IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, November 2024

Speech Emotion Recognition Using Feedforward Neural Network

Deepali Sale[1], Anand Bhagat[2], Pranit Bhalekar[3], Rohit Gorde [4], Mahendra Gayakwad[5]
Associate Professor[1]
Students [2,3,4,5]
Dr. D. Y. Patil College of Engineering and Innovation, Pune, India.

Abstract: *Speech Emotion Recognition (SER) has gained increasing attention due to its application in fields such as human-computer interaction, healthcare, customer service automation, and affective computing. This review focuses on the application of Deep Neural Networks (DNNs), specifically Feedforward Neural Networks (FNNs), in detecting and classifying emotions from speech signals. Recent advancements in deep learning have significantly enhanced the ability of machines to interpret emotional cues from auditory data. This paper discusses the motivation, objectives, and scope of employing DNN FNN architectures for SER. We also address the technical feasibility and potential of this approach for real-time applications. By reviewing existing literature, this study aims to provide insights into the current progress, challenges, and future prospects of SER systems.*

Keywords: Speech Emotion Recognition (SER), Feedforward Neural Networks (FNN), Deep Neural Networks (DNN), Affective Computing, Human - Computer Interaction (HCI), Real- time Systems

I. INTRODUCTION

Speech plays a pivotal role in human communication, not only conveying information but also emotions and intentions. As systems capable of understanding speech become more prevalent, recognizing the speaker's emotional state has become a key area of research. Speech Emotion Recognition (SER) systems aim to identify emotions like happiness, sadness, anger, and fear from speech signals. The ability to automatically recognize emotions enhances human-computer interactions by making them more intuitive and empathetic. Traditional SER systems typically relied on handcrafted features and classical machine learning models such as Support Vector Machines (SVM) and Hidden Markov Models (HMM). However, these methods faced limitations in handling the high variability of speech data and failed to achieve high accuracy across diverse datasets. The rise of deep learning has led to the exploration of DNNs for SER. Feedforward Neural Networks (FNNs), which consist of multiple fully connected layers, have shown promise in capturing the intricate patterns present in speech signals. In this paper, we focus on reviewing the use of FNNs in the domain of SER. We examine their architecture, strengths, and limitations compared to other deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). We also discuss the challenges of adapting FNN-based SER systems for real-time deployment.

II. MOTIVATION

Human emotions are an essential aspect of communication, influencing the meaning and context of spoken language. With the growing demand for more intelligent and responsive systems, integrating emotional understanding into human-computer interactions is crucial. Traditional machine learning techniques for Speech Emotion Recognition (SER), such as Support Vector Machines and Hidden Markov Models, have demonstrated limited performance in capturing the complex, non-linear relationships present in speech data. These models often rely on manually engineered features, which are prone to inconsistencies due to variations in speakers, recording conditions, and emotional expressions.Deep Neural Networks (DNNs), particularly Feedforward Neural Networks (FNNs), offer a more robust solution by learning hierarchical features directly from the data, bypassing the need for manual feature extraction. The architecture of FNNs allows the network to recognize subtle patterns in speech signals, improving emotion classification accuracy. As real-time systems and affective computing become more relevant in areas such as virtual

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, November 2024

assistants, customer service, and healthcare,the need for accurate and efficient SER systems is more pressing than ever. Our motivation lies in addressing these gaps by employing DNN-based models to develop more reliable and scalable SER systems.

III. OBJECTIVES

The main objective of this project is to design and implement a Speech Emotion Recognition system using DNN Feedforward Neural Networks. Specifically, we aim to:

- Investigate the effectiveness of FNNs for extracting meaningful features from speech data and classifying emotional states.
- Evaluate the performance of our proposed model across multiple datasets, including real-world noisy environments, to ensure robustness and generalization.
- Optimize the model to achieve real-time processing capabilities, making it suitable for integration into interactive systems.
- Explore future extensions, such as incorporating multimodal emotion recognition (e.g., combining facial expressions) to enhance accuracy and applicability across various domains.

IV. LITERATURE SURVEY

In recent years, the field of Speech Emotion Recognition (SER) has gained traction due to advancements in deep learning and its applications across human-computer interaction, healthcare, and education. Several studies have explored architectures like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for capturing emotional nuances in speech. For instance, Zhang et al. [1] employed a CNN model using Mel-frequency cepstral coefficients (MFCCs) as input features, achieving substantial improvements over traditional machine learning models. However, the CNN's high computational requirements posed challenges for real-time applications, which our project addresses by implementing a lightweight CNN that maintains a balance between efficiency and performance.

Chen and Wang [7] utilized Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs) to capture temporal dependencies in emotion-laden speech on the EMO-DB dataset, resulting in significant accuracy gains. Nevertheless, these RNN-based methods often encounter computational complexity, making them less ideal for real-time SER applications. Our project mitigates this by integrating an optimized model, designed to reduce processing time while retaining RNN advantages in temporal feature extraction.

In another approach, Liu et al. [8] developed a hybrid CNN-LSTM model to combine spatial and temporal feature extraction capabilities on the CREMA-D dataset. While this hybrid model achieved promising results, it was limited by the dataset size, raising concerns about generalizability. Our project addresses this issue by employing data augmentation techniques and parameter optimization to enhance both generalization and efficiency.

Khan et al. [9] introduced a lightweight CNN model optimized for real-time emotion detection, achieving lower computational complexity compared to conventional models. However, accuracy was sacrificed, especially when compared to larger, more complex models. Our project aims to improve accuracy by refining feature selection and model parameters without increasing computational demands.

Smith et al. [10] presented a noise-robust Deep Neural Network (DNN) capable of recognizing emotions within noisy environments. While effective under certain noise conditions, the model struggled to generalize to diverse noise types. To enhance robustness, our project incorporates pre-processing techniques that allow for a wider array of acoustic settings.These studies demonstrate that while deep learning has advanced SER, challenges remain regarding real-time application, noise robustness, and model generalization. Our work contributes by addressing these issues with a lightweight, noise-robust model optimized for efficiency across varied acoustic environments.

Feasibility

The feasibility of our Speech Emotion Recognition (SER) project is evaluated based on three key factors: technical feasibility, economic feasibility, and operational feasibility.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, November 2024

Technical Feasibility

Our SER project leverages established technologies in machine learning, particularly deep learning techniques like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). These technologies are compatible with widely used software libraries such as Keras and TensorFlow, making development and model training both accessible and efficient. The use of MFCCs (Mel-Frequency Cepstral Coefficients) as primary input features is also well-suited to capturing the essential elements of speech for emotion recognition. Given the availability of robust computational resources (e.g., cloud-based GPU options), our project can be executed effectively without excessive hardware requirements. Furthermore, the deployment of the model in a web-based application is feasible with existing platforms, making it accessible to a broad user base.

Economic Feasibility

The economic feasibility of the SER project is favourable. Open-source libraries (Keras, TensorFlow, Librosa) minimize software costs, while pre-existing datasets such as RAVDESS and CREMA-D are freely available for research purposes, eliminating the need for data collection expenses. Moreover, cloud services offer cost-effective solutions for scalable computation, allowing us to use GPU resources on-demand and reducing infrastructure costs. This project's potential for real-world applications in customer service, healthcare, and education suggests a return on investment through practical deployment and integration, making it economically viable.

Operational Feasibility

Operationally, the SER project is designed to be user-friendly and accessible. The web-based interface ensures that users can easily interact with the application by uploading audio files or recording directly. The model's training and implementation pipeline is designed for real-time or near-real-time performance, which meets user expectations for responsiveness in emotion recognition systems. Additionally, the lightweight model architecture is intended to work efficiently in environments with limited resources, making it deployable across different systems without significant operational burdens. The project aligns with current needs in various industries, supporting its long-term operational feasibility.

V. METHODOLOGY

Data Collection

We use publicly available datasets like RAVDESS which contain label audio samples for various emotions such as anger, happiness, sadness, and surprise. These datasets are suitable due to their balanced audio quality and variety in emotional expressions.

Data Pre-processing

The audio files are pre-processed to extract essential features. We utilize:

- **Feature Extraction**: Mel-Frequency Cepstral Coefficients (MFCCs) are extracted from the audio signals as they capture timbral texture, which is crucial for differentiating emotions in speech.
- **Standardization**: Features are standardized to ensure uniform scaling, which enhances model performance.
- **Encoding Labels**: The emotion labels are converted into categorical values using one-hot encoding, enabling compatibility with neural network models.

Model Architecture

The model used in our project is a Deep Neural Network (DNN) built using a Sequential model architecture. This model consists of:

- **Input Layer**: Accepting 40-dimensional MFCC feature vectors.
- **Hidden Layers**: Two dense layers with 128 and 64 neurons, respectively, both activated using ReLU (Rectified Linear Unit) activation. Dropout layers (with a rate of 0.5) are included after each hidden layer to prevent overfitting.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, November 2024

- **Output Layer**: A dense layer with 8 neurons (representing each emotion) and a softmax activation function to provide probabilities for each emotion class.

Model Compilation and Training

The model is compiled with the categorical cross-entropy loss function, an Adam optimizer, and accuracy as the evaluation metric. Training is conducted over 50 epochs with a batch size of 32, and validation is done using a separate portion of the dataset (20% split for testing).

Evaluation

After training, the model's performance is evaluated on the test dataset. Key performance metrics like accuracy are recorded, and confusion matrices are analyzed to assess the model's effectiveness in classifying emotions accurately.

Deployment

We deploy the trained model in a web application where users can record or upload an audio file. The application processes the file, extracts features, and inputs them into the model for prediction. The predicted emotion is then displayed to the user.

User Interface (UI) Design

A simple, user-friendly web interface allows users to interact with the SER system. Using technologies like Flask for the backend, the UI facilitates real-time emotion detection by accepting user audio inputs.

Algorithm

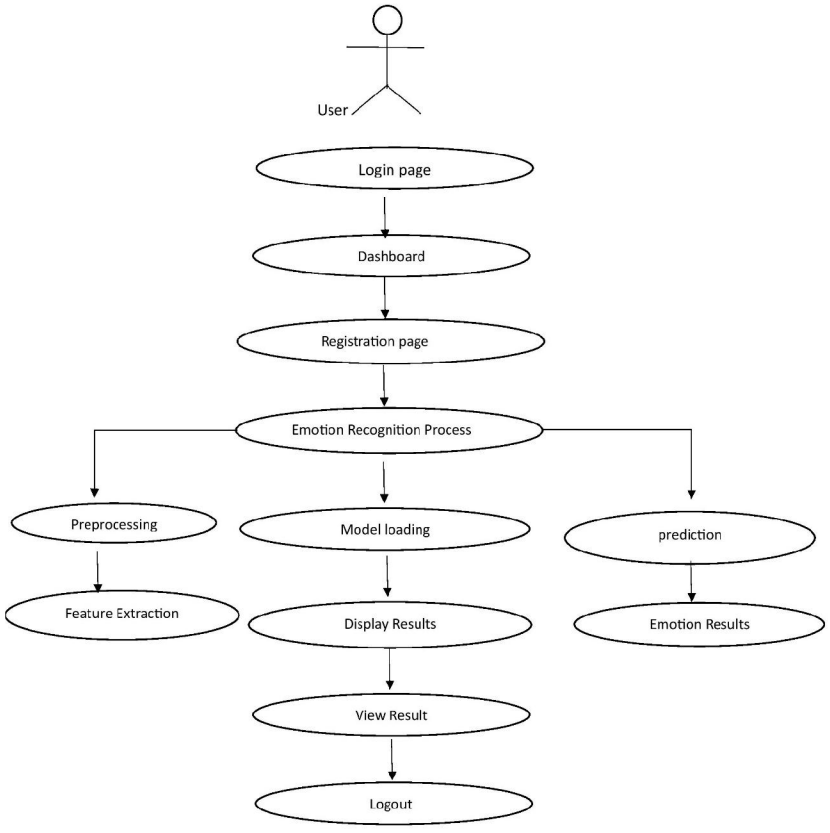

Fig. 1. Flowchart Diagram Representing SER

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, November 2024

In this Speech Emotion Recognition (SER), we employed a sequence of algorithms to process and classify emotions from audio data. Initially, we applied pre - processing techniques, such as noise reduction, to clean the audio input. We then extracted Mel-frequency cepstral coefficients (MFCCs), which are crucial for capturing the unique characteristics of speech. The extracted features were normalized using Standard Scaler to improve model performance. We utilized a Feedforward Neural Network (FNN) for emotion classification, with soft max activation in the final layer to obtain probability distributions across emotion classes. Finally, we evaluated the model's accuracy and effectiveness through metrics such as precision, recall, and F1 score.

VI. ARCHITECTURE

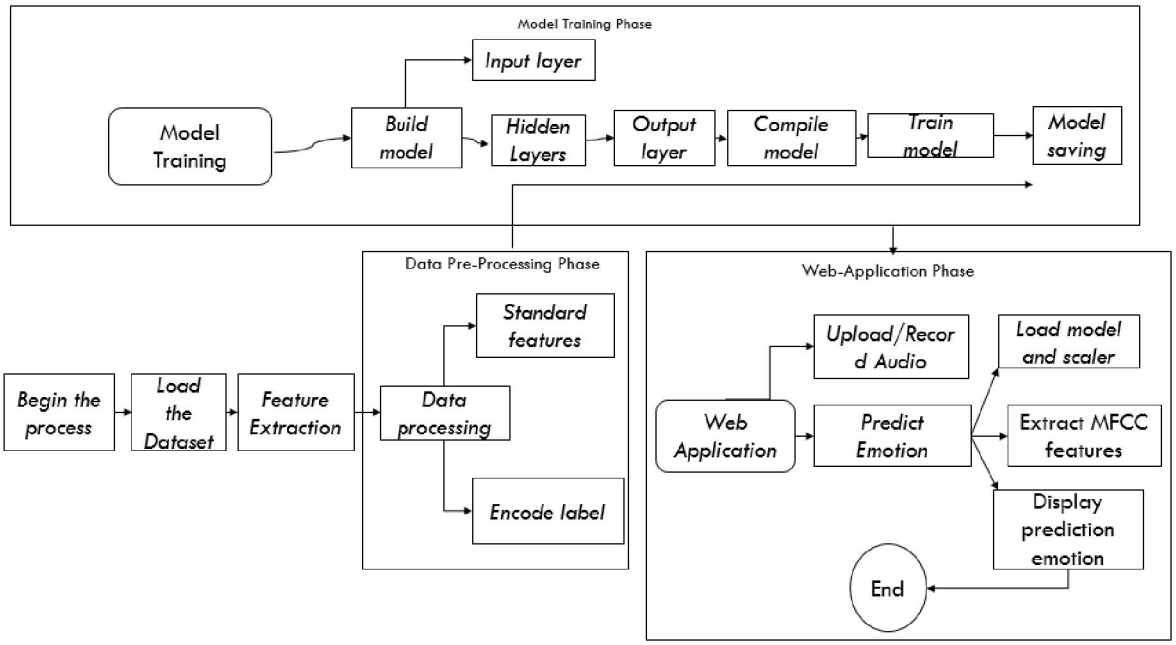

Fig. 2. Architecture of SER

VII. CONCLUSION

This review highlights the promising role of Feedforward Neural Networks in the domain of Speech Emotion Recognition. FNNs provide a flexible and powerful tool for learning from speech data, offering improvements over traditional machine learning models. By leveraging deep learning architectures, SER systems can more accurately capture the complexities of human emotions, potentially leading to more responsive and adaptive human-machine interactions. However, challenges remain, particularly in optimizing these models for real-time use cases. While FNNs are easier to implement than more complex architectures like RNNs or CNNs, they may still require significant computational resources and model optimization to meet the demands of real-time applications. Furthermore, there is scope for further exploration into hybrid models that combine the strengths of different neural network architectures to improve the overall performance of SER systems. Future research should focus on addressing these challenges and expanding the applicability of FNN-based SER systems to more diverse contexts, including multi-modal emotion recognition and cross-linguistic studies. With continuous advancements in deep learning and computational hardware, the future of SER systems looks promising, with potential applications in various industries, from healthcare to customer service and beyond.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, November 2024

REFERENCES

[1]. Zhang, X., et al., "Emotion Detection Using CNN," 2020.
[2]. "Feature Extraction and Classification of Emotion Recognition Using Deep Learning,"*Journal of Computational Science*, vol. 40, pp. 101073, 2020.
[3]. "Emotion Recognition from Speech Signals Using DNN and SVM,"*Applied Sciences*, vol. 10, no. 15, 2020.
[4]. "Speech Emotion Recognition Using Deep Learning Techniques: A Review,"*International Journal of Advanced Computer Science and Applications*, vol. 11, no. 6, pp. 167-178, 2020. doi:10.14569/IJACSA.2020.0110621.
[5]. "Emotion Recognition from Speech Signals Using DNN and SVM,"*Applied Sciences*, vol. 10, no. 15, 2020.
[6]. "A Novel Approach for Speech Emotion Recognition Based on DNN and GMM,"*IEEE Transactions on Affective Computing*, vol. 12, no. 3, pp. 512-525, 2021.
[7]. Chen, L., and Wang, H., "Temporal Feature Extraction with LSTM for SER," 2021.
[8]. Liu, Y., et al., "Hybrid CNN-LSTM Model for Speech Emotion Recognition," 2022.
[9]. Khan, R., et al., "Real-time Emotion Detection with Lightweight CNN," 2023.
[10]. Smith, J., et al., "Noise Robust SER Using Deep Neural Networks," 2023.
[11]. Kumar, A., & Singh, R. "A Comprehensive Review of Speech Emotion Recognition Systems." 2024