture Model.

III. METHODOLOGY

A. Hindi Emotional Speech Databases

There are numerous emotional speech corpora available for research on voice emotion recognition in a number of languages. The speech corpus created by Agrawal, A., Jain, A., et al. for Hindi speech emotion research is the Indian Hindi Speech Corpus [1]. The experimental analysis uses a portion of a database made up of performed Hindi emotional speech utterances that were recorded with the assistance of 16 Actors in India. The actors are between the ages of 15 and 25. Five Hindi sentences are chosen, with syllable counts ranging from 11 to 18, and word counts from 3 to 6. Then, within a single session, every actor delivers each sentence while expressing a different feeling. Angry, Happy, Neutral, Sad, and surprised are the five fundamental emotions that were taken into account in this study. On different days, five of these sessions are repeated and recorded. The entire recording is done in one spot, in a silent room. The 16 kHz sample rate is used to record the voice signal, which is then represented as a 16-bit number.

Table1: Description of speech corpora Dataset

Sr.No	Specification of database	Statistics of emotion
1	Speakers	16
2	Sentences	5
3	Emotions	Anger, Happy, Neutral, Sad, Surprise
4	Sentence type	Hindi
5	Speech Coding	16khz
6	Sampling Rate	44.1KHz
7	Total Sentences	400

B. Data Pre-processing

Prepare the dataset by gathering and labeling speech data samples, with each sample associated with its corresponding emotion label. Before feature extraction, some audio pre-processing is used to create a robust system. The signal is split into 20 ms frames at a sampling rate of 16 kHz, with a 10 ms overlapped hamming window between each pair of adjacent speech frames.

C. Feature Extraction

Prosodic features are distinctive aspects of speaking that mirror the auditory characteristics of sounds. The following speech characteristics are extracted in this work:

Mel Frequency Cepstral Coefficients (MFCC): It is the perception of frequency's counterpart. After pre-processing, in order to obtain MFCC Calculate the power spectrum: The power spectrum is calculated for each segment using a Fourier transform.

Mel-scale filtering: The power spectrum is then filtered through a set of overlapping triangular filters that are spaced in a non-linear, mel-scale frequency axis.

Logarithmic compression: The logarithm of the filterbank energies is computed to approximate the human perception of loudness.

Discrete cosine transform: The Discrete Cosine Transform (DCT) is applied to the logarithmic filterbank energies to obtain a set of cepstral coefficients.
Selecting the coefficients: The resulting MFCC coefficients can then be further processed or used directly for tasks such as speech recognition or music genre classification.

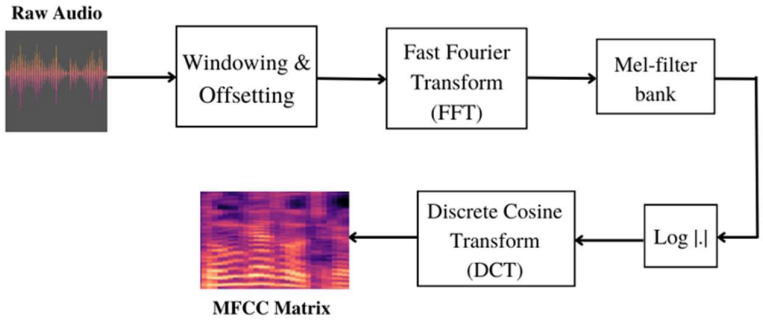

Figure. 1: Block diagram for computing MFCC.

Following are the MFCC Graph for the Different Five Emotions like angry, happy, neutral, sad, and surprise.

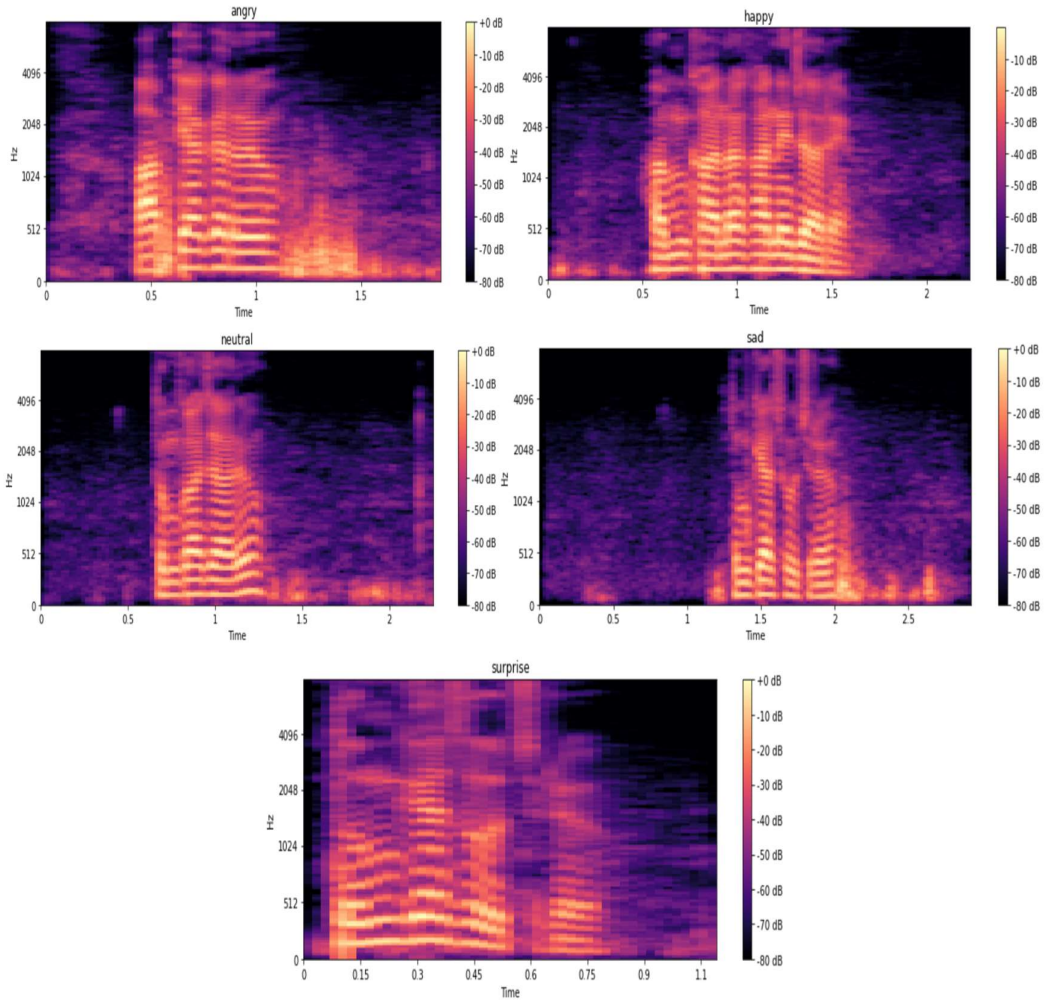

Figure. 2:MFCC features Of Different Emotions (angry, happy, neutral, sad, surprise).

D. Classification

Recurrent Neural Networks (RNNs) have shown promising results in speech emotion recognition tasks due to their ability to model temporal dependencies in speech signals. Due to its capacity to simulate temporal dependencies, a recurrent neural network (RNN) is a particular kind of neural network that works well for handling sequential input, like speech signals. For classification, the Recurrent neural network has been proposed. models have been developed. Based on MFCC data, a classifier model is trained. Emotions are divided into five classes for the purpose of emotion identification, which is a multi-class problem: happy, sad, angry, neutral, and surprised state. For emotion recognition, the train and test of the recurrent neural network (RNN) LSTM sequential model are considered.

The LSTM has links to feedback. Such a recurrent neural network (RNN) is capable of processing not only single data points (such as images), but also complete data sequences (such as speech or video). Because of this quality, LSTM networks are perfect for handling and making predictions about data.

The most widely employed deep learning models for speech emotion detection (CNNs, RNNs, etc.).

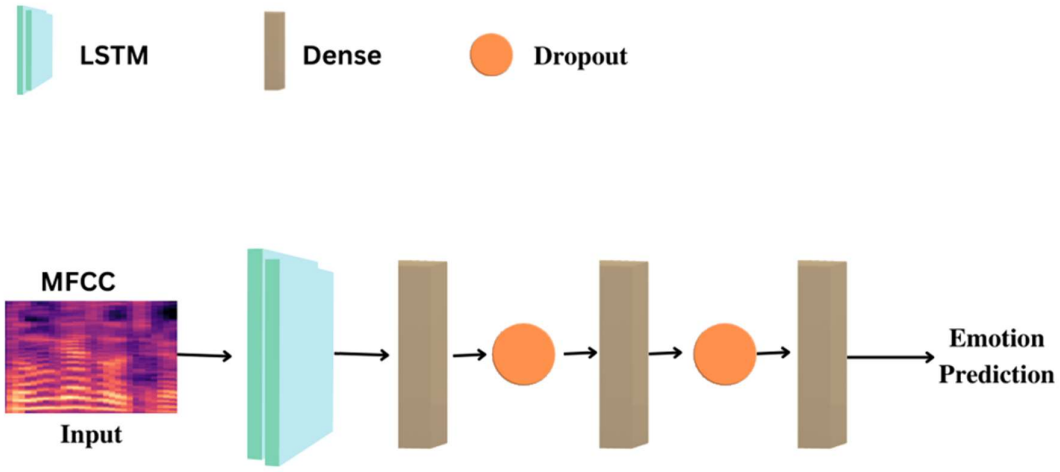

Figure. 3: Proposed Architecture off RNN-LSTM Sequential Model.

IV. RESULTS

We proposed an RNN model with an attention mechanism for speech emotion recognition. The model was trained and tested on Speech Corpora dataset, which contains speech recordings of actors in various emotional states.

We Obtained an accuracy of 89% on the Speech Corpora dataset using their proposed RNN model with attention mechanism. This outperformed several baseline models, including Decision Tree (DT), Random Forest (Rf), and simple RNN models.

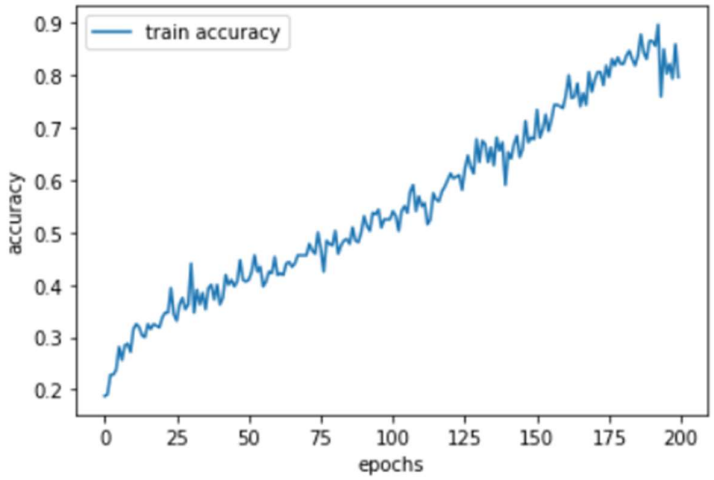

Figure. 4: Proposed Model RNN trainingAccuracy

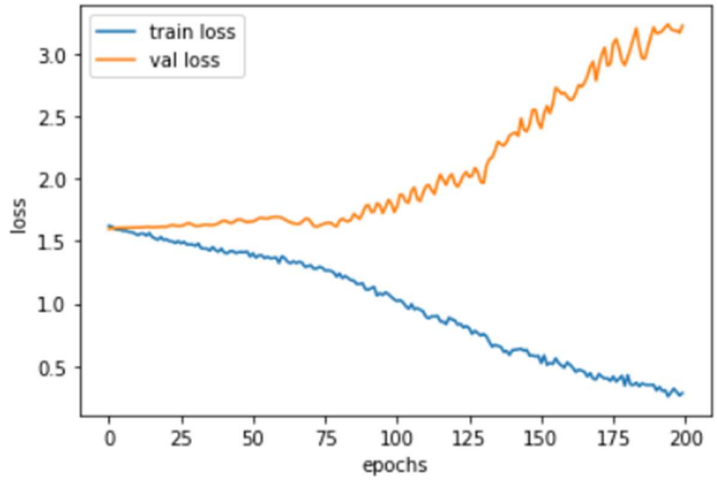

Figure.5: Proposed Model RNN training and validation Loss.

Table2: Comparison Of Proposed Model Accuracy

Model	RNN	Decision Tree	Random Forest
Accuracy	89%	28%	41%

Table3: Comparative AnalysisOfExisting Techniques

Author Name	**Year**	**Methods**	**Accuracy**	**Dataset**
Proposed Work	2023	RNN, DT, RF	89%, 28%, 41%	Speech Corpora
Agrawal, A., & Jain, A. et al. [1]	2020	KNN, NBC, PCA,	87%	Speech Corpora

		LDA		
Mishra, P., & Sharma, R. et al. [2]	2020	MFCC, CNN, SVM	Mean accuracy=92.28%	RAVDESS CREMA-D SAVEE TESS
Gupta, V., Juyal, S.et al.[3]	2020	Random Forest, CNN, SVM, MLPC, Decision Tree	76%, 91.1%, 50%, 45.56%, 79.9%	RAVDESS
El Seknedy, M., & Fawzi, S. et al. [10]	2021	MLP, SVM, Random Forest, Logistic Regression	79.26%, 88.24%, 82.35%.	Corpora
Verma, D., Mukhopadhyay et al. [30]	2016	KNN MLP SVM	68.3% 74.73% 77.57%	Hindi Emotional Speech Databases

V. MODEL EVALUATION METRICS

In machine learning, model evolution metrics are used to assess and contrast the performance of various models during the training and testing stages. Listed below are a few typical measures.

Precision

Out of all the positive predictions the model makes, precision is the percentage of accurate positive predictions. When the cost of false positives is significant, it is helpful.

$$\text{PRICISION} = \frac{\text{TRUE\_POSITIVE}}{(\text{TRUE\_POSITIVE} + \text{FALSE\_POSITIVE})} \quad (1)$$

Recall

Recall quantifies the percentage of accurate positive predictions among all occurrences of positive data that actually occurred. When the cost of false negatives is significant, it is helpful.

$$\text{RECALL} = \frac{\text{TRUE\_POSITIVE}}{(\text{TRUE\_POSITIVE} + \text{FALSE\_NEGATIVE})} \quad (2)$$

F1-score

The harmonic mean of recall and precision is the F1 number. In cases where the dataset is unbalanced, it is a more balanced measure than accuracy.

$$\text{F1 SCORE} = \frac{2 * \text{PRECISION} * \text{RECALL}}{\text{PRECISION} + \text{RECALL}} \quad (3)$$

Accuracy

Accuracy is a measure of the model's success across all classes. It is advantageous when all types are similarly significant. It is determined by dividing the overall percentage by the amount of precise forecasts.

$$\text{ACCURACY} = \frac{\text{TP+TN}}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \quad (4)$$

VI. CONCLUSION

Speech emotion recognition software now includes powerful deep learning methods. Researchers have been able to recognise emotions like happiness, anger, and sorrow with high levels of accuracy by training models like CNNs and RNNs on large datasets of labelled speech samples. Speech emotion recognition has a wide range of possible applications, including virtual assistants, human-computer interaction, and healthcare. Even so, there is still room for improvement, particularly with regard to adapting the model to various populations and addressing linguistic differences. Deep learning-based speech emotion recognition is probably going to become an increasingly important tool for comprehending and interpreting human feelings in spoken language as the field develops. Our proposed RNN-LSTM sequential model has a success rate of 89%. The deep learning algorithms are more accurate than the machine learning algorithms, according to the results we obtained.
In Feature We Can Add Gender Classification and Add new Emotions.

VII. ACKNOWLEDGEMENT

The Hindi Speech Corpora dataset [1] for experimentation was provided by Akshat Agrawal, who offered support for this study.

REFERENCES

1. Agrawal, A., & Jain, A. (2020). Speech emotion recognition of Hindi speech using statistical and machine learning techniques. Journal of Interdisciplinary Mathematics, 23(1), 311-319.
2. Mishra, P., & Sharma, R. (2020, October). Gender-differentiated convolutional neural networks for speech emotion recognition. In 2020 12th International Congress on UltraModern Telecommunications and Control Systems and Workshops (ICUMT) (pp. 142-148).

3. Gupta, V., Juyal, S., Singh, G. P., Killa, C., & Gupta, N. (2020). Emotion recognition of audio/speech data using deep learning approaches. Journal of Information and Optimization Sciences, 41(6), 1309-1317.
4. Lieskovska, E., Jakubec, M., & Jarina, R. (2022, April). RNN with Improved Temporal Modeling for Speech Emotion Recognition. In 2022 32nd International Conference Radioelektronika (RADIOELEKTRONIKA) (pp. 1-5) IEEE.
5. Ho, N. H., Yang, H. J., Kim, S. H., & Lee, G. (2020). Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network. IEEE Access, 8, 61672-61686.
6. Sajjad, M., & Kwon, S. (2020). Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. IEEE Access, 8, 79861-79875.
7. Yang, S., Gong, Z., Ye, K., Wei, Y., Huang, Z., & Huang, Z. (2020). EdgeRNN: a compact speech recognition network with Spatio-temporal features for edge computing. IEEE Access, 8, 81468-81478.
8. Sun, T. W. (2020). End-to-end speech emotion recognition with gender information. IEEE Access, 8, 152423-152438.
9. Braunschweiler, N., Doddipatla, R., Keizer, S., &Stoyanchev, S. (2022). Factors in Emotion Recognition With Deep Learning Models Using Speech and Text on Multiple Corpora. IEEE Signal Processing Letters, 29, 722-726.
10. El Seknedy, M., & Fawzi, S. (2021, December). Speech Emotion Recognition System for Human Interaction Applications. In 2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS) (pp. 361-368). IEEE.
11. Shanta, S. S., Sham-E-Ansari, M., Chowdhury, A. I., Shahriar, M. M., & Hasan, M. K. (2021, September). A Comparative Analysis of Different Approach for Basic Emotions Recognition from Speech. In 2021 International Conference on Electronics, Communications and Information Technology (ICECIT) (pp. 1-4). IEEE.
12. Bharti, D., &Kukana, P. (2020, September). A hybrid machine learning model for emotion recognition from speech signals. In 2020 International Conference on Smart Electronics and Communication (ICOSEC) (pp. 491-496). IEEE.
13. Matin, R., & Valles, D. (2020, October). A speech emotion recognition solution-based on support vector machine for children with autism spectrum disorder to help identify human emotions. In 2020 Intermountain Engineering, Technology and Computing (IETC) (pp. 1-6). IEEE.
14. Lee, K. H., Choi, H. K., & Jang, B. T. (2019, October). A study on speech emotion recognition using a deep neural network. In 2019 International Conference on Information and Communication Technology Convergence (ICTC) (pp. 1162-1165). IEEE.
15. Zhang, H., Huang, H., & Han, H. (2020). Attention-based convolution skip bidirectional long short-term memory network for speech emotion recognition. IEEE Access, 9, 5332-5342.
16. Lian, Z., Liu, B., & Tao, J. (2021). CTNet: Conversational transformer network for emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 985-1000.

17. Bahreini, K., Nadolski, R., &Westera, W. (2016). Data fusion for real-time multimodal emotion recognition through webcams and microphones in e-learning. International Journal of Human–Computer Interaction, 32(5), 415-430.
18. Han, J., Zhang, Z., Ren, Z., & Schuller, B. (2019). EmoBed: Strengthening monomodal emotion recognition via training with crossmodal emotion embeddings. IEEE Transactions on Affective Computing, 12(3), 553-564.
19. Avila, A. R., Akhtar, Z., Santos, J. F., O'Shaughnessy, D., & Falk, T. H. (2018). Feature pooling of modulation spectrum features for improved speech emotion recognition in the wild. IEEE Transactions on Affective Computing, 12(1), 177-188.
20. Gideon, J., McInnis, M. G., & Provost, E. M. (2019). Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG). IEEE Transactions on Affective Computing, 12(4), 1055-1068.
21. Yi, L., &Mak, M. W. (2020). Improving speech emotion recognition with adversarial data augmentation network. IEEE transactions on neural networks and learning systems.
22. Xiao, Y., Zhao, H., & Li, T. (2020). Learning class-aligned and generalized domain-invariant representations for speech emotion recognition. IEEE Transactions on Emerging Topics in Computational Intelligence, 4(4), 480-489.
23. Luo, H., & Han, J. (2020). Nonnegative matrix factorizationbased transfer subspace learning for cross-corpus speech emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 2047-2060.
24. Oliveira, J., &Praça, I. (2021). On the usage of pre-trained speech recognition deep layers to detect emotions. IEEE Access, 9, 9699-9705.
25. Parthasarathy, S., &Busso, C. (2020). Semi-supervised speech emotion recognition with ladder networks. IEEE/ACM transactions on audio, speech, and language processing, 28, 2697-2709.
26. Zhao, H., Ye, N., & Wang, R. (2020). Speech emotion recognition based on hierarchical attributes using feature nets. International Journal of Parallel, Emergent and Distributed Systems, 35(3), 354-364.
27. Song, P., Zheng, W., Yu, Y., &Ou, S. (2020). Speech emotion recognition based on robust discriminative sparse regression. IEEE Transactions on Cognitive and Developmental Systems, 13(2), 343-353.
28. Fonnegra, R. D., & Díaz, G. M. (2017, December). Speech emotion recognition based on a recurrent neural network classification model. In International Conference on Advances in Computer Entertainment (pp. 882-892). Springer, Cham.
29. Joy, E., Joseph, R. B., Lakshmi, M. B., Joseph, W., & Rajeswari, M. (2021, March). Recent survey on emotion recognition using physiological signals. In 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 1858-1863). IEEE.
30. Verma, D., Mukhopadhyay, D., & Mark, E. (2016, August). Role of gender influence in vocal Hindi conversations: A study on speech emotion recognition. In 2016 International Conference on Computing Communication Control and automation (ICCUBEA) (pp. 1-6). IEEE.