International Journal for Modern Trends in Science and Technology
Volume 12, Issue 05, pages 157-164
ISSN: 2455-3778 online
Available online at: http://www.ijmtst.com/vol12issue05.html
DOI: https://doi.org/10.5281/zenodo.19836528

Automated Emotion Recognition from Speech Signals using Deep Learning

G. Iswarya, G. Sai Veera Sri Santosh, G. Anusha, G. Vyuhita, G. Chanikya Durga Vara Prasad, B. Madhav Rao

Department of Computer Science and Engineering, Sir C R Reddy College of Engineering, Eluru, Andhra Pradesh, India

To Cite this Article
G. Iswarya, G. Sai Veera Sri Santosh, G. Anusha, G. Vyuhita, G. Chanikya Durga Vara Prasad & B. Madhav Rao (2026). Automated Emotion Recognition from Speech Signals using Deep Learning. International Journal for Modern Trends in Science and Technology, 12(05), 157-164. https://doi.org/10.5281/zenodo.19836528

Article Info
Received: 28 March 2026; Revised: 24 April 2026; Accepted: 26 April 2026.

KEYWORDS

Speech Emotion Recognition, Deep Learning, CNN-LSTM, MFCC, wav2vec 2.0, Audio Signal Processing

ABSTRACT

The growing need of smart systems able to comprehend human emotions has brought about major development on speech-based analysis. Human speech has a deep emotional content and this can be utilized to improve the human-computer interaction, communication systems and behaviour analysis. Proper recognition of emotions based on speech cues can result in more adjustive and receptive technologies. This paper describes a deep learning-based speech-based automatic emotion recognition system. The suggested system identifies acoustic features like Mel-Frequency Cepstral Coefficients (MFCCs) of audio input through which important features of human voice are extracted. It uses a hybrid Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) with network to extract both spatial and temporal patterns within speech signals to successfully classify emotions. Besides the hand-crafted features, the system also includes representations of wav2vec 2.0 which obtains high-level contextual embedding's out of raw audio. These pre-trained features are complementary to CNN-LSTM pipeline and help to acquire more semantic and acoustic information that might not be entirely covered by MFCCs. Combining these complementary feature representations helps the model better classify emotions in a variety of speech conditions. The system is trained and tested with the help of the RAVDESS, TESS, SAVEE datasets that comprises various samples of emotional speech. The model has been proved to be highly accurate in the classification of several emotional states according to the results of the experiments. The proposed approach highlights the appropriateness of deep learning algorithms in speech emotion recognition and demonstrates its potential application in the intelligent system and interactive technologies.

1. INTRODUCTION

The world today is undergoing rampant growth in the psychological pressure with the rapid advancement in technology, changing relations in the social sphere, and the growing work demand towards people.

Although a moderate amount of pressure has the ability to promote performance and motivation, in the long-term, it may harm emotional stability, cognitive functioning, and well-being. Previous

researches have revealed that chronic mental stress correlates with many health risks such as anxiety disorders, depression, sleep disorders, and heart diseases [4][13].

To overcome these issues, finding effective automated detection systems has become a significant field of study in the fields of healthcare and artificial intelligence [10]. Conventional assessment techniques mostly depend on qualitative methods like questionnaires and clinical interview and physiological measurements like heart rate variability, skin conductance measure and brain activity indicators [13]. Despite being useful methods, the techniques are commonly complicated, involve specialized machinery, and involve the participation of experts, make them less applicable to continuous and real-time monitoring.

Speech has recently become a non-invasive behavioural cue that can be used to study the emotional state of man [8][9]. In comparison with physiological measurements, speech is easy to record without the need to have extra sensors and intrusion. Human voice holds a lot of acoustic data that reveals the psychological and emotional states of being. Differences in vocal components e.g. pitch, intensity, speed of speech, rhythm, articulation patterns etc. also act as valuable cues that can be used to determine emotional states. These features make it possible to create smart systems that can automatically interpret speech cues to identify the emotions with a high level of accuracy. Moreover, develops in deep learning have boosted the capacity to learn meaningful speech data patterns. The complex relationships between the acoustic features and emotional states can be learned effectively with the help of modern models and allow more accurate and scalable emotion recognition systems. Consequently, speech-based analysis has emerged as an effective method of constructing real-time, efficient and accessible intelligent systems of understanding human behaviour.

I. RELATED WORK

Human emotional state detection is a significant field of research because of its use in monitoring of healthcare, productivity at work, and in human computer interaction systems. One of the avenues of study has been on speech signals as a potential tool in detecting emotional and psychological states. Initial studies were mainly based on classic machine learning methods with manually designed acoustic features based on speech cues [4], [13].

The earliest methods were centered on the extraction of features like Mel-Frequency Cepstral Coefficients (MFCC), pitch and energy that were later classified using classical classifiers like Support Vector Machines (SVM), k-Nearest Neighbors (KNN) and Decision Trees to classify the emotional states [1], [4].

CNNs have been successfully used in spectrogram representations of speech signals, whereby spatial and frequency-based features can be extracted. It has been demonstrated that CNN-based models are much more effective than traditional machine learning methods in emotion recognition tasks [3], [11]. Alongside CNNs, Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) networks have also proven to be able to model sequential dependencies in speech signals with great ability [5]. LSTM models are powerful to capture temporal variations of audio information, which is important to perceive emotional dynamics in speech. Moreover, the Bidirectional LSTM architectures are also better because they process sequences in both directions; this increases the contextual understanding [10].

Access to benchmark datasets like RAVDESS and IEMOCAP has also enabled the development of speech emotion recognition, as they are a variety of emotional speech sample to be trained and evaluated [2], [14]. Also, other tools like openSMILE and libraries like Librosa have been instrumental in deriving meaningful acoustic features to deep learning models [8], [12].

Continuing on these advances, in more recent work, hybrid deep learning models have been investigated, which are a combination of CNN and LSTM networks, to successfully encode both spatial and temporal information in speech signals. These methods have proven to be more precise and resilient at recognizing emotions [6], [3]. This paper presents a speech-based emotion recognition system based on the use of deep learning. The framework involves the MFCC feature extraction and a hybrid CNN-LSTM network to process audio samples and rank the emotional states. This is done to improve the functioning of automated

emotion recognition systems and shows the efficiency of deep learning method in speech analysis in real time.

II. PROPOSED ALGORITHM

The proposed system introduces a deep learning-based speech emotion recognition framework that uses acoustic cues that are derived based on audio signals. It examines these attributes to adequately categorize speech under various emotional categories. In the general form of the system, the system is organized into a few main steps, namely, data acquisition, preprocessing, feature extraction, model training, and final emotion classification, which provides a well-structured and effective processing pipeline.

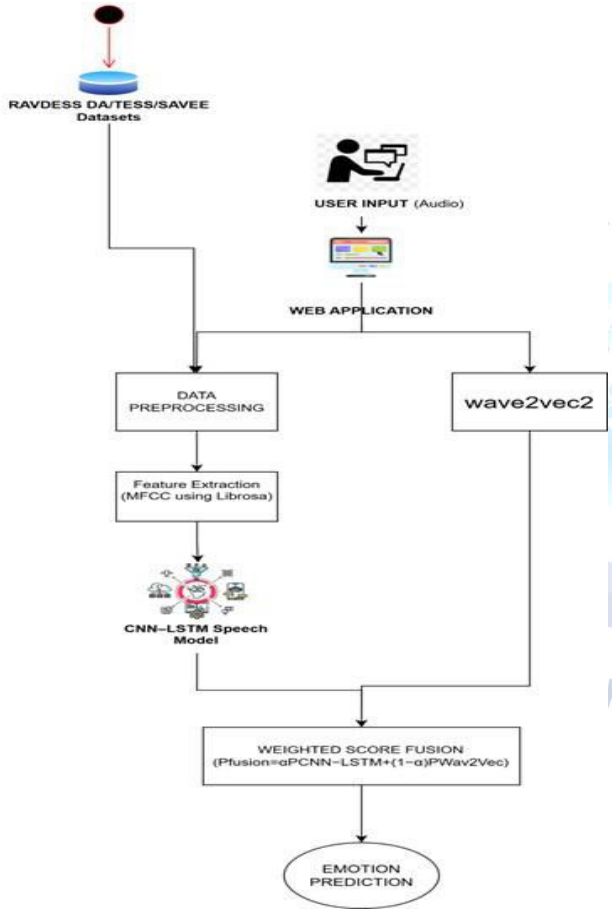

Fig. 1. System Architecture

A. Dataset Description

The datasets involved in the research are RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song), TESS (Toronto Emotional Speech Set) and SAVEE (Surrey Audio-Visual Expressed Emotion). The RAVDESS dataset is made up of 1440 audio samples of 24 speakers, with 8 different categories of emotion. TESS dataset consists of about 2800 audio samples of 2 speakers of 7 emotional classes whereas SAVEE consists of about 480 samples of 4 speakers of 7 emotional classes. All three datasets are presented in the form of WAV files and have high-quality and noise-controlled samples of emotional speech.

All these datasets provide variety in the speakers, accents, and emotional manifestations, which is needed to create a powerful emotion recognition system. The difference in gender, speaking style, and strength of emotions also contribute to the richness of the data. Also, the use of several datasets will reduce overfitting and enhance the overall generalization capability of the model with data that are not seen. This variety allows the model to acquire more realistic acoustic patterns and has a positive effect on its performance in the real world.

Table 1. Speech Datasets Used in the Proposed System

Dataset	Number of Samples	Speakers	Emotional Classes	Format
RAVDESS	1440	24	8 emotions	WAV
TESS	2800	2	7 emotions	WAV
SAVEE	480	4	7 emotions	WAV

B. Data Preprocessing

Background noises and silence, amplitude and duration variations are all part of the raw audio signals and can severely deteriorate the performance of speech based models. Such inconsistencies add undesirable variability and decrease the reliability of the extracted features. To address these concerns, the audio data is preprocessed with a comprehensive pipeline to enhance the overall level and uniformity of the audio data prior to feature extraction.

The initial is to cut down on background noise using a pre-emphasis filter, a form of high-pass filtering which seeks to trim the low-frequency signal but enhance the high-frequency information that is important to speech. The step enhances signal clearness and intelligibility, which elevate key speech characteristics and bring them more to the fore and can be processed downstream.

After noise reduction, a peak normalization is used to normalize the amplitude to make sure that all audio samples share the same signal intensity. This avoids bias due to variance in recording volume and enhances stability of the model. In addition to this any audio signal is resampled to a common sampling rate of 22,050 Hz band-limited interpolation, such that it can be compatible with other datasets and share the same time resolution.

The elimination of silence is then implemented in order to eliminate non-informative parts of the audio signal. This is carried out by an energy-based thresholding method, i.e., decibel (dB) trimming, which only retains those

line-elements that contain some interesting speech data. The model is able to drop unnecessary silent parts and concentrate more on the emotional cues in the speech that are important.

Finally, the audio samples are zero-padded to identical length to make shorter signals the same length, and truncated to make longer signals the same length to make the model the same size. In addition, every audio file is converted to the same format of WAV that is standardized in the aspects of encoding in order to allow easy processing and reliable extraction of features. The combination of these preprocessing steps reduces variability, signal quality and robustness, stability and workability of the proposed speech emotion recognition system.

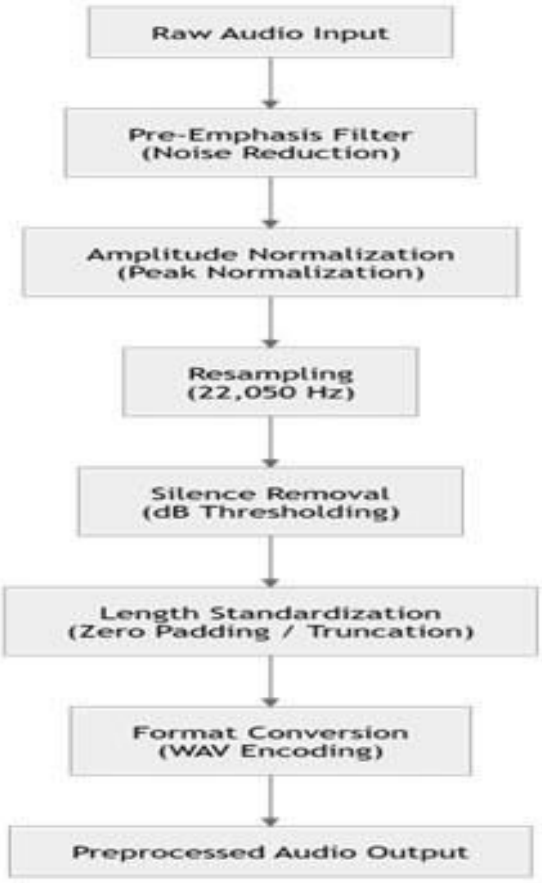

Fig. 2. Audio preprocessing pipeline.

C. Audio Feature Extraction

The preprocessed audio signals are extracted into meaningful representations through feature extraction to facilitate the learning of meaningful features by the model. Out of the several acoustic features that exist, Mel-Frequency Cepstral Coefficients (MFCCs) are mostly used because of their capacity to preserve the perceptually significant spectral attributes of speech signals.

It starts with the extraction process that separates the audio signal into small overlapping frames, and a windowing function is applied to minimize spectral leakage. Each framed signal is then converted to the frequency domain via the Fourier Transform. A frequency response is then filtered through a Mel-scaled filter bank that condenses frequency content according to human hearing. The filter bank energies are then subjected to a logarithmic operation, followed by a Discrete Cosine Transform (DCT) to result in a small set of MFCC features.

Besides MFCCs, spectrograms of the audio signals capture time-frequency representations of the signal and give a visual and quantitative representation of the changing frequency content of the signal over time. These representations are used to complement MFCC features and keep the variations of time and enrich the information extracted.

The extracted MFCC features are then normalized into a fixed size representation of 130 temporal frames and 40 cepstral coefficients. It guarantees consistency in all samples, no matter their initial length. Lastly, the features are transformed into a four-dimensional tensor, which can be used as input to deep learning models, especially convolutional and recurrent architectures.

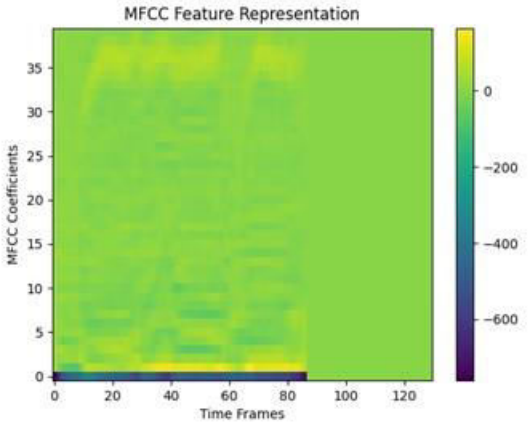

Fig. 3. MFCC or spectrogram of speech signal.

D. Deep Learning Model Architecture.

The suggested system has a hybrid CNN-LSTM architecture that identifies spatial and temporal features of speech signals. CNN layers learn local patterns of MFCC inputs, and LSTM layers learn sequential relationships in the data. The model input comprises of MFCC features of dimension (130 x 40 x 1) that are initially subjected to convolutional layers to acquire spatial representations.

The initial layer (Conv2D) has 32 (3 × 3) filters with ReLU activation and a MaxPooling layer reduces the dimension.

A second Conv2D with 64 filters is implemented and another MaxPooling layer is applied to ensure a further refined features. The resulting feature maps are reshaped into a sequential format and fed into LSTM layers, with a 128-unit LSTM layer (return sequences) and a 64-unit LSTM layer to learn temporal features.

Lastly, the extracted features are fed into a Dense layer of 64 units and a Dropout layer is implemented to minimize overfitting. The model gives the result of the predicted emotion according to the learned representations.

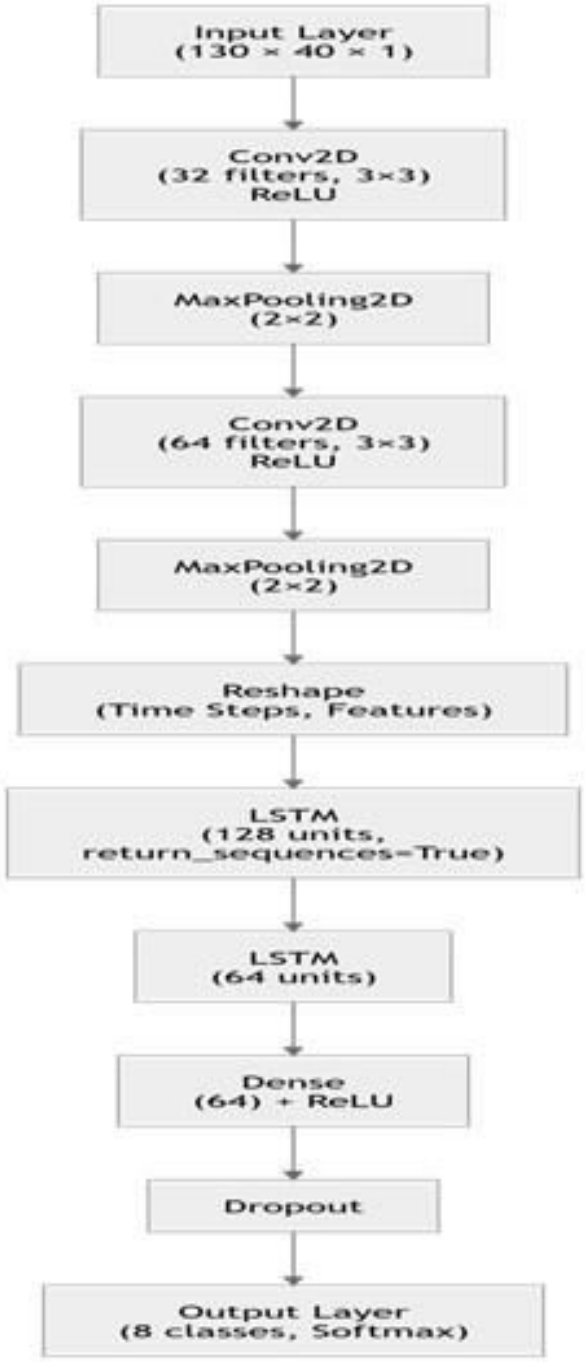

Fig. 4. CNN-LSTM model of speech emotion recognition.

E. Wav2Vec2 Based Feature representation

Besides the CNN-LSTM architecture, the suggested system has a Wav2Vec2-based model that learns to extract high-level contextual representations directly out of raw audio signals. Wav2Vec2, unlike conventional feature extraction models like MFCC, is trained via self-supervised learning to learn powerful speech representations, learning both phonetic and semantic information about speech through audio samples.

1) The input to the Wav2Vec2 model is the raw audio waveform, which is then coded into latent representations by a convolutional feature encoder and then transformed by a transformer-based context network. These embeddings are useful in capturing long range dependencies and contextual relationships in the speech signal. The features extracted would subsequently be processed into a classification layer or a combination of the outputs of the CNN-LSTM model by a weighted score fusion approach to improve the overall performance.

Combining Wav2Vec2 and CNN-LSTM, the system is able to enjoy the advantages of both handcrafted spectral features and deep contextual representations, leading to a higher accuracy and strength in speech emotion recognition tasks.

G. Reliability-Aware Fusion Method

The prediction outputs of CNN, LSTM, and Wav2Vec2 networks are fused using a weighted combination technique that factors in the reliability of the model while assigning weights to the output predictions. The weight assigned is based on the level of confidence of the model when processing the input data. The formula for the fused output is:

$$E_{\{final\}} = \alpha E_{\{CNN-LSTM\}} + (1 - \alpha) E_{\{Wav2Vec2\}}$$

- **Efinal**: Final predicted emotion score after combining all models
- **ECNN–LSTM**: Emotion prediction score obtained from the hybrid CNN–LSTM model
- **EWav2Vec2**: Emotion prediction score obtained from the Wav2Vec2 model
- α: Weighting factor (ranges between 0 and 1) that controls the contribution of the CNN–LSTM model
- $(1-\alpha)$: Complementary weight assigned to the Wav2Vec2 model

F. Emotion Classification

Classification of the speech data into emotional classes such as happiness, sadness, anger, neutral, fear, disgust, surprise, and calmness is done using the system. Final determination of the emotion is done through the fused prediction score

resulting from the fused outputs from CNN, LSTM, and Wav2Vec2. Emotion having the highest fused score is selected as the final output. Such an approach ensures consistent classification of speech data regardless of how it was expressed, and both emotions and their scores are given as outputs by the system.

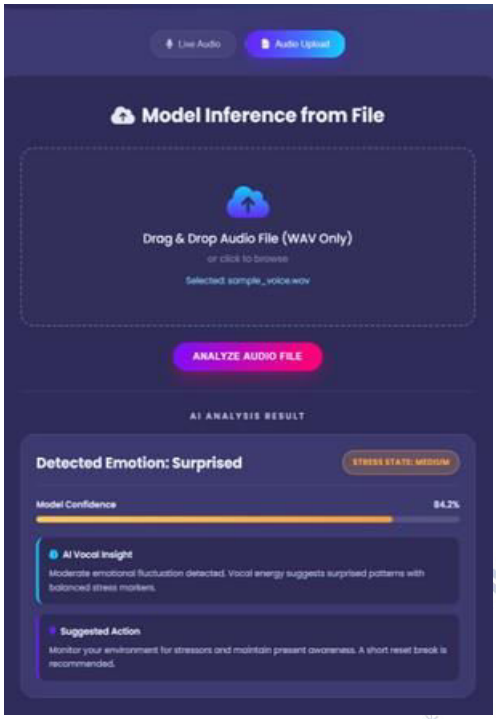

Fig. 5. Output interface showing predicted emotion and confidence score.

III. Results and Discussion

A. Performance of Speech Emotion Recognition Model

The effectiveness of the proposed hybrid model, a combination of CNN-LSTM and Wav2Vec2 implemented with a reliability-aware fusion strategy is reported to be strong in speech emotion recognition. An overall accuracy of the model is 96.2% and the generalization is strong across all classes of emotions. These findings suggest that the spectral-temporal features are greatly improved in their performance when used together with the contextual embeddings

Table 2 provides a more detailed analysis of the results in class-wise form, indicating the precision, recall, and F1-score of each emotion. The model demonstrates a high performance in most classes, and the F1-scores are above 94% on most of the emotions. However, interestingly, the calm mood has the best result in the F1-score 99 which is an outstanding classification capability. Conversely, the fearful class has the lowest F1-score of 90% which can be explained by the fact that it has a similarity with other expressions of emotion that results in a slight misclassification.

Table 2. Classification Performance for Each Emotion Class

Emotion	Precision	Recall	F1-Score
Angry	100%	91%	95%
Calm	97%	100%	99%
Disgust	95%	98%	96%
Fearful	86%	95%	90%
Happy	100%	94%	97%
Neutral	93%	95%	94%
Sad	94%	97%	95%
Surprised	99%	95%	97%

B. Confusion Matrix Analysis.

The confusion matrix shows how the model is performing in classifying the various emotions.

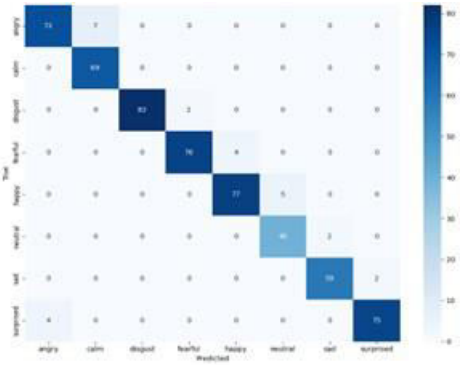

Fig. 6. Confusion matrix of Hybrid model.

C. ROC and Precision–Recall Analysis

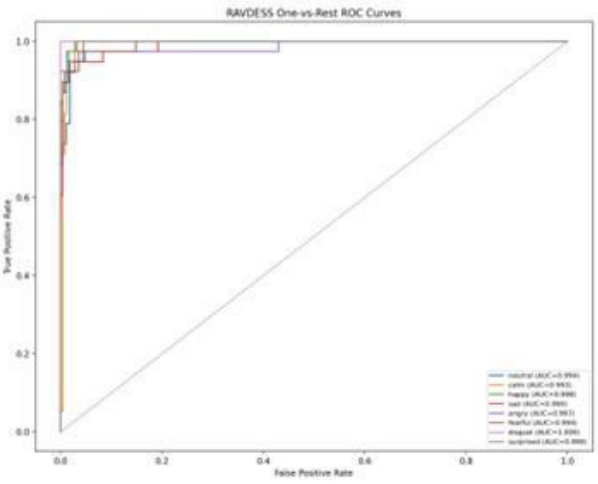

Figure 7. ROC Curve for Hybrid Model

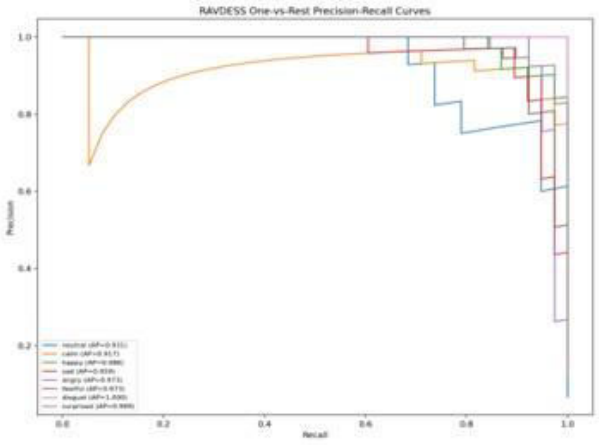

Figure 8. Precision–Recall Curve Hybrid Model

D. Comparison with Existing Methods

Table 3. Comparison with Existing Emotion Detection Methods

Approach	Dataset	Accuracy
SVM	Emotional Speech	78%
DNN	Speech Dataset	83%
CNN	RAVDESS	85%
Deep Learning	Speech Dataset	90%
CNN–LSTM	Speech Dataset	**95.49%**
Wav2Vec2	Speech Dataset	**91.66%**
Proposed CNN–LSTM + Wav2Vec2 (Fusion)	Speech Dataset	**96.2%**

E. Discussion

The findings indicate that the proposed hybrid model CNN-LSTM with wav2vec 2.0 can be used to capture the spatial and temporal information of the speech signals. The model is highly accurate on most of the emotional categories with F1-scores of more than 94% in most of the classes.

The calm emotion shows the highest performance with F1-score of 99%, which means it can be classified well. The fearful group has the lowest performance with an F1-score of 90 and this could be because of similarity with other emotional expressions.

Generally, the model shows a strong overall performance with an average F1-score of about 95% that reflects a strong and reliable performance on the speech emotion recognition tasks.

IV. CONCLUSION

This paper a system proposes that uses learning to recognize emotions in speech. It looks at the sounds in signals to figure out how someone is feeling. The system uses s Mel-Frequency Cepstral Coefficients and a special kind of architecture that combines CNN and LSTM to understand the sounds and how they change over time. It also uses a Wav2Vec2-based model to get an understanding of the context of the audio.

The system then combines the results of these models in a way that considers how reliable each result's which makes the whole system work better.

The results of the experiments show that this system is good at recognizing emotions in speech with an accuracy of 96.2%. It works well for all kinds of emotions with most of them being recognized more than 95% of the time. The system is especially good at recognizing when someone is feeling calm with an accuracy of 99%.

Overall, the system is good at recognizing emotions in speech with a performance of about 96%. This means it is robust and can be used in different situations.

In conclusion the system that combines kinds of features to recognize emotions in speech is effective. It is an accurate and reliable system that could be used in many real-world applications, such as computers that can understand how humans are feeling, analysing behaviour and smart audio systems. The speech emotion recognition system has a lot of potential, for these kinds of applications.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," Advances in Neural Information Processing Systems, pp. 1097–1105, 2012.

[2] S. Livingstone and F. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS)," PLOS ONE, vol. 13, no. 5, 2018.

[3] H. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," Neural Networks, vol. 92, pp. 60–68, 2017.

[4] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognition, vol. 44, no. 3, pp. 572–587, 2011.

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] Ramyakrishna Kadiyala, etc , "Psychological Stress Detection from Voice using Deep Learning," International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE), vol. 14, no. 3, pp. 2027–2038, Mar. 2026.

[7] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH emotion challenge," Proceedings of Interspeech, pp. 312–315, 2009.

[8] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: The Munich versatile and fast open-source audio feature extractor," Proceedings of ACM Multimedia, 2010.

[9] D. O'Shaughnessy, Speech Communications: Human and Machine, IEEE Press, 2000.

[10] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," Proceedings of Interspeech, 2014.

[11] Sarkar et al., "A review of speech emotion recognition using deep learning," IEEE Access, vol. 8, pp. 11171–11186, 2020.

[12] BHANDARI, "GENERATING LOG-MEL SPECTROGRAM USING LIBROSA," SIGNAL PROCESSING STACKEXCHANGE, 2021.

[13] Z. Zeng, M. Pantic, G. I. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 1, pp. 39–58, 2009.

[14] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," Language Resources and Evaluation, vol. 42, no. 4, pp. 335–359, 2008.