International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

A Comparative Analysis of CNN-LSTM and MFCC-LSTM for Sentiment Recognition from Speech Signals

Suman Lata[1*], Neha Kishore[1], Pardeep Sangwan[2]

Submitted: 04/02/2024 **Revised**: 11/03/2024 **Accepted**: 18/03/2024

Abstract: Sentiment analysis of speech is a rapidly evolving field with immense potential for Human Computer Interaction (HCI). As technology improves and addresses current challenges, we can expect a future where computers interact with us on a deeper emotional level, creating a more natural and intuitive user experience. Sentiment analysis of speech allows computers to understand the emotional tone behind a user's words. This unveils a powerful tool for designing more natural and empathetic HCI systems. While sentiment analysis often focuses on written text, speech offers a richer sentimental landscape. Voice tone for example- **s**arcasm, frustration, excitement, speech patterns of speaking speed, hesitation, emphasis & non-verbal cues like laughs, sighs, or grunts can add emotional context missed in the text. This research proposes a hybrid architecture that combines Convolution Neural Network (CNN) with Long Short Term Memory (LSTM) and leverages linear stack of deep stride layers to enhance the accuracy metrics of sentiment recognition system by speech signals. Convolution neural network capture spatial features efficiently from spectrograms while LSTM networks excel at modeling temporal dependencies. This system classifies seven sentiments such as happiness, disgust, sadness, angry, neutral, fear and pleasant surprise from the speaker's utterances. The proposed work utilizes Toronto Emotional Speech Set (TESS) dataset. Experimental results demonstrate that the hybrid CNN-LSTM architecture achieves high accuracy rate of 98 % which is slight improvement in our previous work utilizing MFCC+LSTM having 96% accuracy in recognizing sentiments, outperforming other state-of-the-art methods. Notably, the model achieved these results utilizing a relatively smaller size (1.8 MB), highlighting its computational efficiency.

***Keywords**: Convolution neural network, Long short-term memory network, Sentiment recognition, and Deep learning, IoT (Internet of Things), Hearing aids.*

1. Introduction

An essential aspect of human existence are sentiments. Sentiments are fundamental to human interaction, shaping our daily lives and influencing how we perceive the world around us. We express them not only through facial expressions and gestures, but also through the subtle nuances of our speech. Speech emotion recognition (SER) focuses on analyzing these vocal cues to understand a speaker's emotional state. The foundation of SER lies in the connection between sentimental state of an individual and their corresponding physiological changes[1]. Most sentiments trigger specific physiological responses, which in turn affect how we produce sound. Many sentiments like fear, surprise, or happy etc. trigger physiological changes that alter sound production mechanism. These changes can affect the vibration of our vocal cords, the shape of the vocal tract, and ultimately, the way our voice sounds[2]. While facial expressions have dominated research in emotional communication, speech offers a valuable yet understudied source of information. Traditionally, sentiments were considered subjective and difficult to quantify, hindering research in this area. However, recent advancements have sparked renewed interest in sentiment recognition. Sentiment recognition holds immense potential for real-world applications[3]. Hearing aids equipped with sentiment recognition systems can improve communication for individuals with autism. Call centers could leverage SER to identify frustrated callers and route them to human support. Educational technology could adapt its presentation style based on a student's emotional engagement. Safety in autonomous vehicles by detecting a driver's emotional state through voice can trigger emergency features in self-driving cars. Voice interfaces are becoming increasingly common in IoT (Internet of Things) devices like smart speakers[4]. SER can improve user experience by understanding emotional cues in voice commands. Speech audio is a reliable source for emotion data due to its ease and affordability of acquisition compared to other biological signals. The rise of speech-enabled devices across various applications has fueled a surge in research on Speech Emotion Recognition (SER). As a result, the majority of researchers are currently focusing on SER. Numerous scholars and researchers are motivated to study SER systems since it is considered a significant endeavour in the field of Human-Computer Interaction (HCI). However, for successful SER systems, three key challenges need to be addressed: 1) utilizing high-quality emotional speech databases, 2) developing efficient feature extraction methods, and 3) designing robust deep learning classifiers with well-suited algorithms and

[1]Maharaja Agrasen University, Himachal Pradesh, India
[2] Maharaja Surajmal Institute of Technology New Delhi, India

models[5]. The limited amount of emotional information present in speech signals poses a challenge. So choosing the right emotional speech database ensures the system is trained on realistic data. For speaker-independent and speaker-dependent experiments, popular sentiment based audio datasets like TESS (Toronto Emotional Speech Set), RAVDESS (The Ryerson Audio-Visual Database of Emotional Speech and Song), SAVEE (Surrey Audio-Visual Expressed Emotion), and Emo-DB (Berlin Database of Emotional Speech) were typically utilized[6]. Feature extraction techniques like pitch, energy, frequency, LPC, LPCC and Mel-Frequency Cepstral Coefficients (MFCC) are used to extract relevant emotional information from speech signals. While analyzing speech for emotions, researchers often utilize a vast array of features hoping to capture every nuance. However, combining too many features can increase redundancy and computational complexity. Selecting the most informative features is essential for optimal performance[7]. The last step of the Speech Emotion Recognition system usually involves classifying the audio input data. Conventional algorithms like HMM, GMM, Artificial Neural Networks, Support Vector Machines (SVM) etc. can be used to classify emotions based on extracted features. Because of complex preprocessing, limited input flexibility, static nature, real world performance with background noise & speaker variations, the performance of traditional SER systems can degrade significantly[8]. The emergence of deep learning offers promising solutions to these hurdles. Deep learning models like Convolutional Neural Networks (CNNs), directly work on raw dataset with dynamic learning approach[9]. By leveraging the capabilities of deep learning, SER systems can become more robust, flexible, and cost-effective, paving the way for wider adoption in real-world applications.

2. Literature Review

Digital signal processing is a vast research area. Researchers have recently developed some effective methods for SER that use digital audio speech signals to determine a speaker's emotional state. A typical speech sentiment recognition system has two key components: feature extraction and classification. Feature extraction aims to capture high-level emotional cues from speech data. Classification then uses these features to identify emotions accurately. Researchers have developed various methods and algorithms for this purpose. Traditionally, some studies relied on hand-crafted, low-level features to train models like MFCC, Formants, Pitch and Neural Networks (NNs) for SER. However, recent research has shifted towards deep learning approaches that directly process raw audio signals, potentially leading to improved recognition accuracy.

a) Shallow Feature-Based Speech Emotion Recognition-This traditional approach relies on manually selecting and extracting specific features from the speech signal that are believed to be relevant to sentiments. Common examples of these hand-crafted features include-MFCC, pitch, and formants. These features are then fed into a machine learning classifier, such as a Support Vector Machine (SVM), K- nearest neighbor etc. to predict the sentiment. Namrata Dave et.al [10]evaluate various feature extraction techniques like - PLP-RASTA (PLP-Relative Spectra) , Perceptual Linear Prediction (PLP), Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Codes (LPC) for speech emotions and because MFCC is based on the idea of a logarithmically spaced filter bank combined with the idea of the human auditory system, it responds better than LPC parameters. For classification and recognition, methods such as Vector Quantization (VQ), Dynamic Time Warping (DTW), Support Vector Machine (SVM), and Hidden Markov Model (HMM) can be employed. Gabrielle K. Liu et al.[11]investigated Gammatone Frequency Cepstral Coefficients (GFCCs), demonstrating a slight improvement over MFCCs when combined with additional features like jitter and shimmer. Zhen-Tao Liu et al. [12] proposed a method to identify hidden emotional features based on correlations within a Chinese speech dataset (CASIA). They used an Extreme Learning Machine (ELM)-based decision tree for classification and achieved 89.6% recognition rate. Fahad et al. [13]explored feature selection based on glottal and MFCC features for training Deep Neural Network (DNN) models for SER and concluded that accuracy had been improved by 7.13 % by using hybrid model. Saurabh Sahu et.al[14]proposed an auto-encoder approach for SER. An auto-encoder can extract high-dimensional hidden features, while a sparse network extracts lower-dimensional sparse features. Lianzhang Zhu et.al[15]explored novel approaches for Chinese speech emotion dataset. A new classification technique was utilized in which Deep Belief Network (DBN) serving as a deep feature extractor and support vector machine (SVM) as a classifier. When combined, they produce a better outcome than that of using only SVM or DBN. Hand-crafted features have limited ability to capture complex relationships & might not capture the full range of emotional cues present in speech. To address this, researchers have adopted various hybrid techniques.

b) Convolution Neural network Based Speech Emotion Recognition -These approaches offer the ability to automatically learn discriminative features directly from raw speech data. Researchers are increasingly utilizing Convolutional Neural Networks (CNN) to extract effective features for sentiment recognition systems.

Zhang et al.[16]proposed a method using a pre-trained CNN model (Alex Net) to extract deep features from speech. Using these extracted features, an SVM classifier was employed to identify sentiments in the speech data. George et al.[17]explored a technique for spontaneous SER that combined a CNN with a Long Short-Term Memory (LSTM) network using RECOLA natural emotion database. The CNN learned discriminative features from the entire speech utterance, while the LSTM captured the sequential nature of the speech signal to identify sentiments. Dong Yu et al.[18]demonstrate that that DNNs can extract more discriminative and invariant features at higher layers. DNN learned features exhibit reduced sensitivity to variations in the input features. Due to this characteristic, DNNs are able to generalize more effectively than shallow networks, and CD-DNN-HMMs are able to recognize speech in a way that is more resilient to variations in the speaker, surroundings, or bandwidth. Guihua Wen et al.[19]employed Deep Belief Networks (DBNs) for SER. Low-level features were first extracted and then fed into DBNs to learn high-level, discriminative features for emotion classification using an SVM classifier. Fang Bao et al.[20]suggest a novel Cycle-GAN(Generative Adversarial Networks)-based architecture that, guarantees discriminability among generated samples & ensures similarity between real and synthetic generated features. Authors demonstrated that as compared to a classifier trained solely on real feature vectors, a neural network classifier trained on a combination of real and synthetic feature vectors performs better in terms of classification. Noushin Hajarolasvadi et al.[21]proposed a method that segmented speech signals into frames, extracted MFCC features, and converted them into spectrograms. Key-frames (k most discriminant frames) representing the entire utterance were selected using k-means clustering. Finally, a 3D CNN was trained to predict speech emotions.

Some research gaps identified from literature are as follows:

1) Feature extraction effectiveness: While sentiment analysis typically focuses on positive and negative sentiment, recognizing specific emotions (e.g., anger, happiness, sadness) from speech is an area that requires more attention. The relative effectiveness of these approaches in capturing sentiment specific nuances in audio is not fully understood.

Our Approach- CNN automatically learn features from audio.

2) Handling Temporal dependencies: Traditional methods rely on statistical models like HMMs which may not capture complex temporal dynamics effectively.

Our Approach- LSTM is used to capture complex temporal dynamics.

3) Data Augmentation and Synthesis: Limited labeled datasets for sentiment analysis from speech signals constrain the performance of deep learning models. Also sentiment analysis models may suffer from data imbalance and bias, leading to skewed results.

Our Approach – Some data augmentation techniques and synthetic data generation methods are used in this research to expand training datasets and improve model accuracy and robustness.

2.1 Motivation and Scope

Sentiment analysis of spoken speech signals is a challenging and evolving research area, and various acoustic features have been used in this context. Some challenges in identifying sentiments from speech cues are summarized as follows:

1) Identifying sentiments from speech can be complex, and age adds another layer of difficulty. Sentiments expressed by elderly individuals may be conveyed differently compared to younger people. These variations in speech patterns can make it harder for models to accurately recognize emotions in adults.

2) Imbalanced datasets is another big challenge when the number of speech samples representing certain emotions is significantly lower compared to others. For instance, a dataset might have a large number of "silence" frames labeled as a single emotion category. The model becomes heavily influenced by the dominant emotional class (e.g., silence). When encountering new silence frames, the model's predictions will likely be biased towards that majority class, even if a different emotion is present.

3) Emotion labels are often assigned to entire speech utterances (sentences or phrases). However, within a single utterance, emotions can fluctuate. Labeling every frame within an utterance with the same emotion can be misleading, as the emotional content might change throughout the speech.

4) Mel-frequency cepstral coefficients (MFCCs) are a popular choice for feature extraction in SER. However, they have a significant drawback: they treat each element within a speech frame independently, ignoring the relationships between neighboring elements. This can lead to the loss of important emotional information embedded in the consequent frames.

2.2 Major Contributions

1) The primary contributions of this work focuses on recognizing specific sentiments for example- happy, sad, anger, fear, disgust, neutral, pleasant surprise from speech by using CNN-LSTM hybrid deep learning technique utilizing CNN-LSTM.

2) The CNN captures local patterns, and the LSTM can learn how these patterns evolve over time, potentially leading to more accurate sentiment recognition.

3) Furthermore, sentiments expressed vocally can differ between individuals due to cultural background and environmental factors. Someone raised in a culture where expressing strong sentiments is discouraged might use subtle vocal cues compared to someone from a more expressive culture. This hybrid model gives promising results, considering potential speaker-specific variations.

4) The key lies in innovative implementation for building robust and accurate sentiment recognition systems.

The structural organization of this work is as follows:

A review of state of arts on sentiment analysis and recognition with traditional techniques is given in section II. Proposed model methodology is explained in section III. Section IV covers the proposed hybrid architecture used for sentiment analysis. Further, section V provides the experimental results. Finally, postscript conclusion is presented in section VI.

3. Methodology

The main aim of this study is to recognize sentiments of an individual by processing audio speech signals. Leveraging deep learning algorithms, sentiment recognition systems can analyze and understand human emotions. Moreover the motive of this research is to make an efficient SER system with high accuracy and reduced false rate. Previous research in this area often relied on analyzing word choice, for sentiment recognition. This approach typically classifies sentiments into just three categories: positive, negative, and neutral. These method has limitations in capturing the nuances of human sentiments. MFCCs effectively capture the spectral characteristics of speech signals, which are informative for recognizing sentiments. While MFCCs can be a good starting point for feature extraction, CNNs have the capability to learn even more complex and nuanced features directly from the raw data through their convolutional layers[22]. This translates to faster training times and lower computational demands compared to more complex deep learning architectures. CNNs excel at identifying patterns within data, making them well-suited for analyzing the emotional cues embedded in speech. Their relatively simple structure and efficient use of parameters makes them advantageous for real-time applications. The proposed model is simple & utilizes fewer parameters to train the model. Conventional Speech Emotion Recognition (SER) models follow a three-step process: speech signal pre-processing, speech sentiment feature extraction, and sentiment recognition shown in figure 1. Extraction of proper emotional characteristics and sentiment classification are the key aspect of SER model which directly affect the models performance.

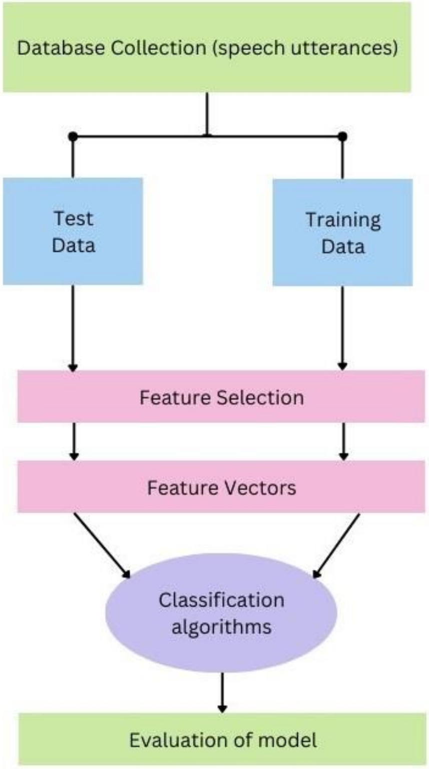

Fig 1. Basic steps for sentiment recognition system

3.1 Implementation

The proposed model is executed by hybrid technique in which features are extracted by CNN and the sentiments are recognized by LSTM technique. The process starts with gathering relevant datasets, which are then used to train the developed model. Based on the data, all decisions and outcomes generated by the developed model are made. Here we have used TESS dataset which is publically available on keggal repository. The University of Toronto provided this dataset. Two Toronto-native actresses were located and signed on, considering that English is their native tongue, both women speak it with ease. Additionally, they have some degree of musical training and are well-educated, both actresses' thresholds fall within the typical range. A list of prearranged target words was given to them to speak. Seven different sentiments were recorded: fear, sadness, disgust, happiness, anger, pleasant surprise, and neutral. After collecting the audio files the next step is data visualization. It is a graphical depiction of data and information. Imagine an audio signal as a three-dimensional landscape. Time stretches along one axis, loudness (amplitude) varies on another, and frequency creates the height, representing the different sound pitches. Data visualization here plays a crucial role by transforming raw audio data into visual representations that aid in understanding and analyzing sentiments. Tools like wave plot, charts, graphs, and spectrograms make it easier to reveal trends, uncover hidden patterns, spot outliers in the audio data that might be difficult to spot with numerical analysis alone. For example, a gradual increase in intensity over time might indicate rising anger in a speaker's voice. Here librosa library is utilized for analyzing and extracting features. It facilitates various visualizations like spectrogram, wave plots etc. It depict the frequencies present in the audio signal at different points in time. By analyzing the distribution of color within a spectrogram, we can identify patterns related to pitch, formants, and other characteristics that convey emotional information. For example, figure 2 and 3 represents wave plots & corresponding spectrograms for happy and disgust sentiments respectively. An algorithm based model is constructed with sampling rate of 22 KHz in the third step. Input data is divided into training and testing dataset. Since speech utterances are converted into images i.e spectrogram, a CNN model is constructed to extract the relevant features by using Keras in python. Next layer of model is LSTM, that capture the temporal dependencies between speech segments, which can be crucial for understanding sentiments. Finally the sentiments are recognized by using this hybrid model. Steps for dataflow process is depicted in figure 4.

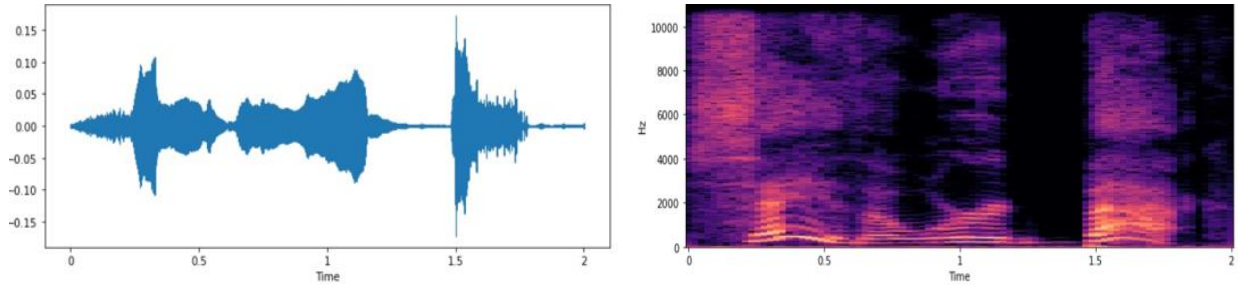

Fig 2. Wave plot and Spectrogram for speech sample with happy sentiment

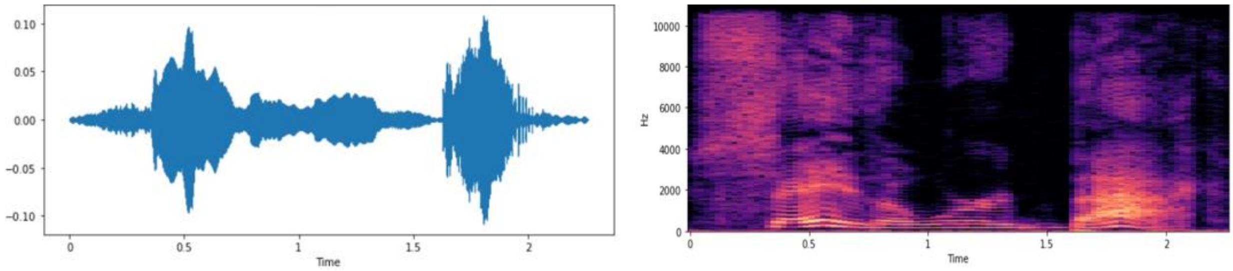

Fig 3. Wave plot and Spectrogram for speech sample with disgust sentiment

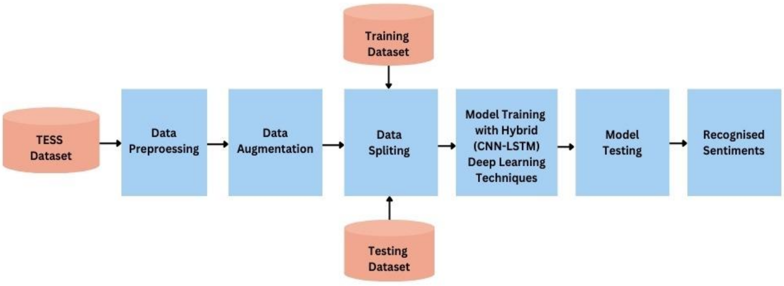

Fig 4. Dataflow process of proposed sentiment recognition system

3.2 Basic Framework & Pseudo Code of the Proposed Algorithm

The proposed model experimental approach is described below:

Step1. Load the training data set (TESS) using Keras API.

Step2. Convert audio samples into spectrograms.

Step3. Flatten the input image dimensions.

Step4. Normalize the image using one hot encoding using Keras.

Step5. Build a linear stack of convolution layers and max pooling layers with kernel size 3x3, pool size 2x2 & activation Leaky ReLU

Step6. Reshape for LSTM model.

Step7. LSTM layers for sequential processing

Step8. Flatten dimensions

Step9. Apply dense layers with activation ReLU & later, apply activation softmax for classification.

Step10. Compile the model.

Step11. Split the data into training and testing data sets.

Step12. After normalization train the model**.**

Step13. Evaluate the model**.**

The pseudo code for the proposed methodology is given below:

Input: Speech audio data
Import :**tensor flow**; Load : Dataset :**TESS**
Convolution layer : Number of neurons(64)
Define**Conv2D**(filters; kernel size=3x3;activation='leaky Relu')
Input shape : (128x70x1)
Max-pooling layer : Pool size(2,2); Down sampling
Output: **Flatten layer**(final convolution layer)
Define **LSTM Layer**: number of neurons
Dense layers ;dropout;
Output layer : Softmax activation function
Compile model : loss function : categorical cross-entropy; **optimizer**: Adam ;Normalization
Train model: specified no. of epochs
Monitor training :accuracy &loss
Evaluate model : validation dataset

4. Proposed Hybrid CNN-LSTM Architecture

In the proposed hybrid architecture shown in figure 5, the audio samples were converted into spectrogram and Convolution Neural Network, a deep learning algorithm is applied for feature extraction and classification is performed by LSTM technique. TESS data set having audio samples of seven sentiments from keggal repository is utilized. For analysis 400 recordings of each sentiment class are taken which will further augmented to increase the accuracy in training phase. So 800*7=5600 input samples were extracted from the data set and converted into spectrograms. Image array shape of the samples is 128x70x1. Since the model is built step-by-step, adding convolutional layers, pooling layers, and fully connected layers in a sequential order, sequential API in python is utilized. In this proposed model each input image will pass through a series of convolution layers with filters also called Kernels of size 3x3, with activation function leaky Relu. To reduce the dimensionality max pooling is utilized with pool size 2x2. To achieve better results three convolution layers and max pooling layers are used. Then we flattened our matrix into vector, reshape it and feed it into a fully connected layer LSTM. 64 neurons are at input node of LSTM layer followed by two dense layers with 30% dropout. For final output of 7 sentiments Softmax activation function is applied at last layer.

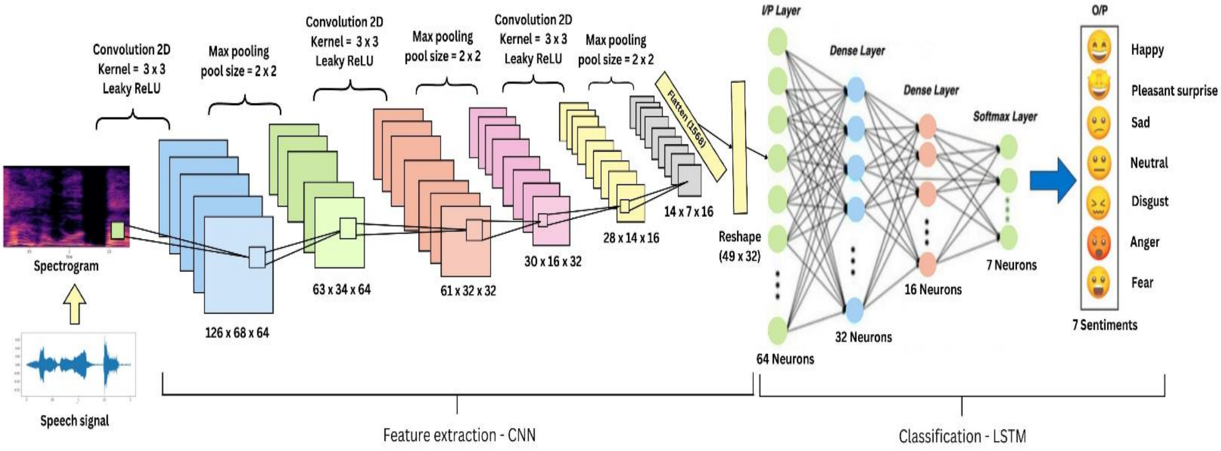

Fig 5. Proposed hybrid architecture for sentiment analysis

4.1 Model Organization and Computational Setup

The proposed deep learning model architecture was implemented in Python using libraries like librosa and TensorFlow, along with other necessary resources. Speech data with offset 0.5, sampling rate- 22 KHz & duration of 3ms was processed to generate spectrograms. The generated spectrograms were then strategically divided: 80% allocated for training the model and 20% reserved for testing its performance on unseen data. The model was trained on 7 epochs with 0.001 learning rate. The batch size is 32 and get training accuracy 0.9810 with loss value 0.0543. Notably, the model achieved these results with a relatively small size (1.8 MB), highlighting its computational efficiency. Table 1 represents the pre-trained CNN feature maps and the relationship between CNN and LSTM for proposed model.

Table 1. CNN-LSTM model training parameters summary

Model Type	Layer Type	Output Shape	Parameters
Sequential (API) with Input dimensions - 5600	Conv 2D	(none,126, 68, 64)	640
	Maxpooling	(none, 63,34,64)	0
	Conv 2D	(none, 61,32,32)	18464
	Maxpooling	(none, 30,16,32)	0
	Conv 2D	(none, 28,14,16)	4624
	Maxpooling	(none, 14, 7, 16)	0
	Flatten	(none, 1568)	0
	reshape	(none, 49, 32)	0
	LSTM	(none, 49,64)	24832
	Flatten	(none, 3136)	0
	Dense	(none, 32)	100384
	Dense	(none, 16)	528

	Dense	(none, 7)	119
Total Parameters	149,591		
Trainable Parameters	149,591		
Non-Trainable Parameters	0		

5. Experimental Results

Table 2 represents various evaluation parameters used in the proposed model. A comparison with our previous work in terms of precision, F1-score, recall is also shown in this table. It has been observed that hybrid model CNN-LSTM gives better accuracy than our previous MFCC-LSTM hybrid model. Simulation parameters are summarized in table 3.

Table 2. Architecture performance evaluation & comparison with CNN-LSTM & MFCC LSTM

Nature	**Result with CNN-LSTM Model**			**Result with MFCC-LSTM Model**		
Sentiments	**Precision**	**Recall**	**F1 score**	**Precision**	**Recall**	**F1 score**
Angry	0.97	0.97	0.97	0.94	0.98	0.96
Disgust	0.98	0.96	0.97	0.92	0.97	0.94
Fear	0.99	1.00	1.00	1.00	0.96	0.98
Happy	0.98	0.95	0.96	0.90	0.95	0.93
Neutral	0.99	0.99	0.99	1.00	1.00	1.00
Pleasant Surprise	0.96	0.96	0.96	0.98	0.85	0.91
Sad	0.96	1.00	0.98	0.97	0.98	0.97
Macro Avg.	0.98	0.98	0.98	0.96	0.96	0.96
Weighted Avg.	0.98	0.98	0.98	0.96	0.96	0.96
Accuracy	**0.98**			**0.96**		

To enhance the performance of model three convolution layer along with max pooling layer is used with 3x3 filter size. Figure 6 shows the training & validation accuracy of the proposed work, which is higher than our MFCC-LSTM model. Figure 7 presents loss value of the model. Figure 8 elaborates the overall prediction performance of model in terms of confusion matrix. For analyzing the confusion matrix row normalization is preferred for the model which shows different matrix values in different colors to better illustrate the distribution of the sentiments.

Table 3. Simulation parameters used in proposed work

Model Parameters	**Value/Type**
Model optimizer	ADAM
Loss Function	Categorical Cross Entropy
Learning Rate	0.001
Epochs	7
Pool size	Max pooling (2x2)
Batch Size	32
Kernel Size	3x3
Activation function	Leaky ReLU
Activation Output	Softmax
Encoder	One hot encoder
Image Array Shape	128x70x1

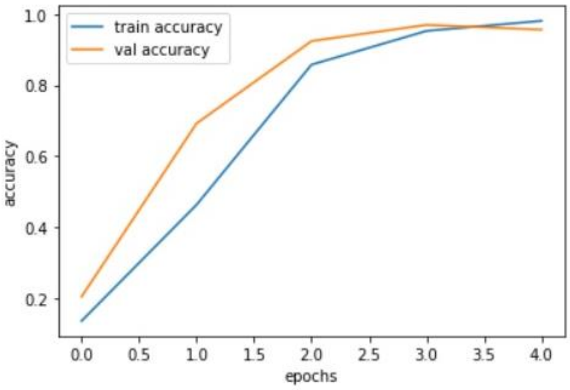

Fig 6. Accuracy of the proposed model

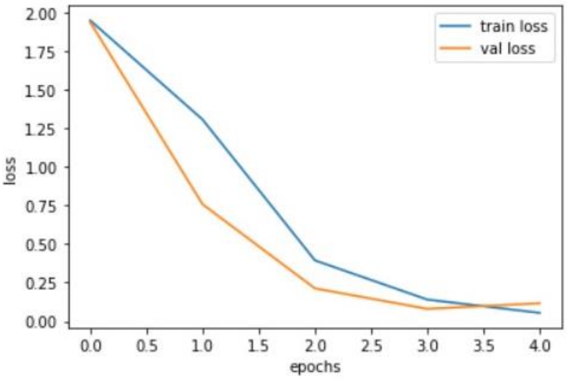

Fig 7. Loss value of the proposed model

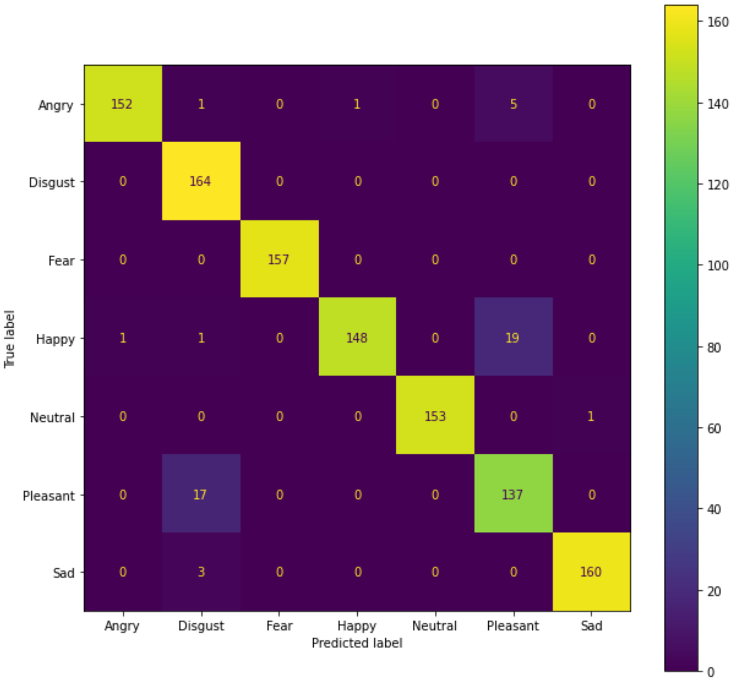

Fig 8. Confusion matrix for seven sentiments

The distinct colors on the diagonal cells (representing correct predictions) offer a clear picture of the model's performance, which is an exact indication that model performs relatively well as compare to MFCC-LSTM technique. Figure 9 depict a bar chart that describe the accuracy score for each emotion class. As inferred from bar chart, accuracy values of recognizing sentiments set – {disgust, fear, neutral, sadness} is very high. Whereas sentiments that belong to set- {happy, pleasant surprise, angry} shows less accuracy. The comparison of this work with other previous researches is presented in table 4.

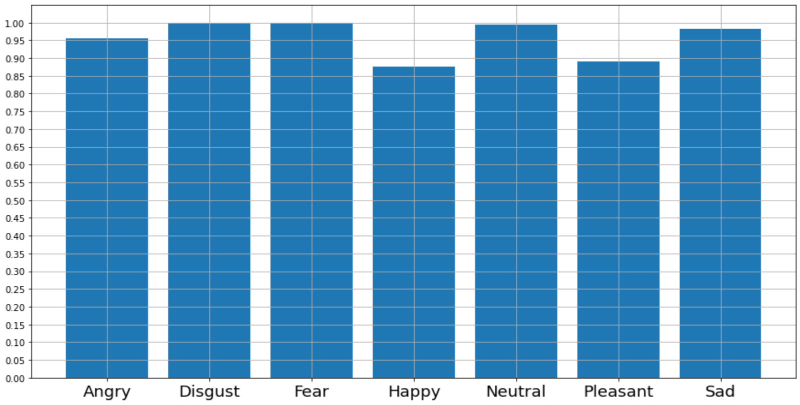

Fig 9. Bar chart depicting accuracy score of different sentiment classes

Table 4. Comparison of proposed work model with previous models.

Reference	Data Set	Model used	Recognized Emotions	Accuracy
[23]	CREMA-D, RAVDESS, SAVEE, TESS	CNN	6(happiness, anger, sadness, fear, disgust, and neutral)	58 %
[24]	TESS	CNN-LSTM	7(fear, happy, sad, disgust, pleasant surprise, neutral, angry)	62 %
[25]	TESS,RAVDESS	CNN-MFCC	7(fear, happy, sad, disgust, pleasant surprise, neutral, angry)	78 %
[21]	SAVEE, RML	CNN- K mean	6(happiness, anger, sadness, fear, disgust, and neutral)	81 %
[16]	EMO-DB, RML, eNTERFACE05	CNN-SVM	6(happiness, anger, sadness, fear, disgust, and neutral)	87 %
Our previous work	TESS	MFCC-LSTM	7(fear, happy, sad, disgust, pleasant surprise, neutral, angry)	96 %
Our proposed work	TESS	CNN-LSTM	7(fear, happy, sad, disgust, pleasant surprise, neutral, angry)	98 %

6. Conclusion

In previous studies sentiments were recognized on the bases of hand-crafted features. This research proposed a hybrid architecture utilizing the key strength of CNN & LSTM, deep learning techniques for sentiment recognition. Toronto Emotional Speech dataset is utilized. The raw audio signal after preprocessing (converted into spectrograms) fed to CNN inputs. CNN extracts the features from the preprocessed data through its stacked convolutional layers. The extracted feature vectors were input to fully connected LSTM architecture with various dense layers. The proposed hybrid model is able to learn the spatial features efficiently from spectrograms while LSTM networks excel at modeling temporal dependencies. The hybrid model's effectiveness is shown through the results obtained from the experiments. The accuracy rate of the proposed model is 98%. Further, proposed work aims to give better performance with reduced false rate. Findings show that by leveraging deep learning techniques the model size is reduced to 1.8 MB which provide decreased computational time and can be applied to real world scenario where speech utterances are not static.

However need of large amount of data, data diversity, data quality and black-box nature of deep learning techniques are always key hurdles for sentiment recognition systems. Speech signals often co-occur with other modalities such as text or facial expressions, which can provide complementary information for sentiment analysis. In future CNN-LSTM models can be extended to incorporate multimodal inputs, allowing the model to leverage information from multiple modalities simultaneously, further enhancing the accuracy and robustness of sentiment analysis systems.

References

[1] J. Rong, G. Li, and Y. P. P. Chen, "Acoustic feature selection for automatic emotion recognition from speech," *Inf. Process. Manag.*, vol. 45, no. 3, pp. 315–328, 2009, doi: 10.1016/j.ipm.2008.09.003.

[2] M. S. Likitha, S. R. R. Gupta, K. Hasitha, and A. U. Raju, "Speech based human emotion recognition using MFCC," *Proc. 2017 Int. Conf. Wirel. Commun. Signal Process. Networking, WiSPNET 2017*, vol. 2018-Janua, pp. 2257–2260, 2017, doi: 10.1109/WiSPNET.2017.8300161.

[3] J. Paul *et al.*, "A survey and comparative study on negative sentiment analysis in social media data," *Multimed. Tools Appl.*, 2024, doi: 10.1007/s11042-024-18452-0.

[4] R. Mohd Hanifa, K. Isa, and S. Mohamad, "A review on speaker recognition: Technology and challenges," *Comput. Electr. Eng.*, vol. 90, no. January, p. 107005, 2021, doi: 10.1016/j.compeleceng.2021.107005.

[5] D. Deshwal, P. Sangwan, and D. Kumar, "A structured approach towards robust database collection for language identification," *Proc. - 2020 21st Int. Arab Conf. Inf. Technol. ACIT 2020*, pp. 19–24, 2020, doi: 10.1109/ACIT50332.2020.9299963.

[6] M. Gupta and S. Chandra, “Speech emotion recognition using MFCC and wide residual network,” *ACM Int. Conf. Proceeding Ser.*, pp. 320–327, 2021, doi: 10.1145/3474124.3474171.

[7] P. Sangwan, D. Deshwal, and N. Dahiya, “Performance of a language identification system using hybrid features and ANN learning algorithms,” *Appl. Acoust.*, vol. 175, p. 107815, 2021, doi: 10.1016/j.apacoust.2020.107815.

[8] B. T. Atmaja and A. Sasou, “Sentiment Analysis and Emotion Recognition from Speech Using Universal Speech Representations,” *Sensors*, vol. 22, no. 17, 2022, doi: 10.3390/s22176369.

[9] S. Kwon, “A CNN-Assisted Enhanced Audio Signal Processing,” *Sensors*, 2020.

[10] Namrata Dave, “Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition,” *Int. J. Adv. Res. Eng. Technol.*, vol. 1, no. Vi, pp. 1–5, 2013, [Online]. Available: www.ijaret.org

[11] G. K. Liu, “Evaluating Gammatone Frequency Cepstral Coefficients with Neural Networks for Emotion Recognition from Speech,” pp. 2–6, 2018, [Online]. Available: http://arxiv.org/abs/1806.09010

[12] Z. T. Liu, M. Wu, W. H. Cao, J. W. Mao, J. P. Xu, and G. Z. Tan, “Speech emotion recognition based on feature selection and extreme learning machine decision tree,” *Neurocomputing*, vol. 273, pp. 271–280, 2018, doi: 10.1016/j.neucom.2017.07.050.

[13] M. S. Fahad, A. Deepak, G. Pradhan, and J. Yadav, “DNN-HMM-Based Speaker-Adaptive Emotion Recognition Using MFCC and Epoch-Based Features,” *Circuits, Syst. Signal Process.*, vol. 40, no. 1, pp. 466–489, 2021, doi: 10.1007/s00034-020-01486-8.

[14] S. Sahu, R. Gupta, G. Sivaraman, W. AbdAlmageed, and C. Espy-Wilson, “Adversarial auto-encoders for speech based emotion recognition,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2017-August, no. ii, pp. 1243–1247, 2017, doi: 10.21437/Interspeech.2017-1421.

[15] L. Zhu, L. Chen, D. Zhao, J. Zhou, and W. Zhang, “Emotion recognition from chinese speech for smart affective services using a combination of SVM and DBN,” *Sensors (Switzerland)*, vol. 17, no. 7, 2017, doi: 10.3390/s17071694.

[16] S. Zhang, S. Zhang, T. Huang, and W. Gao, “Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching,” *IEEE Trans. Multimed.*, vol. 20, no. 6, pp. 1576–1590, 2018, doi: 10.1109/TMM.2017.2766843.
“CNN 6 72019.pdf.”

[17] D. Yu, M. L. Seltzer, J. Li, J. T. Huang, and F. Seide, “Feature learning in deep neural networks – Studies on speech recognition tasks,” *1st Int. Conf. Learn. Represent. ICLR 2013 - Conf. Track Proc.*, pp. 1–9, 2013.

[18] G. Wen, H. Li, J. Huang, D. Li, and E. Xun, “Random Deep Belief Networks for Recognizing Emotions from Speech Signals,” *Comput. Intell. Neurosci.*, vol. 2017, 2017, doi: 10.1155/2017/1945630.

[19] F. Bao, M. Neumann, and N. T. Vu, “CycleGAN-based emotion style transfer as data augmentation for speech emotion recognition,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2019-September, pp. 2828–2832, 2019, doi: 10.21437/Interspeech.2019-2293.

[20] N. Hajarolasvadi and H. Demirel, “3D CNN-based speech emotion recognition using k-means clustering and spectrograms,” *Entropy*, vol. 21, no. 5, 2019, doi: 10.3390/e21050479.

[21] A. M. A. B, V. Palade, M. England, and R. Iqbal, *A Combined CNN and LSTM Model*. Springer International Publishing, 2018. doi: 10.1007/978-3-319-99740-7.

[22] M. Zielonka, A. Piastowski, A. Czyżewski, P. Nadachowski, M. Operlejn, and K. Kaczor, “Recognition of Emotions in Speech Using Convolutional Neural Networks on Different Datasets,” *Electron.*, vol. 11, no. 22, 2022, doi: 10.3390/electronics11223831.

[23] M. H. Farouk, “Emotion Recognition from Speech,” *SpringerBriefs Speech Technol.*, pp. 31–32, 2014, doi: 10.1007/978-3-319-02732-6_7.

[24] C. Hema and F. P. Garcia Marquez, “Emotional speech Recognition using CNN and Deep learning techniques,” *Appl. Acoust.*, vol. 211, p. 109492, 2023, doi: 10.1016/j.apacoust.2023.109492.