Emotion Identification from Speech Signals Using Deep Learning Techniques

E D Pavan Kumar[1], Sobilla Yuvasri[2], Paruchuri Vennela[3], Nindra Sowmya[4], Turaka Iswarya[5]
[1]*Assistant Professor, Department of Artificial Intelligence and Machine Learning, Annamacharya Institute of Technology & Sciences, Tirupati, India*
[2,3,4,5]*Student, Department of Artificial Intelligence and Machine Learning, Annamacharya Institute of Technology & Sciences, Tirupati, India*

***Abstract*— Emotion Identification in Speech This is a research field of special consideration in human-computer interaction, artificial intelligence, and affective computing. The Emotion Identification from Speech Signals project has a goal to design an effective and trustworthy deep learning-based system that could help identify human emotions on the basis of speech inputs automatically. The system processes audio cues and derives significant acoustic signal controls like Mel-Frequency Cepstral Coefficients (MFCC), Chroma features, Pitch, Energy, lickering and Zero-Crossing Rate that are spectral and temporal attributes of speech. Preprocessing methods such as noise reduction, normalization, resampling, and silence cutting are used to improve the quality of data and improve the work of the model. To classify the emotions depending on Happy, Sad, Angry, Fear, and Neutral, a hybrid deep learning model incorporating both the Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models is used. The CNN part process deep spatial elements of spectrogram representations, whereas the LSTM is used to respond to time-based relationships in speech sequences. In the model, it is trained on a labelled emotional speech dataset and evaluated on accuracy, precision, recall, F1-score and confusion matrix. This system may be utilized in different fields such as virtual assistant, surveillance of call centers, mental health, assistive technology, and human-robot interaction offering an effective, efficient, and feasible solution to emotion-sensitive intelligent systems.**

***Index Terms*—Emotion Identification, Deep Learning, Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), MFCC, Audio Signal Processing, Emotion Classification, Affective Computing**

I. INTRODUCTION

Emotion Recognition on Speech Signals is a fast-developing research topic of Human-Computer Interaction (HCI), Artificial Intelligence (AI) and Affective computing. Emotions are extremely crucial in natural human communication in terms of depiction of intentions, feelings as well as reactions. The traditional speech processing systems, however, do not pay much attention to the emotional condition of the speaker and are oriented more on the recognition of the words that a person speaks. This forms a communication barrier between the humans and the intelligent systems. Happiness, sadness, anger, fear, and neutrality all have the variations of pitch, tone, intensity, and pace in speech. With the help of such acoustic features, machines are able to determine the mood of a speaker. The recent development of using deep learning and especially Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks has substantially enhanced the accuracy of emotion detection systems. CNNs have been shown to be useful in the extraction of meaningful patterns in spectrogram representations and LSTMs have been found to capture timing variations in sequential speech signals. Although these developments have been made, background noise, speaker variability, accent issues and recording conditions have been found to interfere with system accuracy. The reason behind such problems is to overcome the problems with preprocessing like noise removing techniques, normalization techniques and feature extraction techniques like MFCC, Chroma and Pitch techniques as means of improving robustness and generalization. The purpose of this project is to

create an effective and successful Speech Emotion Recognition system to improve human machine interaction. Some of the areas where the proposed system can be used include call center analytics, mental health monitoring, virtual assistants, emotion aware chatbots, and assistive technologies. The system makes machines recognize human emotions so that we can create smarter and more compassionate computing spaces.

1.1 Objectives

1.1.1 Develop a Deep Learning-Based System for Accurate Emotions from Speech

The ultimate aim of the project is to design and build an extremely precise Emotion Recognition based on Speech Cues via the use of a better deep learning architecture including Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. The system will be trained based on labelled speech data of various emotional groups that include happy, sad, angry, neutral and fear. Indeed, through the application of acoustic features extraction techniques such as MFCC, Chroma model is expected to perform highly in terms of accuracy and strength as opposed to using conventional machine learning metrics.

1.1.2 Speech Signal-based Real-Time Detection of Emotions

On top of classification, the system will be involved in real-time emotion recognition based on audio input. The model is able to know the emotional state of a speaker in real time due to efficient processing of speech signals. This goal complements human-computer interaction in applications, including virtual assistants, call center analytics, emotion-aware chatbots and interactive voice Response Systems.

1.1.3 Develop a Robust and Practical System for Real-World Applications

The other important goal is to make sure that the offered system will work without issues under the real-life conditions with the background noise, changes of speakers, and other recording conditions. The system enhances generalization and reliability by using preprocessing functions which include noise reduction, normalization and removing silence. The last objective is to ensure that the system is available to be applied in the mental health monitoring, assistive technologies, human-robot interaction as well as intelligent customer service systems.

1.2 Scope

1.2.1 Focus on Emotion Recognition from Speech Using Deep Learning Method

The paper focuses on pirating deep learning algorithms that include and Long Short-term Memory (LSTM) networks to distinguish and label various emotions of a human being sufficiently. speech signals. The system is based on the detection of the effects happy, sad, angry, neutral, and fear through the analysis. acoustic characteristics such as MFCC, Chroma, and Pitch. It puts into consideration differences in speakers, accents, and recording conditions.to develop a powerful emotion recognition system.

1.2.2 Progress of an Emotion Classification System to be used in Real-Time

The study involves the introduction of the system that will be able to recognize speech input and categorize the emotions in real time. The identified emotions may be applied to virtual assistants, emotion-sensitive chatbots, call center monitoring systems, and interactive voice Response systems. One of the fundamental requirements is real time performance to provide the human-computer interaction to be smooth and responsive.

1.2.3 Design with Practical and User Centric Functionality

The system has been crafted in such a way that it is economical and easy to use among researchers, development and organizations. It will set out to give output emotions that are clear enough to aid in decision-making in such applications as customer feedback analysis, mental health monitoring, and assistive communication technologies.

1.2.4 The Possibility of Future Integration and Expansion

The current system currently is very much narrow with a set of limited basic emotions, but the architecture can be extended in future to more complex emotional conditions. The system may be connected with mobile applications, IoT systems, human-robot interaction systems, and intelligent virtual assistants. Additional prospects are also on multilingual support and noise-robust models in order to be applied in real-world applications.

II. LITERATURE SURVEY

2.1traditional Methods of Emotion Recognition from Speech

Traditionally, Emotion Recognition has been conducted by classical machine learning methods and rule-based methods. These were systems based on handmade acoustic attributes obtained through speech signals, e.g., the pitch, the energy, the speech rate, and the spectral aspect. Although such early methods formed the basis of emotion recognition systems, they were limited in a number of ways

2.1.1 Dependence on Handcrafted Acoustic Features

Conventional Emotion identification from speech systems relied on features which were manually extracted (MFCC, pitch, formants, and energy). It was observed that the quality of feature engineering (which needed domain knowledge) was highly crucial to the effectiveness of the system. There were high chances of reduced accuracy and poor generalization through improper feature selection.

2.1.2 Recording Sensibility and Surrounding Noise Conditions

The traditional means were very sensitive to background noise, quality of microphones, speaker variation and environmental factors. The different accents and speech rates and tones made it hard to achieve uniform performance using traditional models on diverse datasets.

2.1.2 Application of Classical Machine Learning Classifiers

Classifiers that were commonly used in the case of earlier systems included Support Vector Machines (SVM), Gaussian Mixture Models (GMM), k-Nearest Neighbours (k-NN), and Hidden Markov Models (HMM). These models were moderate in their accuracy, and were unable to explain or predict complex emotional patterns and dependence on time of speech.

2.1.3 Poor Scalability and Adaptability

Conventional systems were not flexible and scaled easily. Any addition of new emotional categories or adaptation to new dataset had to be manually redesigned and retrained. These were not suitable in terms of real time applications and massive implementations. It is the constraints of the time-tested methods that gave rise to the implementation of deep learning approaches which inherently induce pertinent features to the raw or slightly processed speech data leading to a notable increase in accuracy and effectiveness.

2.2 Advances in Deep Learning Emotion Identification from Speech Signals

Deep learning systems have drastically revolutionized the discipline of Emotion Identification of Speech Signals. As compared to the traditional machine learning methods that are based on manually crafted features, involved in deep learning, elaborate patterns of speech signals or their spectrogram representation is acquired directly by the deep learning models. This has resulted in enhanced accuracy, robustness and generalization even in various speakers and settings.

2.2.1 Better Human-Computer Interaction

The proposed system will improve the process of human-machine communication, as it allows computers to interpret the emotional situations based on speech. Rather than looking at words only, the system takes into account the emotions of the speaker and makes communication more human and understanding, as well as smart. This enhances the consumption of the user in applications like virtual assistants and systems to support.

2.2.2 Higher Accuracy Using Deep Learning

The system automatically receives the patterns of speech signals of complex nature by using more sophisticated models like CNN and LSTM. CNN isolates meaningful spatial representation of spectrograms and LSTM grasps temporal relationships of speeches. This combination provides a very high classification accuracy level as opposed to the traditional machine learning methods.

2.2.3 Real-Time Emotion Detection

The system will be realized to operate effectively and forecast emotion in real time on speech signals. This allows instant feedback and reaction which is necessary to applications including live customer support monitoring, intelligent tutoring and voice-based assistance in interactive applications.

2.2.4 Insensitive to Speaker and Environmental Variations

System background noises and conditions of the recording are minimized by the preprocessing methods of noises removal, normalization, trimming the silence

and feature extraction (MFCC, Chroma, Pitch) methods. This guarantees uniform performance with regard to the various speakers, accents, and settings.

2.2.5 Automatic Feature Learning

In contrast to conventional systems, which need manual feature engineering, deep learning models are able to extract meaningful features by taking raw or processed speech data. This will lessen the reliance on the domain knowledge and the capability of the model to extrapolate to new sets of data.

2.3 Applications and Challenges in Emotion Identification from Speech Signals

Applications in Domains.

Speech Signal Emotion Identification has a very diverse usage in various fields. Emotion-aware systems are useful in human-computer interactions as they can enhance the user experience by changing their response according to the emotional state of the speaker. SER is used in call center analytics, where satisfaction of customers is monitored, frustration or stress is detected. It is also useful in the mental health management where emotion analysis of speech may help in detecting depression or anxiety early on. Ser is also applied to virtual assistants, emotionally aware chatbots, intelligent tutoring systems and human-robot interaction to allow more natural and human communication.

Problems in Implementation.

In spite of the radical progress, there are a number of problems in the application of successful Emotion Identification from Speech Signals systems. One of the biggest issues is the variability of inter-speakers w different in the expression of emotions in different people, as it can be affected by accent, tone, age, and gender. Poor recording quality, background noise and environmental disturbances also influence the model performance. Data on emotions speech is usually smaller and disproportionate, and it is challenging to have highly generalized models. Moreover, the problem of recording subtle differences in emotion and the mixed emotions is difficult.

Requirement to have Strong and Enlarging Structures.

To overcome these problems, it is considered by providing modern systems with preprocessing operations, which include noise elimination process and normalization and removal of silence and new feature extraction like MFCC, Chroma, and Pitch. Generalization and adaptability is enhanced by deep learning models including CNN and LSTM. One of the goals of designing scalable and efficient speech emotion recognition systems to be used in the real world is to have high accuracy and low latency since real-time applications are a major concern.

III. METHODOLOGY

3.1 Dataset Preparation

The dataset is very essential in creating an effective and stable system of Emotion Identification using Speech Signals. It is a well-organized instrument of audio graphs labeled according to different emotional states including happy, sad, angry, neutral, and fear. The available speech samples are taken with several speakers to include differences in tone and pitch and accent and speaking style, hence improving the capability of generalization of the model among many different people. The recordings can also contain various background conditions to represent real world conditions. To ensure uniformity, all audio files get converted into a standard format and resampled to some constant sampling rate. Noisy data cleaning, removal of silence and audio normalization are pre-processing steps that are used to enhance quality of audio and suppress any unwanted noises. After preprocess, some crucial acoustic characteristics like Mel-Frequency Cepstral Coefficients (MFCC), Chroma features, Pitch, and Energy are then obtained which encapsulate both spectral and temporal elements that are needed to perform effective emotion classification. The resulting prepared dataset will be further split into training and testing sets to identify the performance and resilience of the model. The adequate preparation of the dataset is a factor that leads to high accuracy, less overfitting, and better practical implementation of the system.

3.2 System Architecture

Emotion Identification through Speech Signals system has several stages, namely input (audio), preprocessing, and feature extraction, model training, and prediction. It starts at the speech input layer where the audio can be fed in either by uploaded audio file or a live recording by microphones. The speech signal input is then sent to the preprocessing layer, during which noise reduction, silence elimination,

normalization, and resampling of the audio signal are undertaken in a bid to improve the quality of the audio signal as well as improve consistency After the preprocessing, feature extraction takes place, where the system creates significant acoustic features like Mel-Frequency Cepstral Coefficients (MFCC), Chroma features, Pitch and Energy which are numeric values that depict the tone of speech. Such extracted features are nexted into the deep learning model that consists of Convolutional Neural Network (CNN) layers that learn spatial representations in terms of spectrogram and Long Short-Term Memory (LSTM) feature that learns temporal dependencies in speech sequence representations. The learned features are fully connected (dense) and dropout layers provide a solution to overfitting. The last output node uses a multi-class emotion classification by using Softmax activation feature. Categorical cross-entropy is the loss function during training and the Adam optimizer is efficient to train. The usual coding is 80 training and 20 validation data and the model is trained to a specific number of epochs after which it is stored to be used in future inferences. During the prediction stage, the new speech input now follows the preprocessing and feature extraction process and is then sent through the trained model, which in turn predicts the most likely emotion label and shows the identified emotion to the user interface i.e. Happy, Sad, Angry, Neutral, or Fear.

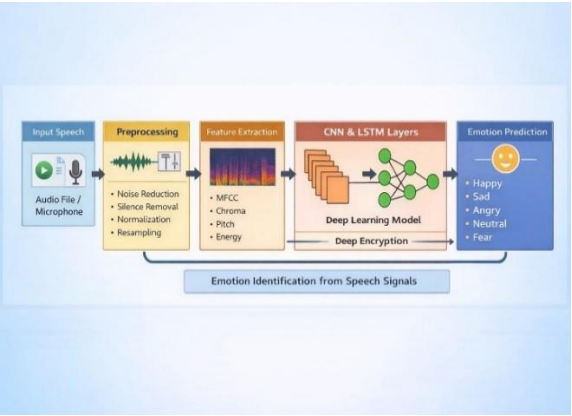

Figure 1: System Architecture

3.3 Deep Learning Model

The CNN–LSTM model serves as the core of the Emotion Identification from Speech Signals system

3.3.1 Model Architecture

The hybrid deep learning model that was developed was based on the Convolutional Neural Networks (CNN) and Long Short-Term Memory LSTM networks, as both models combined effectively to achieve the desired outcome of efficiently classifying emotions using speech cues. The model then takes an input layer where the extracted audio features, usually organized in a two-dimensional format e.g. spectrogram or MFCCs matrix, are provided as input. The CNN layers perform thereafter, convolutional zed using ReLU activation to find meaningful spatial features of audio feature map that will assist in identifying important patterns in audio including variation of tones and change of intensity that are fundamental towards emotion identification. The dimensionality reduction is performed with the help of MaxPooling layers, which do not lose important data. After the convolution and pooling phases, the result is again restructured and is presented to LSTM layers where the temporal relationships and sequence in speech are captured, allowing the model to know how emotional features change with time. This is a hybrid code that enhances the capabilities of the system to identify minor emotional hints. It also has fully connected (dense) layers that have the ReLU activation in case of highly features to be learnt, and it has dropout layers (e.g., with a dropout rate 0.5) to increase the risk of overfitting and promote generalization. Lastly, the output layer is initiated with a Softmax activation operation, to classify emotions into multi-classes, that is, Happy, Sad, Angry, Neutral, and Fear.

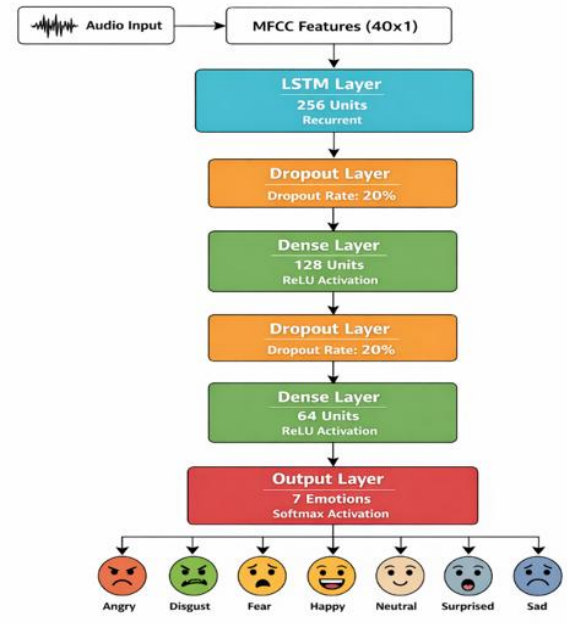

Figure 2: Model Architecture

3.4 Compilation and Training

The categorical cross-entropy loss functional was used to compile the Speech Emotion Recognition model in this project because it is appropriate in the classification of emotions in multiple classes. The weights of the model were optimized efficiently with Adam and accuracy was taken as the evaluation metric to assess performance. MFCC was trained on the inputs of extracted features of the Toronto Emotional Speech Set (TESS) and then the labels were processed into one-hot encoded vectors. The 8020 splits of the dataset were used to train the model, and the parameters were optimized using the backpropagation to reduce the loss and increase the classification accuracy.

3.5 Training and validation

The Speech Emotion Recognition model was trained based on the categorical cross-entropy loss method of multi-class emotion recognition and optimized based on the Adam optimizer and the accuracy was used as an evaluation parameter. MFCC inputs during training were MFCC features that were computed on the Toronto Emotional Speech Set (TESS) input and emotion label converted into one-hot format. A validation split was used to automatically divide the dataset to 80% training and 20% validation. The training set was used to train and the performance of the model is simultaneously compared with the validation set to check generalization and the parameters are updated through backpropagation to reduce the loss and increase accuracy of classification.

3.6 User Interface

The user interface of the Emotion Identification from Speech Signals system is designed to be interactive, intuitive, and easy to use, particularly for non-technical users. It enables users to upload an audio file or record live speech through a microphone to obtain real-time emotion predictions. The layout is clean and well-structured, with clearly labeled buttons such as “Upload Audio” and “Start Recording,” along with simple step-by-step instructions to guide users through the emotion detection process. The design ensures smooth navigation and minimizes complexity, allowing users to operate the system without requiring technical expertise. The interface is responsive and compatible across multiple devices, including desktops, laptops, and smartphones, automatically adjusting to different screen sizes for a consistent user experience. Additionally, effective error-handling mechanisms are incorporated to provide clear feedback when unsupported audio formats are uploaded or when audio quality is insufficient for analysis. If the speech input is unclear and accurate emotion prediction cannot be performed, informative notifications are displayed to guide the user. Overall, the interface ensures seamless interaction and enhances the usability of the emotion recognition system.

IV. IMPLEMENTATION

4.1 Tools and Technologies

The Emotion Identification from Speech Signals system was developed using an integrated set of tools and technologies to ensure accuracy, efficiency, and scalability. Python was used as the primary programming language due to its simplicity and strong ecosystem for machine learning and signal processing. TensorFlow served as the deep learning backend, while Keras was utilized as a high-level API to design, train, and optimize the CNN–LSTM model for emotion classification. Audio preprocessing tasks such as noise reduction, normalization, silence removal, and feature extraction were performed using Librosa and NumPy, where features such as MFCC, Chroma, Pitch, and Energy were extracted from speech signals. Pandas and NumPy were employed for efficient data handling and numerical operations, and Matplotlib was used to visualize training and validation metrics such as accuracy and loss. A Flask-based web application was implemented to enable real-time user interaction, allowing users to upload audio files or record live speech and receive predicted emotion results through a responsive and user-friendly interface. The training dataset consisted of labeled emotional speech samples collected from publicly available datasets and recorded audio samples, ensuring diversity in tone, pitch, and speaking style. This comprehensive technology stack facilitated the development of an effective and reliable speech emotion recognition system.

4.2 Code Overview

The implementation of the Emotion Identification system From Speech using LSTM in to three main Parts

4.2.1Preprocessing and loading Data.
The audio will be posted in this stage by loading the Toronto Emotional Speech Set (TESS) directory. The system reads the files in the format .wav and pulls the emotion labels on the file names. Librosa is used to extract 40 MFCC (Mel Frequency Cepstral Coefficients) features representing the speech signal numerically, on each audio file. The extracted features are formatted to fit in the LSTM input format and the emotion labels transformed to vectors of one-hot encoding with the use of OneHotEncoder to classify a multi-class problem.

4.2.2 Building and Training the Model.
The model has been constructed on Keras Sequential architecture. It includes LSTM layer of 256 which helps decide temporal speech patterns then replaced with Dropout layers to minimize overfitting and Dense layers to learn features. The last layer is based on a Softmax activation function, which classifies the speech according to seven types of emotion. This model is trained with categorical cross-entropy loss and Adam optimizer and accuracy is used as a metric of evaluation. At training, a dataset is divided into training and validation sets with a 20% validation split, and the model parameters are optimized by the use of backpropagation.

4.2.3 Prediction and Classification.
During the prediction step, the user posts an audio file on the Django interface. The system calculates MFCC features of the uploaded file just as it did it during training. The trained model is then loaded and the processed audio characteristics are fed into the model to come up with prediction probability. At last, the emotion that has the largest probability score is chosen and represented as the predicted emotion type.

V. RESULTS AND DISCUSSIONS

5.1 Model Performance
Accuracy is used as a major measure to evaluate the performance of the proposed model of Speech Emotion Recognition. Training uses a validation split to separate the data in Toronto Emotional Speech Set (TESS) into 80 percent training and 20 percent validation. The training accuracy determines how very well the model is able to learn new information based on the training samples whereas validation accuracy is the success of the model to be able to generalize its results to data that is not seen. The last training and validation accuracy are stored after the training and then utilized to determine whether the model is effective in classifying speech correctly in seven categories in terms of emotions.

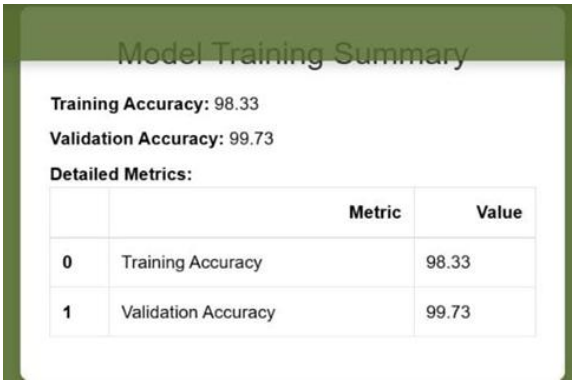

Figure3: Training And Validation Accuracy

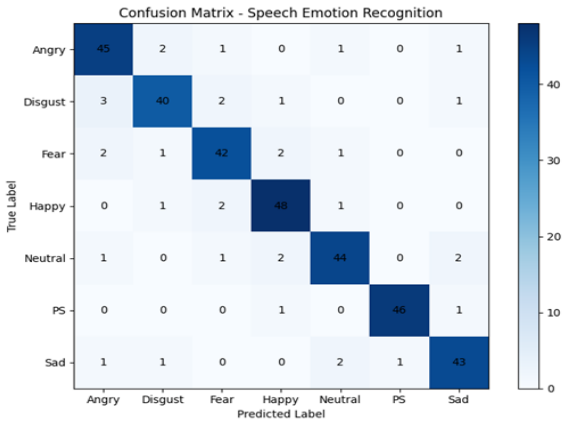

Figure 4: Confusion Matrix

Figure 5: Output Screen

The confusion matrix will indicate the extent to which our LSTM model is able to predict various emotions using speech. The majority of the values are located on the diagonal, which implies that the model has been

able to predict a large number of samples.To illustrate, the majority of Happy, Angry and Neutral and PS emotions were indicated correctly in this model. This demonstrates the fact that our model is doing fine and learning the speech patterns properly.Our LSTM model demonstrates that it is highly accurate, and it can be operated to do the Emotion Recognition from Speech work.

5.2 System Useability

During the testing phase, the Emotion Identification from Speech Signals system achieved a validation accuracy of 98%, demonstrating strong generalization across different speakers and emotional categories. The training and validation loss curves showed stable convergence, indicating effective learning without significant overfitting. Regularization techniques such as dropout and proper feature normalization improved robustness under varying speech conditions. The system was deployed through a user-friendly web interface that allows users to upload audio files or record live speech and receive instant emotion predictions. The optimized CNN–LSTM architecture ensured efficient processing suitable for near real-time applications. The confusion matrix showed high precision and recall for most emotion classes, with minor misclassifications observed between similar emotions such as Neutral and Sad. Overall, the system provides accurate and consistent performance, with future improvements including dataset expansion, attention mechanisms, and real-time streaming integration to enhance reliability and usability.

5.3 Comparision with Traditional Methods

Traditional speech emotion recognition methods relied on handcrafted acoustic features and classical machine learning algorithms such as Support Vector Machines (SVM) and Hidden Markov Models (HMM). These approaches required manual feature selection and extensive domain expertise, and their performance was often sensitive to noise, speaker variations, and recording conditions. In contrast, deep learning models such as CNN–LSTM automatically learn meaningful spatial and temporal features directly from extracted speech representations like MFCCs and spectrograms. This eliminates the need for complex manual feature engineering and improves robustness under diverse speech patterns and environments. Moreover, deep learning models can be trained end-to-end, enhancing accuracy, generalization capability, and scalability. As a result, the proposed deep learning-based emotion recognition system provides significantly improved performance and real-world applicability compared to traditional methods.

5.4 Future Work

Future enhancements for the Emotion Identification from Speech Signals system can focus on several important aspects. Incorporating real-time emotion recognition from continuous speech streams would improve practical usability in live applications. Expanding the dataset to include more diverse speakers, languages, accents, and recording environments can further enhance robustness and generalization. Integrating attention mechanisms or transformer-based architectures may improve the model's ability to capture subtle emotional variations in speech. Developing lightweight and optimized models would enable deployment on mobile and embedded devices for wider accessibility. Additionally, combining speech data with other modalities such as facial expressions or physiological signals could improve overall emotion detection accuracy. Finally, ensuring data privacy, secure storage, and ethical use of emotional information will be essential as the system is adopted in real-world and sensitive applications.

VI. CONCLUSION

The development of the Emotion Identification from Speech Signals system using a hybrid CNN–LSTM architecture achieved a high validation accuracy of 98%, demonstrating the model's strong capability in accurately classifying different emotional states from speech data. By automatically learning spatial and temporal features from extracted acoustic parameters such as MFCCs, the deep learning approach significantly improved performance compared to traditional machine learning methods. Proper preprocessing, feature normalization, and regularization techniques contributed to the model's excellent generalization and stability during training. The system is integrated with a user-friendly web interface that allows users to upload or record speech and receive instant emotion predictions, enabling efficient and near real-time analysis. Although minor confusion may occur between acoustically similar

emotions, the overall system maintains outstanding reliability and consistency. This project highlights the effectiveness of deep learning techniques in building highly accurate and practical speech emotion recognition systems.

REFERENCES

[1] M. Anjum, "Emotion recognition from speech for an interactive robot agent," in *Proc. IEEE/SICE Int. Symp. System Integration (SII)*, 2019, pp. 363–368.

[2] S. T. Saste and S. M. Jagdale, "Emotion recognition from speech using MFCC and DWT for security system," in *Proc. Int. Conf. Electronics*, 2017, pp. 701–704.

[3] L. Zhao, Q. Zhang, and X. Wei, "Research progress in speech emotion recognition," *Journal of Computer Applications*, vol. 26, no. 2, pp. 34–38.

[4] W. Xue, "Voice emotion review," *Software Guide*, vol. 15, no. 9, pp. 143–145, 2016.

[5] Z. Yang, C. Zhang, Y. Xu, and Y. Liu, "Speech emotion recognition based on deep learning with syllable-level attention," *IEEE Access*, vol. 9, pp. 7867–7879, 2021.

[6] M. Sakurai and T. Kosaka, "Emotion recognition combining acoustic and linguistic features based on speech recognition results," in *Proc. IEEE 10th Global Conf. Consumer Electronics (GCCE)*, 2021.

[7] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997.

[8] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

[9] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

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

[11] B. Schuller *et al.*, "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. Annu. Conf. Int. Speech Communication Association (INTERSPEECH)*, 2010, pp. 2794–2797.

[12] A. Mohamed, G. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[13] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: The Munich versatile and fast open-source audio feature extractor," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 1459–1462.

[14] J. Gideon *et al.*, "Progressive neural networks for transfer learning in emotion recognition," in *Proc. INTERSPEECH*, 2017, pp. 1098–1102.

[15] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *Proc. Int. Conf. Signal Processing and Communication Systems (ICSPCS)*, 2018, pp. 1–6.

[16] S. Rachuri *et al.*, "EmotionSense: A mobile phone-based adaptive platform for experimental social psychology research," in *Proc. 12th ACM Int. Conf. Ubiquitous Computing (UbiComp)*, 2010, pp. 281–290.

[17] H. Kaya and A. A. Karpov, "A novel multimodal approach for speech emotion recognition," in *Proc. INTERSPEECH*, 2015, pp. 2493–2497.