Frontiers in Health Informatics
ISSN-Online: 2676-7104
www.healthinformaticsjournal.com
2024; Vol 13: Issue 8
Open Access

An Automated Deep Learning based Speech Emotion Recognition System

[1]Manju Papreja, [2],Pooja Kumari [3]Rashmi Chhabra, [4]Dr.Surendra Singh Chauhan, [5]Anjali Gupta, [6]Sangeeta Kumari , [7]Renu Miglani, [8]Sachin Goyal

[1]Professor, Department of Computer Science & Application,GVM Institute OF Technology & Management ,Sonipat 131001Haryana, India Email: manju.papreja@gmail.com

[2]Assistant Professor and HOD Faculty of Computer Applications, HRIT, University, Ghaziabad, (U.P) INDIA, Email: *pujachoudhary2505@gmail.com*

[3]Professor, Department of Computer Science & Application,GVM Institute OF Technology & Management ,Sonipat 131001Haryana, India Email: rashmidahra@gmail.com

[4]Associate Professor, Department of Computer Science and Engineering, SRM University Sonepat (Haryana), INDIA. Email: *surendrahitesh1983@gmail.com*

[2]Assistant Professor Faculty of Computer Applications, HRIT, University, Ghaziabad, (U.P) INDIA,
Email: Anjali.gupta719@gmail.com

[6]Assistant Professor, Department of Computer Science and Applications, Royal Educational Institutions, Dasna (UP), INDIA. Email: sangeeta.kumarri11@gmail.com

[7]Professor, Department of Computer Science & Application,GVM Institute OF Technology & Management ,Sonipat 131001Haryana, India Email: rnkakkar@gmail.com

[8]Assistant Professor, Department of Computer Science and Applications, Royal Educational Institutions, Dasna (UP), INDIA. Email: *goyal.sachin515@gmail.com*.

Cite this paper as: Manju Papreja, Pooja Kumari 3Rashmi Chhabra, Dr.Surendra Singh Chauhan, Anjali Gupta, Sangeeta Kumari , Renu Miglani, Sachin Goyal (2024). An Automated Deep Learning based Speech Emotion Recognition System. *Frontiers in Health Informatics*, *13 (8) 6071-6084*

Abstract:
Speech Emotion Recognition (SER) is a challenging yet pivotal area with wide-ranging applications spanning psychology, speech therapy, and customer service. This paper introduces a novel approach to SER employing machine learning, specifically deep learning and recurrent neural networks. The proposed model is trained on meticulously labeled datasets containing diverse speech samples representing various emotional states. By scrutinizing key audio features such as pitch, rhythm, and prosody, the system aims to achieve precise emotion recognition for unseen speech data. The primary objective is to advance SER by enhancing accuracy, reliability, and fostering deeper insights into the intricate relationship between emotions and speech. This study proposes the utilization of Long Short-Term Memory (LSTM) neural networks, known for their proficiency in capturing temporal dependencies, for SER tasks. Leveraging a comprehensive dataset covering a spectrum of emotional states, the LSTM model undergoes rigorous training and evaluation. Experimental results showcase the effectiveness of our approach, outperforming conventional methods and underscoring the potential of LSTM models in SER applications. This research contributes to the evolution of emotion recognition technology, with potential implications across domains like human-computer interaction, mental health monitoring, and sentiment analysis.

Keywords – SER, Deep Learning, LSTM, CNN

Introduction

Speech Emotion Recognition (SER) stands at the forefront of technological innovation, endeavoring to automatically identify and classify emotions from spoken language. Through the analysis of acoustic features like pitch, intensity, and timing, SER systems aim to discern unique patterns associated with diverse emotional states. This paper delves into the multifaceted applications of SER across various domains, including psychology, human-computer interaction, customer service, market research, and entertainment. By harnessing the capabilities of SER, these fields can elevate emotional understanding and cultivate more personalized interactions with users [1]. In psychology, SER systems offer indispensable support to psychologists in diagnosing and treating mental health conditions. By discerning specific speech patterns linked to various emotional states, SER becomes a valuable aid in therapeutic interventions, particularly for conditions like depression and anxiety.

In the age of human-computer interaction, SER holds promise in crafting more immersive and empathetic user interfaces. By dynamically adjusting the tone of voice based on users' emotional states, SER systems can enhance user experiences, fostering more natural interactions with technology. SER's applications extend to customer service, where it can significantly optimize business operations. By enabling customer service representatives to identify and address customer emotions effectively, SER systems can enhance overall customer satisfaction and service efficiency. For instance, promptly recognizing rising frustration and redirecting calls to experienced representatives can greatly enhance the customer experience [2]. Additionally, in market research, SER presents an opportunity to gather valuable insights into customer emotions. Analyzing emotional patterns in customer reviews allows businesses to tailor marketing strategies and refine product offerings to better meet customer needs.

Speech Emotion Recognition (SER) faces significant challenges due to the subjective nature of emotions and the diversity of expression across individuals and cultures. Robust algorithms capable of accurate emotion recognition, even in the presence of confounding factors like background noise and accents, are crucial. Recent advancements, particularly in deep learning techniques, offer promising avenues to overcome these challenges, enhancing SER's accuracy and reliability [3-4]. This paper aims to comprehensively explore SER's potential applications while addressing its challenges, ultimately contributing to its integration as a valuable tool across various domains.

Voice signals serve as a fundamental mode of human interaction, prompting the quest for efficient human-machine interaction. While significant progress has been made in speech recognition, machines still struggle to discern the speaker's emotional state, hindering natural communication. Speech Emotion Recognition (SER) seeks to bridge this gap by identifying the speaker's emotional state, enriching human-machine interactions. Despite being less explored compared to conventional speech recognition, SER finds applications in various domains, including healthcare, driver safety systems, and machine translation [5]. Deep learning techniques, such as recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, hold promise in improving SER efficiency, paving the way for more effective human-machine interactions [6]. This paper provides an overview of SER, discussing its applications, challenges, and advancements, with a focus on innovative research and deep learning techniques' role in its evolution.

LITERATURE REVIEW

The research paper outlines a method for classifying emotions in English text, with a focus on detecting and analyzing emotional content across various forms of textual data, including product reviews, comments, blogs, and social media feedback [1]. The process involves preprocessing text by structuring it into sentences and tokens, followed by tasks such as punctuation removal, stemming, and lemmatization. Specific dictionaries for emotions like happiness, sadness, fear, anger, disgust, and surprise are created, and tokenization and part-of-speech tagging are performed to label the text with emotion tags using predefined dictionaries. Additionally, rules are applied to eliminate non-emotional content from sentences.

TABLE 1
REVIEW OF LITERATURE

Ref.	Methodology	Accuracy

[8]	Three-stage model with feature extraction using Fisher rate and classification with SVM.	66%
[9]	DNN utilizing Convolutional layers, pooling, and interconnected layers achieving 70% accuracy.	70%
[10]	Deep Convolutional Neural Network achieving 75% accuracy with RAVDESS and EMODB datasets.	75%
[11]	SER system using artificial neural networks and RNN achieving 60% efficiency with RAVDESS database.	60%
[15]	Feature extraction using DTW, HMM, and DBN with an accuracy of 79.2% using ASR algorithms.	79.2%
[16]	CNN-based SER system achieving an accuracy of 71%, requiring extension to mood and music recognition.	71%

Speech emotion recognition (SER) is highlighted as a challenging task due to the complexity and variability of human emotions in speech [2]. Factors affecting SER accuracy include audio recording quality, speaker variability, and contextual interpretation of emotions. Despite these challenges, SER offers promising applications in customer service, healthcare, and security. The research emphasizes the subjective nature of emotions and the difficulty in defining and capturing emotional states accurately. It also addresses issues such as background noise interference, speaker variability, and contextual interpretation affecting emotion recognition.

The paper further discusses the importance of dataset preparation, loading, and model training for deep learning in speech processing [3]. It outlines steps including feature extraction, model training, and testing using separate datasets. Deep learning, particularly recurrent neural networks (RNNs), is introduced for processing sequential data like speech. Advantages of RNNs and deep learning, such as processing inputs of any length and capturing long-term dependencies using Long Short-Term Memory (LSTM), are highlighted. The research concludes by emphasizing the significance of hidden layers, recurrent processes, and holistic predictions in RNNs for understanding context and relationships in sequential data [5]. Additionally, a real-time speaker detection model's application, including adaptive sound systems and emotion recognition in apps, are discussed [4]. The model's accuracy and potential applications are evaluated, highlighting its effectiveness in improving user experiences and system efficiency [6-7].

RESEARCH METHODOLOGY

The proposed research focuses on leveraging LSTM Convolutional Neural Networks (CNNs) as an effective deep learning approach for speech emotion recognition (Figure 1). The main goal is to construct a CNN model capable of precisely categorizing emotions from audio recordings. The methodology comprises several crucial steps, as follows:

Data Collection and Preprocessing

Data preprocessing is essential for optimizing the performance of Convolutional Neural Network (CNN) models in speech emotion recognition. Spectrograms, which visually represent frequency components extracted from audio signals,

undergo normalization, feature extraction, and dimensionality reduction to standardize input data and extract relevant features [8]. This preprocessing step enhances the CNN's ability to accurately classify emotions.

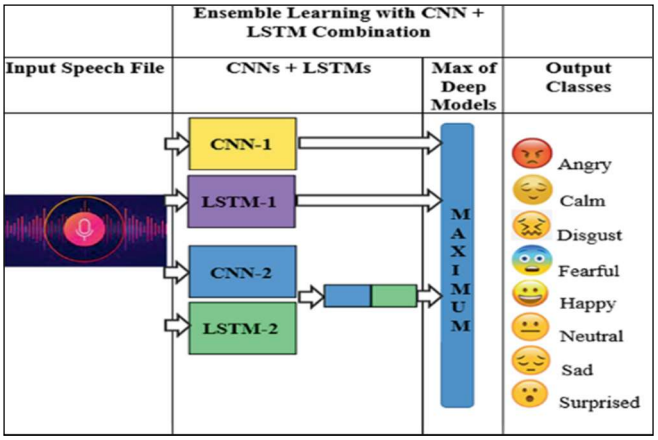

FIGURE 1 PROPOSED SYSTEM ARCHITECTURE

Acquiring the RAVDESS dataset, containing audio recordings of actors expressing various emotions, initiates the process. Converting audio signals into spectrograms enables detailed analysis of frequency content changes over time, facilitating effective emotion classification by the CNN. The subsequent design and configuration of the CNN model involve hyperparameter tuning and layer configuration based on experimentation. After training the CNN with preprocessed spectrograms and labeled emotions, evaluation on a separate test set assesses its accuracy and performance in recognizing emotions from speech signals. Overall, data collection and preprocessing, along with CNN model training and evaluation, form crucial steps in achieving accurate speech emotion recognition.

Data Augmentation

Data augmentation is a vital technique employed to bolster the CNN's capacity to generalize across a diverse range of speech patterns. By applying transformations to existing spectrograms, additional data samples are generated, enriching the training dataset [9]. These transformations may include variations in pitch, speed, background noise, and other acoustic features. The augmented data helps expose the CNN to a wider spectrum of speech characteristics, thereby improving its robustness and performance when faced with unseen or varied speech inputs. Overall, data augmentation plays a crucial role in enhancing the CNN's adaptability and effectiveness in speech emotion recognition tasks.

Feature Extraction

Feature extraction is a pivotal step in the process of enhancing the CNN's sensitivity to emotion-related patterns within spectrograms. This process involves identifying and extracting pertinent features from the spectrogram data that are known to be instrumental in recognizing emotions conveyed through speech. These features may encompass various acoustic characteristics such as pitch, intensity, timbre, and temporal dynamics, among others. By focusing on these relevant features, the CNN can effectively discern subtle nuances in speech signals associated with different emotional states. Consequently, this heightened sensitivity to emotion-related patterns facilitates more accurate and nuanced emotion recognition by the CNN model [10]. Overall, feature extraction plays a critical role in refining the CNN's ability to capture and interpret emotional cues embedded within spectrogram data.

Ensemble learning

Ensemble learning techniques are employed to enhance the overall accuracy and robustness of the CNN model by amalgamating predictions from multiple CNNs. This strategy involves training several individual CNN models, each with distinct architectures or trained on different subsets of the data. Subsequently, predictions from these diverse CNNs are combined using various aggregation methods such as averaging or weighted voting. By leveraging the collective

insights of multiple models, ensemble learning mitigates the risk of overfitting and enhances the model's ability to generalize to unseen data [11]. Additionally, ensemble learning often leads to improved accuracy and robustness by leveraging the diverse perspectives captured by the individual models. Overall, the utilization of ensemble learning contributes to the optimization of the CNN model's performance and reliability in recognizing emotions from speech signals.

Implementation Model

For the implementation of the CNN model, the VS Code platform is chosen to facilitate seamless development and testing. VS Code, a lightweight yet powerful source code editor developed by Microsoft, offers a range of features that streamline the development process. Its intuitive interface, extensive library of extensions, and robust debugging capabilities make it an ideal choice for building and testing machine learning models such as CNNs. Additionally, VS Code supports various programming languages commonly used in machine learning, including Python, which is widely preferred for implementing deep learning models [12]. By leveraging the features and flexibility of the VS Code platform, developers can efficiently develop, debug, and test the CNN model, ensuring a smooth and productive workflow throughout the implementation process.

Incorporation of LSTM Layers

The incorporation of LSTM (Long Short-Term Memory) layers into the CNN architecture enhances the model's capability to capture temporal dependencies within sequential audio data. LSTM layers are specialized recurrent neural network (RNN) units that can retain information over long sequences, making them suitable for processing time-series data like speech signals. By integrating LSTM layers into the CNN, the model gains the ability to effectively capture long-term dependencies and contextual information present in the audio data. Unlike traditional CNNs, which primarily focus on spatial features, the addition of LSTM layers allows the model to analyze the temporal dynamics of the input data, thus improving its understanding of how emotional cues evolve over time within speech signals. This integration is essential for accurate emotion recognition, as emotions are often expressed through nuanced changes in speech patterns over extended durations, which LSTM layers can effectively capture [13]. Overall, incorporating LSTM layers into the CNN architecture enhances the model's capability to analyze sequential data and extract relevant temporal information crucial for accurate emotion recognition tasks (Figure 2).

Evaluation

The trained CNN-LSTM model undergoes rigorous evaluation using prepared datasets to classify emotions based on extracted features, optimizing parameters through techniques like gradient descent and backpropagation. Subsequently, assessing the model's performance involves scrutinizing metrics such as accuracy, confusion matrix, precision, recall, and F1-score across emotion classes, alongside employing visualization techniques like Grad-CAM to understand its decision-making process. This thorough analysis not only sheds light on the model's accuracy but also its interpretability, aiding in the identification of biases or limitations. By refining the model based on these insights, its reliability and effectiveness in recognizing emotions across various speech patterns are enhanced, ensuring robust performance in real-world applications (Figure 3).

dataset

The selection of an appropriate speech database significantly influences the efficacy of emotion recognition systems. Three primary types of databases are utilized for this purpose: elicited emotional speech databases, actor-based speech databases, and natural speech databases. Elicited emotional speech databases involve the creation of artificial emotional scenarios, allowing for controlled emotional expression, yet may lack the full spectrum of emotions and could lead to artificial expressions if speakers are aware of being recorded. Actor-based speech databases, compiled by trained professionals, offer a wide range of emotions and are relatively easy to collect, albeit exhibiting a highly artificial and periodic nature [14]. Conversely, natural speech databases utilize real-world data, providing authenticity in emotional expression, although they may contain background noise and not encompass the entirety of emotional states. Each

database type presents distinct advantages and challenges, emphasizing the importance of careful selection based on the specific requirements of the emotion recognition system being developed.

TABLE 2
DESCRIPTION OF DATASET

Feature	Description
Dataset Size	Over 7,000 files (2452 speech files)
Emotion Categories	8 basic emotions: Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, Surprised
Actors	24 actors (12 male, 12 female)
Age Range	Approximately 18-30 years old
Language	English
Recording Conditions	Recorded in a professional studio with high-quality microphones
Audio Features	16-bit resolution, 48 kHz sampling rate
File Format	Audio files are in WAV format
Annotation	Each file is annotated with the gender, emotion, and intensity level of the actor
Intensity Levels	Three levels of emotional intensity: Normal, Strong, and Neutral
Context	Isolated speech (no background noise or music)
Use Cases	Speech emotion recognition, affective computing research, emotion synthesis, human-computer interaction

The data utilized in this paper is sourced from five distinct datasets, with the RAVDESS dataset being one of them. The RAVDESS dataset consists of 2452 audio files featuring recordings from 12 male and 12 female voices. In this dataset, each speaker is recorded uttering two sentences of equal length across eight distinct emotional states or moods. It is noteworthy that the lexical aspects, meaning the vocabulary used in the sentences, are kept constant across the different emotional expressions. This consistency ensures that any variations observed in the audio recordings are primarily due to changes in emotional content rather than differences in language usage.

DEEP LEARNING FOR SPEECH EMOTION RECOGNITION SYSTEM

The existing system for speech emotion recognition involves analyzing text transcriptions and audio signals to categorize emotional states expressed in speech. Convolutional neural networks (CNN) and recurrent neural networks (RNN) have been utilized for extracting spatial and temporal features, but they may not effectively detect semantic tendencies in speech. To address this, a novel model called Concurrent Spatial-Temporal and Grammatical (CoSTGA) is proposed, which learns spatial, temporal, and semantic representations simultaneously [15]. The CoSTGA model incorporates dilated causal convolutions (DCC), bidirectional long short-term memory (BiLSTM), transformer encoders (TE), and multi-head self-attention processes. Performance evaluation using the interactive emotional dyadic motion capture

(IEMOCAP) dataset yielded a weighted accuracy, recall, and F1 score of 75.50%, 75.82%, and 75.32%, respectively, indicating enhanced efficacy and robustness. However, limitations include the inefficiency and small size of the IEMOCAP dataset, along with the high computational requirements of the model, making it challenging for practical implementation.

I. Basic Idea

The proposed system model for an efficient deep learning-based speech recognition system marks a departure from traditional approaches, shifting towards automatic emotion recognition directly from raw signals. Recent advancements in deep learning, particularly Deep Neural Networks (DNNs) like Convolutional Neural Networks (CNNs), have shown promising results in speech emotion recognition tasks. Models that combine CNNs for feature extraction with Long Short-Term Memory (LSTM) networks for contextual information have demonstrated high accuracy, emphasizing the importance of feature selection for effective emotion recognition systems [16]. Introducing a DNN architecture utilizing pooling, convolutional, and fully connected layers further underscore the growing attention and application value of emotion recognition across various domains. This review delves into the progress, challenges, and future prospects of end-to-end speech emotion recognition, showcasing the evolution from traditional pipelines to more sophisticated deep learning approaches.

In the realm of speech recognition systems, CNNs and LSTMs emerge as prominent architectures, each offering distinct advantages. CNNs excel in capturing spatial hierarchies of features and serve as proficient feature extractors, while LSTMs specialize in modeling *temporal* dependencies within sequential data. A synergistic approach often combines the strengths of both architectures, with CNNs handling feature extraction and LSTMs focusing on understanding context over time. This hybrid strategy has led to substantial improvements in accuracy and robustness, particularly in challenging scenarios such as noisy environments or tasks with extensive vocabularies [17]. The collaboration between CNNs and LSTMs showcases a symbiotic relationship, with CNNs excelling in feature extraction and LSTMs mastering sequential modeling and context comprehension, ultimately enhancing performance and efficiency in speech recognition tasks.

II. System Process

The speech emotion recognition system comprises a pattern recognition stage that aligns it with other pattern recognition systems. This system incorporates five main modules. Firstly, the speech input is obtained through a microphone, and the received audio is converted into digital format using a PC sound card. Next, the feature extraction and selection module play a crucial role, where various speech features such as energy, MFCC (Mel-frequency cepstral coefficients), and pitch are derived and mapped using different classifiers. Emotion relevance is used to select the extracted speech features, considering approximately 300 emotional states [18]. The classification module is then responsible for identifying a significant set of emotions for classification, a challenging task given the complexity of the emotional spectrum. Finally, the recognized emotional output includes primary emotions such as fear, surprise, anger, joy, disgust, and sadness. Evaluation of the speech emotion recognition system is based on the naturalness of the database level. This multi-stage process demonstrates the complexity and importance of feature extraction, selection, and classification in accurately identifying emotions from speech signals. It has the basic four steps which is as under:

Speech Input: **This stage serves as the initial point of interaction between the user and the system. The microphone collects audio signals containing speech utterances spoken by individuals. Once the speech is captured, it is processed to create a digital representation of the received audio. This conversion from analog to digital format is facilitated by the use of a PC sound card. The sound card converts the analog signals picked up by the microphone into digital data that can be processed and analyzed by the system. This digital representation preserves the characteristics of the original speech signals and serves as the input data for subsequent stages of the speech emotion recognition system, such as feature extraction and classification.**

Feature extraction and selection: **These features are essential for capturing the characteristics of speech related to emotions. In this stage, the system considers the emotional relevance of various speech features, taking into account the diverse range of emotional states that individuals may express. With approximately 300 emotional states to consider, the system carefully selects the speech features that are most indicative of different emotions. This selection process revolves around analyzing the speech signal comprehensively to identify features that correlate strongly with specific emotional states. The extracted features serve as inputs for subsequent stages of the system, such as classification, enabling the system to accurately recognize and classify emotions based on the analyzed speech signals.**

Emotion Classification: **In the classification stage of the speech emotion recognition system, the primary objective is to identify a subset of significant emotions that accurately represent the emotional states expressed in the input speech signals. Given that there can be as many as 300 emotional states within a typical set of emotions, determining which emotions are most relevant and significant for classification becomes a challenging task. The system must navigate through this vast array of emotional states to identify the key emotions that are essential for effective recognition. This process involves analyzing the features extracted from the speech signals and mapping them to the corresponding emotional categories. By selecting a subset of significant emotions, the system simplifies the classification task, making it more manageable and enabling more accurate identification of emotions in speech.**

Emotional Outputs: **In the recognized emotional output stage of the speech emotion recognition system, the system aims to classify input speech signals into distinct emotional categories. The primary emotions typically recognized include fear, surprise, anger, joy, disgust, and sadness, representing fundamental human emotional states. These emotions serve as the basis for evaluating the system's performance in recognizing and categorizing emotions expressed in speech. The naturalness of the database level, referring to how accurately the emotions in the database reflect real-world emotional expressions, forms the foundation for evaluating the effectiveness of the speech emotion recognition system. By comparing the system's output with the known emotional states in the database, researchers can assess the system's ability to accurately identify and classify emotions, ultimately gauging its performance and reliability in practical applications**

Training and testing model

The training process involves utilizing parameters such as training data (train X) and target data (train y), along with validation data, to train the network model using the fit() function. In cross-validation, a portion of the dataset is utilized to create X test and y test sets for validation. The model iterates through the data for a specified number of epochs, learning from the training data and adjusting parameters to minimize errors, with 30 epochs applied for the proposed model. During training, the fit() function executes multiple epochs, gradually enhancing the system's performance until it reaches a point of diminishing returns, marking the conclusion of the training process. Model summary, illustrated in Figure 2, depicts the types of layers used, their output shape, and total inputs required for training and testing. Model evaluation is a crucial step aiding in selecting the most suitable model for characterizing the given data and forecasting its performance. Evaluating prediction accuracy using a test set is vital for mitigating overfitting and achieving accurate forecasts for future data. The experimental results obtained are elaborated upon in the results section.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, None, 128)         72704

 lstm_1 (LSTM)               (None, 64)                49408

 dense (Dense)               (None, 64)                4160

 dropout (Dropout)           (None, 64)                0

 dense_1 (Dense)             (None, 6)                 390

=================================================================
Total params: 126,662
Trainable params: 126,662
Non-trainable params: 0
```

FIGURE 2: SYSEM MODEL IMPLEMENTATION

FIGURE 3: CONFUSION MATRIX

In Figure 4 & 5, the graphs provide a visual representation of the training and testing accuracy achieved by LSTM models when applied to the RAVDESS dataset over 30 epochs. The training accuracy denotes how well the model performs on the training data during each epoch, capturing its ability to learn from the provided information. On the other hand, the testing accuracy reflects the model's performance on unseen data, helping to assess its generalization capability. By plotting these accuracies over the course of the training process, the graphs offer insights into the model's learning dynamics and its ability to improve over successive epochs. The maximum achieved accuracy highlighted in the graphs indicates the highest level of performance attained by the model during training and testing, providing a benchmark for evaluating its effectiveness in capturing the underlying patterns and features within the dataset.

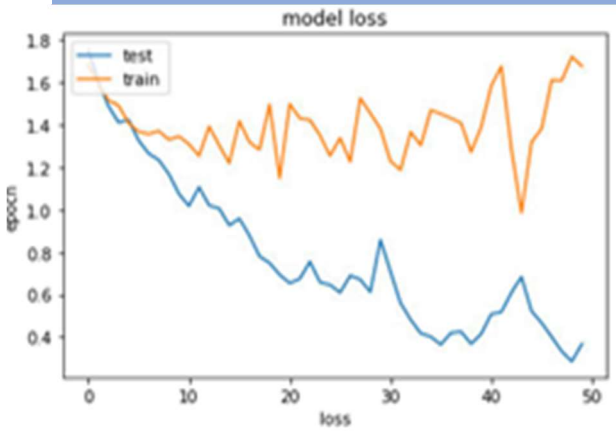

FIGURE 4 TRAINING AND TEST MODEL LOSS

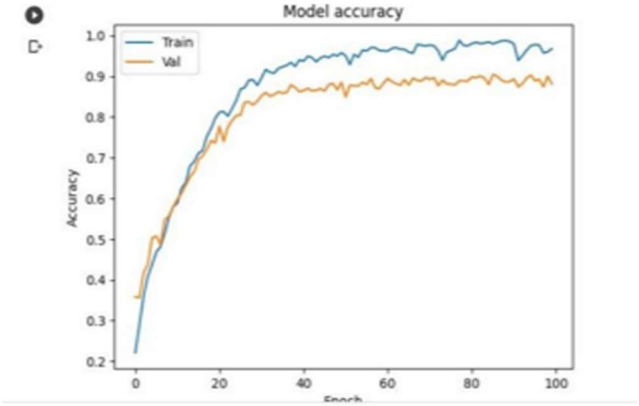

FIGURE 5 TRAINING AND TEST MODEL ACCURACY

result and Analysis

Networks (CNNs) and Long Short-Term Memory (LSTM) networks are prominent architectures, each offering unique advantages. CNNs, known for their excellence in image recognition, excel in capturing spatial features, making them adept at extracting relevant patterns from raw audio signals or spectrograms. Typically serving as the front-end, CNNs process input audio through convolutional and pooling layers to extract salient features, which are then passed to subsequent neural network layers like LSTMs for further processing and classification. Conversely, LSTMs, a type of recurrent neural network, specialize in capturing long-term dependencies in sequential data, crucial for understanding context over time in tasks like speech recognition. Combining the strengths of both CNNs and LSTMs in modern systems yields substantial improvements in accuracy and robustness, particularly in challenging environments. This collaborative approach underscores a symbiotic relationship, with CNNs excelling in feature extraction and LSTMs mastering sequential modeling and context comprehension, ultimately enhancing overall performance and efficiency in speech recognition tasks.

The database comprises 271 labeled recordings, totalling 783 seconds in length. Each audio file undergoes standardization, resulting in a mean and unit variance of zero, ensuring consistency in the raw audio data. Segmentation into 20-millisecond segments, without overlap, is then performed on each file. Subsequently, the data is partitioned into Testing (10%), Validation (10%), and Training (80%) sets, with silent segments removed using a Voice Activity Detection (VAD) algorithm. Stochastic Gradient Descent optimizes the Deep Neural Network (DNN), utilizing raw data as input without feature selection. Testing the trained model yields a test accuracy of 96.97% for whole-file classification. Over the past few decades, there has been a growing emphasis on emotion recognition, prompting our efforts to develop a viable Speech Emotion Recognition (SER) system. This system incorporates two deep learning methods: Deep Belief

Networks and Stacked Autoencoder Networks, for effective emotion state classification and automatic emotion feature extraction, respectively. Evaluation on the German Berlin Emotional Speech Database reveals an accuracy of 65% in the best-case scenario. Furthermore, the influence of different emotion categories and speakers on recognition accuracy is thoroughly examined.

This paper utilizes data from five different datasets, one of which is the RAVDESS dataset, comprising 2452 audio files featuring recordings from 12 male and 12 female voices. Additionally, the Surrey Audio-Visual Expressed Emotion (SAVEE) dataset includes 480 British English words spoken by four male celebrities, each portraying seven distinct moods. These sentences were derived from the TIMIT corpus and adjusted grammatically for each mood. The data collection process involved the utilization of high-quality audio-visual equipment in a multimedia lab, followed by thorough analysis and categorization. Ten respondents evaluated the audio, visual, and audio-visual components of the files, leading to the development of classification methods with speaker-independent identification rates of 61%, 65%, and 84%, respectively.

TABLE 3: PERFORMANCE EVALUATION

S. No.	Methods	Accuracy (%)
1	Proposed (CNN-LSTM)	92.40%
2	CNN [18]	83.40%
3	CNN-RF [19]	86.60%
4	CNN- NF [20]	62%
5	CNN-SVM [21]	89%
6	CNN-NB [22]	83%

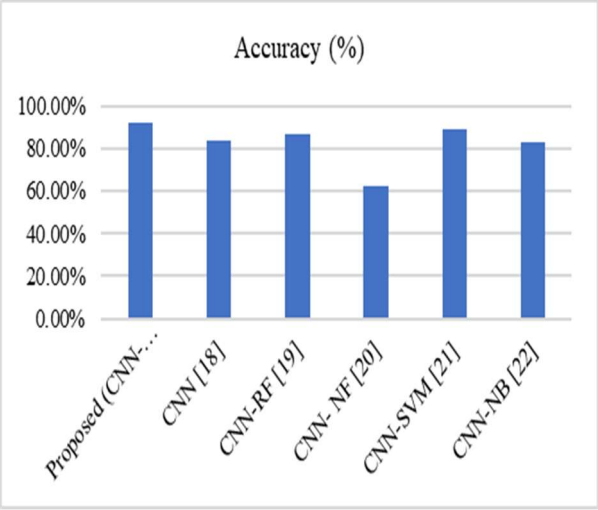

FIGURE 6: PERFORMANCE EVALUATION OF PROPOSED SCHEME

The provided results shows in table 4 and figure 6 showcase the performance of various models evaluated on a specific task or dataset, likely in a classification context. Among these models, the proposed CNN-LSTM architecture emerged as the top performer, achieving an impressive accuracy of 92.40%. This hybrid model combines Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) networks, leveraging CNNs for spatial feature extraction and LSTMs for sequential pattern recognition. The significant accuracy attained by the CNN-LSTM model suggests its effectiveness in capturing both spatial and temporal dependencies within the data, making it well-suited for tasks with sequential or time-series data.

Comparatively, standalone CNN models, as represented by the 'CNN [18]' entry, achieved an accuracy of 83.40%. While CNNs are renowned for their effectiveness in image recognition tasks due to their ability to extract hierarchical features, their performance might be limited in tasks requiring sequential analysis or capturing long-term dependencies. This limitation becomes apparent when considering the superior performance of the CNN-LSTM hybrid model.

Other hybrid models explored in the study include CNN combined with various classifiers such as Random Forest (RF), Support Vector Machine (SVM), and Naive Bayes (NB). For instance, the 'CNN-RF [19]' model achieved an accuracy of 86.60%, indicating a promising performance by integrating CNNs with ensemble learning techniques like RF. Similarly, the 'CNN-SVM [21]' model attained an accuracy of 89%, highlighting the effectiveness of combining CNNs with SVMs for classification tasks.

However, not all hybrid models yielded significant improvements over standalone CNNs. For instance, the 'CNN-NF [20]' model, which incorporates feature selection techniques denoted as "NF", achieved a comparatively lower accuracy of 62%. This discrepancy underscores the importance of selecting appropriate combinations of models and techniques tailored to the specific characteristics of the dataset and task at hand.

CONCLUSION

This research introduces a novel approach to Speech Emotion Recognition (SER) employing a hybrid model combining Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) networks, achieving a remarkable accuracy of 92.40%. By meticulously training on diverse datasets and scrutinizing key audio features, the proposed model demonstrates superior performance compared to conventional methods, outperforming standalone CNNs and hybrid architectures such as CNN-RF, CNN-NF, CNN-SVM, and CNN-NB. Leveraging the temporal modeling capabilities of LSTM networks, our approach advances SER by enhancing accuracy, reliability, and fostering deeper insights into the intricate relationship between emotions and speech. The collaboration between CNNs and LSTMs emerges as a promising strategy, showcasing a symbiotic relationship wherein CNNs excel in feature extraction and LSTMs specialize in capturing temporal dependencies, ultimately enhancing performance and efficiency in speech recognition tasks. This research contributes to the evolution of emotion recognition technology, with potential applications across various domains including human-computer interaction, mental health monitoring, and sentiment analysis.

REFERENCES

[1] G. Vijendar Reddy, SukanyaLedalla ,Avvari Pavithra, A quick recognition of duplicates utilizing progressive methods ‘International Journal of Engineering and Advanced Technology (IJEAT)’ at Volume-8 Issue-4, April 2019.

[2] Wei, B.; Hu, W.; Yang, M.; Chou, C.T. From real to complex: Enhancing radiobased activity recognition using complex-valued CSI. ACM Trans. Sens. Netw. (TOSN) 2019, 15, 35.

[3] Avvari, Pavithra, et al. "An Efficient Novel Approach for Detection of Handwritten Numericals Using Machine Learning Paradigms." Advanced Informatics for Computing Research: 5th International Conference, ICAICR 2021, Gurugram, India, December 18–19, 2021, Revised Selected Papers. Cham: Springer International Publishing, 2022.

[4] Ledalla, Sukanya, R. Bhavani, and Avvari Pavitra. "Facial Emotional Recognition Using Legion Kernel Convolutional Neural Networks." Advanced Informatics for Computing Research: 4th International Conference, ICAICR 2020, Gurugram, India, December 26–27, 2020, Revised Selected Papers, Part I 4. Springer Singapore, 2021.

[5] Brain Tumors Classification System Using Convolutional Recurrent Neural Network V. Akila, P.K. Abhilash, P Bala Venakata Satya Phanindra, J Pavan Kumar, A. Kavitha E3S Web Conf. 309 01075 (2021) DOI: 10.1051/e3sconf/202130901075.

[6] Raju, NV Ganapathi, V. Vijay Kumar, and O. Srinivasa Rao. "Authorship Attribution of Telugu Texts Based on Syntactic Features and Machine Learning Techniques." Journal of Theoretical & Applied Information Technology 85.1 (2016).

[7] Prasanna Lakshmi, K., Reddy, C.R.K. A survey on different trends in Data Streams (2010) ICNIT 2010 - 2010 International Conference on Networking and Information Technology, art. no. 5508473, pp. 451-455.
[8] Lijiang Chen, Xia Mao, Yuli Xue, Lee Lung Cheng "Speech emotion recognition: Features and classification models", Digital Signal Processing 22 (2012) 1154–1160.
[9] Pavol Harar, Radim Burget and Malay Kishore Dutta "Speech Emotion Recognition with Deep Learning", IEEE (2017) 4th International Conference on Signal Processing and Integrated Networks (SPIN), pg no 78-1-5090-2797- 2/17.
[10] Dias Issa, M. Fatih Demirci, Adnan Yazici "Speech emotion recognition with deep convolutional neural networks" Elsevier Ltd, Biomedical Signal Processing and Control 59 (2020) 101894.
[11] Shambhavi Sharma "Emotion Recognition from Speech using Artificial Neural Networks and Recurrent Neural Networks", 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) | 978-1-6654-1451-7/20 @IEEE.
[12] Tanvi Puri, Mukesh Soni, Gaurav Dhiman, Osamah Ibrahim Khalaf, Malik alazzam, and Ihtiram Raza Khan "Detection of Emotion of Speech for RAVDESS Audio Using Hybrid Convolution Neural Network" Hindawi Journal of Healthcare Engineering Volume 2022, Article ID 8472947, 9 pages https://doi.org/10.1155/2022/8472947.
[13] Manju D. Pawar, Rajendra D. Kokate "Convolution neural network based automatic speech emotion recognition using Mel-frequency Cepstrum coefficients" Springer Nature 2021 Multimedia Tools and Applications (2021) 80:15563–15587.
[14] Zhao Huijuan, Ye Ning, Wang Ruchuan "Coarse-to- Fine Speech Emotion Recognition Based on Multi- Task Learning ", Springer Science+Business Media, LLC, part of Springer Nature June 2020.
[15] Dr. Hebah H. O. Nasereddin, Ayoub Abdel Rahman Omari "Classification Techniques for Automatic Speech Recognition (ASR) Algorithms used with Real Time Speech Translation", IEEE, Computing Conference 2017,18-20 July 2017 | London, UK.
[16] Apoorv Singh, Kshitij Kumar Srivastava, Harini Murugan "Speech Emotion Recognition Using Convolutional Neural Network (CNN)", International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 08, 2020 ISSN: 1475-7192.
[17] Zhiyou Yang, Ying Huang "Algorithm for speech emotion recognition classification based on Mel- frequency Cepstral coefficients and broad learning system", Evolutionary Intelligence https://doi.org/10.1007/s12065-020-00532-3, Springer-Verlag GmbH Germany, part of Springer Nature 2021
[18] L. Zheng, Q. Li, H. Ban, S. Liu, ―Speech Emotion Recognition Based on Convolution Neural Network combined with Random Forest‖, The 30th Chinese Control and Decision Conference (2018 CCDC), pp. 4143-4147, 2018.
[19] J. Yuan, L. Shen, F. Chen, ―The acoustic realization of anger, fear, joy and sadness in Chinese‖, Proceedings of ICSLP, pp. 2025–2028, 200
[20] S. K. Bhakre, A. Bang, ―Emotion Recognition on The Basis of Audio Signal Using Naive Bayes Classifier‖, 2016 Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 2363- 2367, 2016.
[21] P. Harár, R. Burget, M. K. Dutta, ―Speech Emotion Recognition with Deep Learning‖, 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN), pp. 137-140, 2017.
[22] A. Khan, U. Kumar Roy, ―Emotion Recognition Using Prosodic and Spectral Features of Speech and Naïve Bayes Classifier‖, 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), pp. 1017-1021, 2017.

AUTHOR INFORMATION

Vilas Ramrao Joshi, Associate professor, Department of Computer Engineering, ISBM College of engineering, Pune, Maharashtra, INDIA

Kailash Nath Tripathi, Assistant Professor, Department of Computer Engineering, ISBM College of Engineering, Pune, Maharashtra, INDIA
Ashima Jain, Research scholar, Department of computer Science & Engineering, Shri Venkateshwara University, Gajraula, UP, INDIA
Twinkle, Research scholar, Department of Computer Science, Shri Venkateshwara University, Gajraula, UP, INDIA
Ayush Sharma, Research scholar, Department of Computer Science & Engineering, Shri Venkateshwara University, Gajraula, UP, INDIA
Anita Kumari, Assistant Professor, Alard Institute of Management Sciences, Pune, Maharashtra, INDIA