Speech Emotion Recognition Using Machine Learning

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

Bachelor of Technology

In

Computer Science and Engineering

School of Engineering and Sciences

Submitted by

Soumya Karuturi-AP22110010175

Jyotika Tammineedi-AP22110010457

Indra Narayana Korrakuti-AP22110010448

Under the Guidance of

Dr. Murali Krishna Enduri

SRM University–AP

Neerukonda, Mangalagiri, Guntur

Andhra Pradesh – 522 240

[November, 2024]

Certificate

Date: 27-Nov-24

This is to certify that the work present in this Project entitled "**Speech Emotion Recognition Using Machine Learning**" has been carried out by **Soumya Karuturi** under my supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology in **Computer Science and Engineering**.

Supervisor

(Signature)

Prof. / Dr. Murali Krishna Enduri

Designation: Asst. Professor

Affiliation: SRM University-AP

Co-supervisor

(Signature)

Prof. / Dr. [Name]

Designation,

Affiliation.

Acknowledgements

We express my deep sense of gratitude and indebtedness to my Project guide Dr. Murali Krishna Enduri, Computer Science and Engineering, School of Engineering and Sciences, SRM University–AP, Neerukonda, Mangalagiri, Guntur, Andhra Pradesh for his guidance and valuable suggestions during the project work.

We are ever so thankful to all of those who have contributed to the successful completion of this project in any way. It is because of their dedication, hard work, patience, and guidance that this has been achieved. A special thanks to our mentor who has been with us from the beginning, offering his advice and support. He was generous enough to give us advice and guide us throughout our project work. We are also immensely grateful to our friends for their support, as well as anyone else who has contributed to this project. Without them, we would not have been able to reach the end goal. The project has come such a long way since it was first conceived, and that is due in large part to the effort of the people who have been involved. We are conscious of the fact that a project of this magnitude could not have been successful without the help of many individuals. We are extremely grateful to all those who have aided us in any way, whether it be directly or indirectly. We thank you all for your support. In summary, we would like to express our sincerest appreciation to all those who have been a part of this project. Without their help, we would not have been able to complete it. We are humbled by the amount of support we have been given, and we thank all those involved for their invaluable contributions.

Table of Contents

Abstract

Recognizing human emotions through machines is a complex and challenging task. Machine learning models seek to automate this process, allowing machines to learn and adapt to emotional cues. Despite advancements, achieving accurate emotion recognition from speech with high performance remains difficult. Emotion recognition and classification are essential in the field of human-computer interaction, with emotions typically identified through an analysis of behaviors such as facial expressions, vocal tone, and body movements. This study investigates two methods of feature extraction to enhance speech emotion recognition. The open-source Audio-Visual Database of Emotional Speech (AVDES) was utilized, containing data representing seven emotions such as angry, disgusted, fearful, happy, neutral, surprised, and sad. A multilayer perceptron (MLP) classifier, a popular supervised learning algorithm, was employed for emotion classification. The performance of the proposed model was compared to similar studies, and the results were thoroughly evaluated. The proposed model achieved an overall accuracy of 97.59% in classifying seven distinct emotions using the AVDES dataset. Furthermore, experimental results demonstrate that the ensemble learning approach delivers highly competitive performance compared to other state-of-the-art methods.

Keywords: Affective computing, Audible feature, Machine learning, MLP, Speech emotions

List of Abbreviations

LSTM	Long Short-Term Memory
MFCC	Mel-Frequency Cepstral Coefficients
STFT	Short-Time Fourier Transform
dB	Decibels
ReLU	Rectified Linear Unit
Hz	Hertz
SR	Sampling Rate
IP	Interactive Python

List of Tables

List of Figures

1. Introduction

1.1 Background

Emotion identification systems play a crucial role in understanding human feelings and behavior (Albahri et al., 2016). Two of the most commonly used methods for emotion expression and recognition are affective speech and facial expressions (Akçay and Oğuz, 2020). Recognizing emotions through facial expressions involves interpreting external visual cues, such as changes in facial muscles or expressions, to infer emotional states. However, such systems are limited as they rely solely on external indicators, which may not accurately represent a person's internal emotional state. This discrepancy often leads to misclassification of emotions, as outward expressions may not align with the individual's actual feelings or intentions (Eyben, 2016). Emotions are increasingly recognized as critical factors in decision-making across various domains, including individual, governmental, and commercial sectors (Schuller et al., 2004). For individuals, emotions shape preferences, relationships, and choices. In the governmental and commercial contexts, understanding collective emotional trends is vital for policymaking, marketing strategies, and customer engagement. The vast amount of information generated and shared on the internet has further underscored the need for advanced analytical techniques to process and utilize this data effectively. Social media platforms, in particular, have become key channels where people openly share their daily activities, reactions, and emotions. These platforms provide a wealth of unstructured data that reflects real-time emotional states, presenting both opportunities and challenges for emotion identification systems (Mustafa et al., 2018). By leveraging such data, emotion recognition technologies can contribute to deeper insights into human behavior, enhancing decision-making processes in a variety of fields.

Emotions are an integral part of daily life, shaping how we perceive and respond to various situations (Zamil et al., 2019). These strong feelings often arise from our experiences or interactions with others and play a significant role in influencing human behavior and decision-making. Emotions are particularly impactful in various domains, such as e-commerce, restaurants, movies, and customer satisfaction with services or products. For instance, emotions often guide consumer choices, from selecting a product to forming opinions about a service. Social media platforms have increasingly incorporated emotional expressions into their user interfaces. A notable example is Facebook, which introduced reaction options like "angry," "happy," "love," and "surprise" to allow users to convey their feelings toward comments, pictures, or events (Mekruksavanich et al., 2020). These tools highlight the growing importance of emotions in digital communication and the need for systems that can effectively interpret emotional expressions. Emotion analysis, a subfield of sentiment analysis, focuses on interpreting judgments, responses, and feelings derived from various data sources, including text. This field has become essential in disciplines such as data mining, web mining, and social media analytics (Liu et al., 2019). Emotions are key indicators of human behavior, providing insights into individuals' attitudes and preferences. The proliferation of social media platforms has amplified the relevance of emotion analysis, as users often express their opinions and feelings on diverse issues through text, enabling the extraction of valuable knowledge. While traditional sentiment analysis primarily focuses on polarity-categorizing text as positive, negative, or neutral-modern approaches delve deeper into understanding specific emotions such as anger, happiness, sadness, and even urgency. These advancements have made emotion analysis a crucial tool in interpreting human behavior in online interactions, offering practical applications in customer service, marketing, and social behavior research. The ability to analyze emotions accurately enables organizations to better understand and respond to their audiences, fostering more meaningful interactions and informed decision-making.

1.2 Statement of the problem

Understanding the implication of the past, present and future patterns of different speeches and audio visuals function is increasingly important in speech recognition. Automatic speech emotion recognition (SER) is a very essential research issue in many fields of study because speech is the most widely used and efficient communication tool. Practical real-time systems have several problems and they are complex and create difficulty to develop the recognition system, because of difficulty to identify what is the correct emotion for the given speech, for a real-time environment emotions are a minimum distance between each other or there is an ambiguity to identify one emotion for the other. It is also hard to identify humans' emotion manually because they express similar emotions in different ways this makes the recognition system difficult and create misclassification of emotions. Another main challenge of SER in the real environment is the existence of noise that is occurred during speech generation and transmitting time. Noise is the main factor in decreasing SER performance. In a previous study, even the researcher use deep learning methods to identify emotions. However, mostly the performance of SER is degraded by the existence of noise in the speech signal, and scarcity of available data especially in the real environment.

Traditional speech emotion recognition models face significant limitations in effectively extracting and classifying emotional features from speech signals. These limitations arise primarily due to the inherent complexity of audio speech signals, which include variations in tone, pitch, and rhythm. To address these challenges, advanced and intelligent models are necessary to extract essential features, minimize misclassification, and reduce the time required for labeling emotions. The performance of SER systems is influenced by several factors, including:

Database Characteristics: The number and diversity of speakers stored in the database significantly impact recognition accuracy.

Feature Extraction Methods: The techniques used to extract relevant features from speech signals are critical to capturing the emotional content accurately.

Classification Techniques: The choice of classification algorithms determines how well the system can differentiate between emotional states.

Speaker Gender: Variations in pitch and tone between male and female speakers can affect recognition accuracy.

Emotional Complexity: A single speech signal may contain overlapping or mixed emotional states, complicating the classification process.

Manually identifying emotions for millions of users and aggregating this information to make rapid and efficient decisions is an increasingly daunting task, particularly with the exponential growth of speech data on social media platforms. As users frequently share their thoughts and emotions through voice-based communications, the demand for automated, scalable, and accurate SER systems becomes even more pressing. Developing models that can overcome these challenges is essential for achieving reliable and efficient emotion recognition in diverse and real-world contexts.

1.3 Objectives

General Objective

The main objective of this study was to develop a speech emotion recognition using machine learning.

1.3.1 Specific Objectives

The specific objectives of this project are as listed below:

- To analyze emotional trends and variations using datasets representing seven distinct emotions.
- To identify and apply suitable machine learning techniques through a systematic review and develop an emotion detection model.
- To examine trends in speech emotions and evaluate the model's performance using a Multi-Layer Perceptron (MLP) approach.

1.4. Project Questions

Based on the stated objectives, the following questions were used to guide the project process and the answer from the findings of the study.

- How to enhance speech noise removal techniques in SER?
- What feature extraction approach is better in SER?
- How to ensemble auditory and spectrogram features MLP model?

1.5. Scope of the study

This study focuses on the assessment of speech emotion recognition by analyzing the emotional states conveyed through human voices. It aims to investigate trends and variations in voice patterns associated with different emotional expressions. The project emphasizes understanding how vocal features correlate with emotional states, providing insights into the dynamics of speech and emotion. The scope also includes evaluating changes in voice patterns to identify and classify emotions, contributing to advancements in emotion recognition systems.

1.6 Significance of the study

Speech emotion recognition holds significant value and finds applications in various fields, including psychotherapy and advertising. In psychotherapy, it aids in understanding and monitoring emotional states, enhancing the effectiveness of therapeutic interventions. In advertising, it helps analyze consumer responses and emotions, allowing for the development of more targeted and impactful marketing strategies.

1.7 Limitations

This study is centered on developing a speech emotion recognition system designed to identify four basic emotions: sadness, anger, happiness, and neutrality, in an uncontrolled environment. Emotions beyond these four categories are not considered within the scope of this system. Additionally, the study does not address discrepancies between a person's internal emotional state and the emotions expressed through speech signals. This includes instances where the displayed emotion does not align with the speaker's true intentions. A significant limitation encountered during the project is the challenge of acquiring authentic, real-world datasets from uncontrolled environments, which may impact the system's training and performance.

2. Related work

Speech Emotion Recognition (SER) entails the development of a model capable of identifying emotions from speech signals (Noroozi et al., 2017). This task utilizes audio data to analyze and interpret human emotions, thereby enhancing human-computer interaction and other real-world applications. The objective is to construct a Deep Learning model proficient in accurately classifying emotions such as anger, happiness, sadness, and others (Schuller et al., 2004). The dataset employed for this research comprises 2,800 audio files, each categorized into one of seven emotions: angry, disgust, fear, happy, neutral, ps (pleasant surprise), and sad. The dataset is balanced containing an equal number of samples for each emotion, thus ensuring unbiased training and evaluation. The Deep Learning model utilized is a Long Short-Term Memory (LSTM) network, a specialized type of Recurrent Neural Network (RNN). LSTMs are particularly suitable for sequential data such as audio due to their ability to learn temporal patterns and dependencies. The model consists of LSTM layers to capture sequential information, dense layers for classification, and dropout layers to mitigate over fitting (Akçay and Oğuz, 2020).

The model is trained on a combination of original and augmented data. Data augmentation techniques, including time stretching, noise injection, and time shifting, are employed to introduce variability to the dataset, thereby enhancing the model's robustness (Schröder, 2001). The training process is optimized using the Adam optimizer and the categorical cross-entropy loss function, which are appropriate for multi-class classification problems (Liu et al., 2019; Mekruksavanich et al., 2020). The model's performance is assessed using metrics such as training and validation accuracy. It achieves a test accuracy of 97.59%, demonstrating its capacity to generalize effectively to unseen data. This high accuracy underscores the model's efficacy in recognizing emotions from speech signals.

Speech Emotion Recognition has diverse applications. It can facilitate the development of emotion-based virtual assistants and chatbots, thereby improving the quality of human-computer interaction. In customer service, sentiment analysis can provide insights into user satisfaction. SER can also be utilized in mental health monitoring, aiding in the tracking of stress or emotional well-being. Furthermore, it has applications in autonomous vehicles, enabling emotion-aware systems for adaptive user experiences.

3. Methodology

The dataset utilized for this investigation comprises 2,800 audio recordings, each categorized with one of seven emotions: angry, happy, sad, neutral, fear, disgust, and ps (pleasant surprise). Each emotion category contains 400 samples, ensuring the dataset is balanced and equitable for model training.

3.1 Feature Extraction

To analyze the audio files, Mel-Frequency Cepstral Coefficients (MFCCs) were extracted, which are widely employed features for processing speech data. These coefficients capture salient characteristics of sound, such as tone and frequency. For each audio file, 40 MFCC features were extracted, summarizing the audio into a compact, meaningful representation.

Data augmentation techniques were also applied to enhance the dataset's diversity and robustness. This involved:

Time Stretching: Altering the audio speed without modifying the pitch.

Noise Injection: Introducing small amounts of random background noise to simulate various environments.

Time Shifting: Adjusting the audio slightly forward or backward in time.

These techniques facilitated an increase in dataset size and improved the model's capacity to handle real-world variability.

3.2 Data preparation

The extracted MFCC features were organized into arrays for model input. Labels (emotions) were converted into one-hot encoded vectors, rendering them suitable for classification. Post-augmentation, the dataset encompassed 5,600 samples:

Original audio samples	2,800
Augmented versions	2,800

3.3 Model design

A Long Short-Term Memory (LSTM) model was employed for emotion classification. LSTM is a type of neural network particularly adept at processing sequences such as speech. The model structure incorporated:

Input Layer: Processes the 40 MFCC features for each audio sample.

LSTM Layer: Captures patterns in the sequence of MFCC features with 256 units (neurons).

Dense Layers: Three fully connected layers with 128, 64, and 32 neurons, utilizing ReLU activation, which facilitates learning complex patterns.

Dropout Layers: Incorporated to mitigate overfitting by randomly deactivating neurons during training.

Output Layer: Utilizes softmax activation to predict probabilities for the seven emotion categories.

3.4 Training and validation

The dataset was partitioned into 80% for training and 20% for testing. Within the training set, 20% was further reserved for validation. The training utilized the Adam optimizer and categorical cross-entropy loss function, which are well-suited for multi-class classification problems. To optimize the training process, the following were implemented:

Model Checkpoints: Automatically preserved the best-performing model.

Early Stopping: Terminated training when validation accuracy ceased to improve to prevent overfitting.

The model was trained for up to 50 epochs, with a batch size of 64. However, training typically concluded earlier due to early stopping.

3.5 Evaluation

Following training, the model was evaluated on the test dataset. It achieved a high accuracy of 97.59%, demonstrating its efficacy in identifying emotions from speech recordings. This result indicates that the LSTM model successfully captured the patterns in the audio data for emotion recognition.

3.6 Visualization

To elucidate the model's performance, the training and validation accuracy across epochs were plotted. The graph illustrated that the model's validation accuracy consistently improved during the initial epochs, with minimal overfitting. Detailed

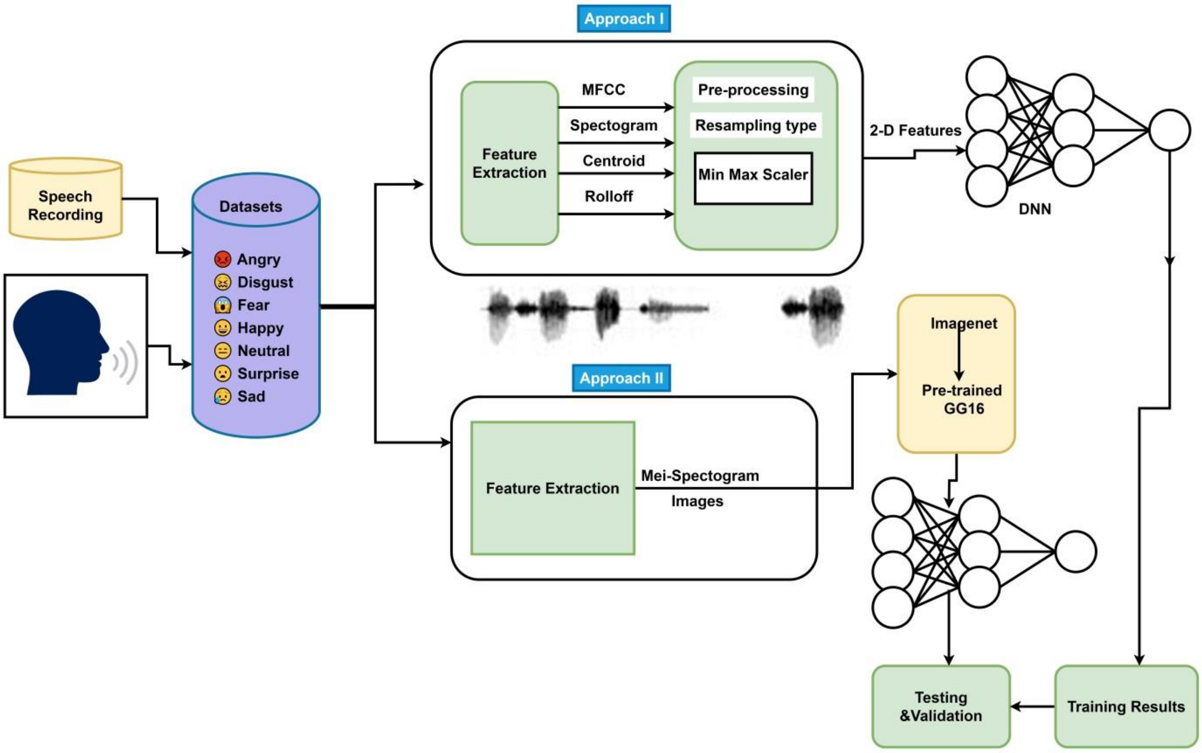


Figure 1. Detailed methodological flow chart of the speech signal analysis trend.

4. Results

The model trained for speech emotion recognition achieved a test accuracy of approximately 97.59%. This result suggests that the model demonstrated high performance in classifying emotions based on the audio data utilized for training. The training process encompassed multiple epochs, during which improvements in validation accuracy were observed. Early stopping was implemented to mitigate overfitting. Model layer details and parameter counts are presented in (Figure 2 and Table 1). The description of Key Features and specifications of the speech emotion recognition system is presented in (Table 2).

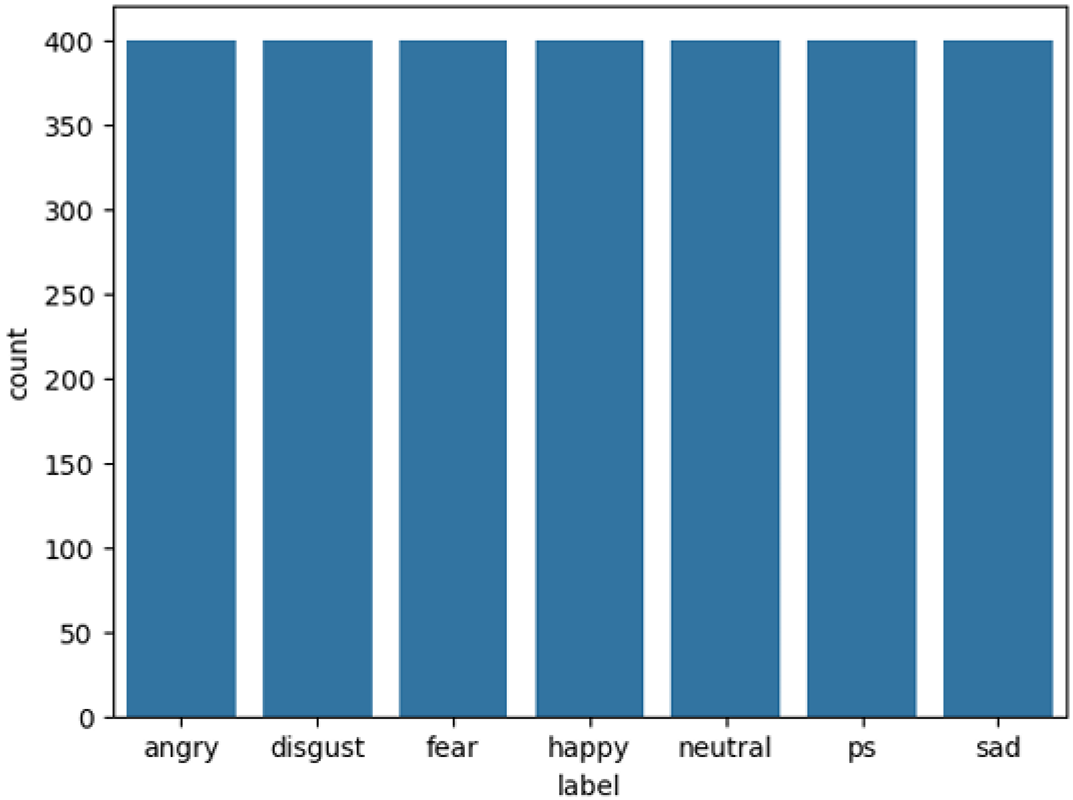
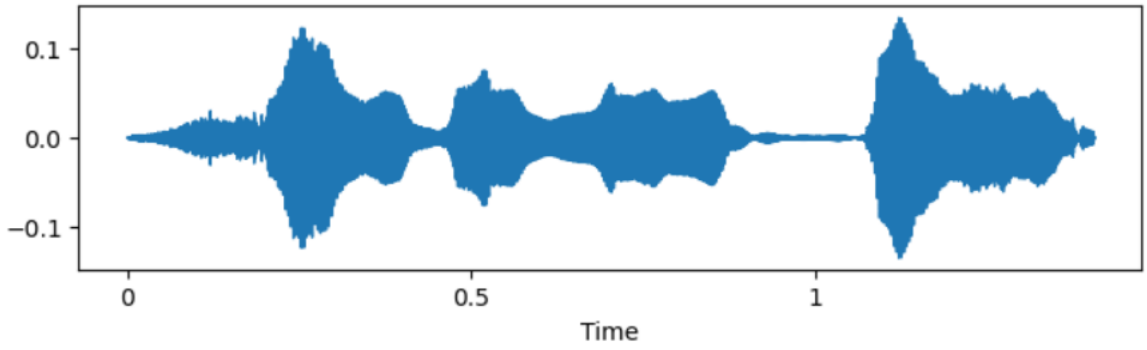
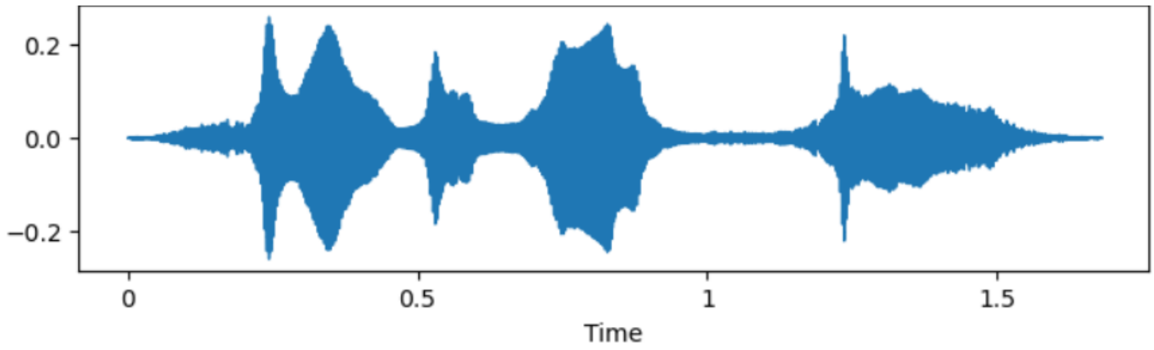


Figure 2. Distribution of emotion labels in dataset.

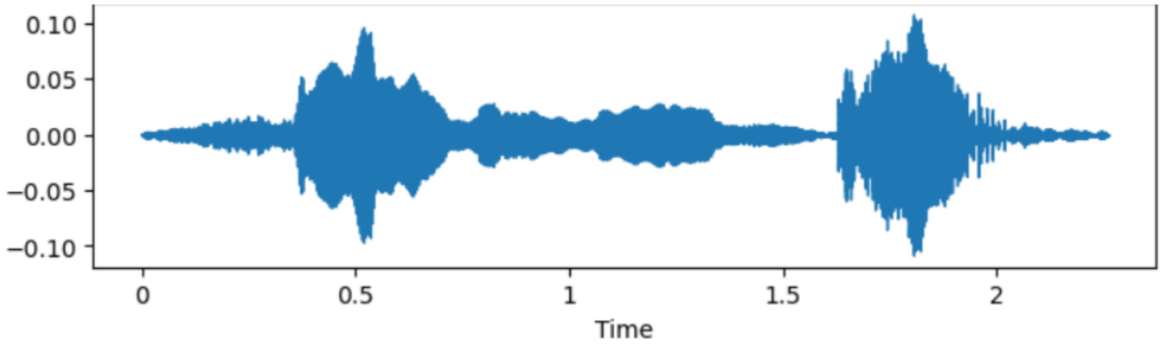
FEAR



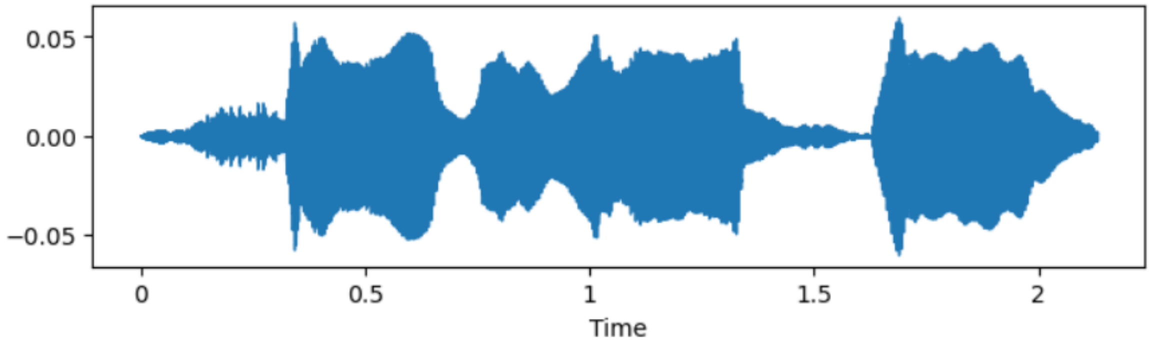
ANGRY



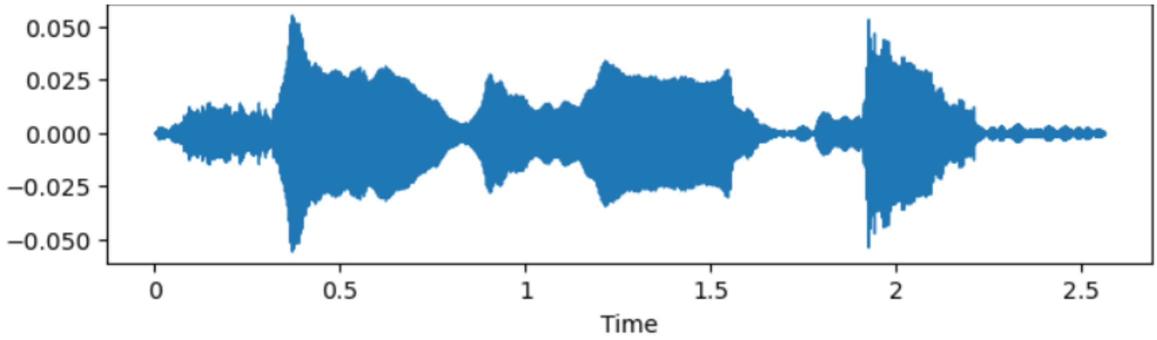
DISGUST



NEUTRAL



SAD



HAPPY

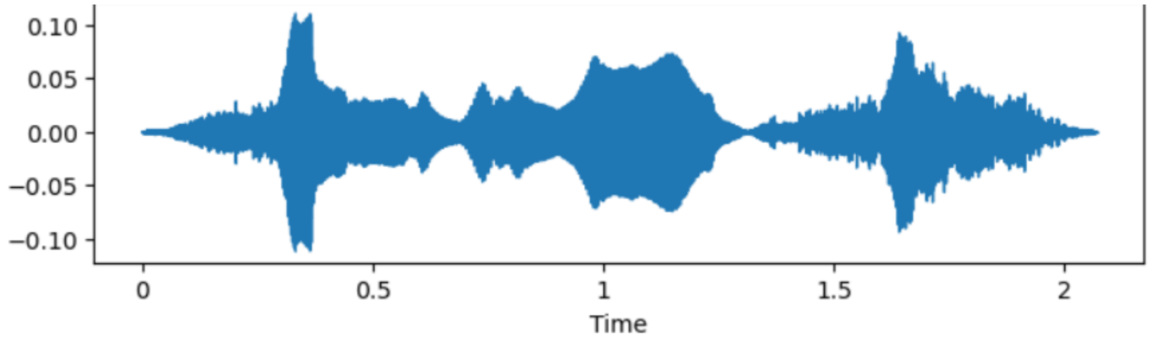


Table 1. Model layer details and Parameter counts.

Layer (Type)	Output shape	Parameter
LSTM (lstm_11)	(None, 256)	264,192
Dense (dense_48)	(None, 128)	32,896
Dense (dense_49)	(None, 64)	8,256
Dense (dense_50)	(None, 64)	4,160
Dropout (dropout_35)	(None, 64)	0
Dense (dense_51)	(None, 32)	2,080
Dropout (dropout_36)	(None, 32)	0
Dense (dense_52)	(None, 7)	231

Table 2. Key Features and specifications of the speech emotion recognition system.

Aspect	Detail description
Dataset Size	2800 audio files
Emotions Included	Angry, Disgust, Fear, Happy, Neutral, Sad, Surprise
Samples per Emotion	400 samples per emotion
Features Extracted	Mel-Frequency Cepstral Coefficients (MFCCs), 40 features
Data Augmentation	Time stretching, pitch shifting, noise injection, time shifting
Model Architecture	LSTM with Dense layers, Dropout for regularization
Model Layers	LSTM (256 units), Dense (128, 64, 32 units), Output Dense (7 units)
Activation Functions	ReLU (hidden layers), Softmax (output layer)
Optimizer	Adam
Loss Function	Categorical Crossentropy
Training Epochs	50 (with early stopping based on validation accuracy)
Test Accuracy	97.59%

5. Conclusions and Future Work

The speech emotion recognition (SER) system utilizes long short-term memory (LSTM) networks to analyze audio signals and classify emotions such as happiness, sadness, and anger. The methodology incorporated Mel-Frequency Cepstral Coefficients (MFCC) for feature extraction and data augmentation techniques to enhance the model's robustness. The system achieved a significant test accuracy of 97.59%, demonstrating its efficacy in emotion identification. These findings underscore the effectiveness of LSTM architectures in capturing temporal patterns within audio data, establishing their suitability for emotion classification tasks. The study contributes to advancing human-computer interaction, with potential applications encompassing mental health monitoring, virtual assistants, and multimedia analysis. In future iterations, the system's reliability in real-world scenarios can be enhanced through the utilization of larger and more diverse datasets, including those incorporating background noise. Further performance improvements may be achieved by implementing advanced models such as Gated Recurrent Units (GRU) or Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) architectures. The range of detectable emotions could be expanded to encompass more nuanced affective states, and the integration of audio with additional modalities, such as facial expressions, could potentially increase the system's accuracy. Future research efforts should also prioritize real-time processing capabilities and cross-cultural performance, to minimize bias and maximize global applicability across diverse languages and cultural contexts.

References

Akçay, M.B., Oğuz, K., (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers Speech Communication, 116, 56−76.

Albahri, A., Lech, M., Cheng, E., (2016). Effect of speech compression on the automatic recognition of emotions. Int. J. Signal Process. Syst. 4, 55–61. doi: 10.12720/ijsps.4.1.55-61.

Bui, H.M., Lech, M., Cheng, E., Neville, K., Burnett, I., (2017). Object recognition using deep convolutional features transformed by a recursive network structure. IEEE Access 4, 10059–10066. doi: 10.1109/ACCESS.2016.2639543.

Eyben, F., (2016). Real-time Speech and Music Classification by Large Audio Feature Space Extraction (Springer, 2016).

Liu, Z.T., Xiao, P., Li, D.Y., Hao, M., (2019). Speaker-independent speech emotion recognition based on CNN-BLSTM and multiple SVMs International conference on intelligent robotics and applications, Springer, 481−491.

Mekruksavanich, S., Jitpattanakul, A., Hnoohom, N., (2020). Negative emotion recognition using deep learning for Thai language 2020 Joint international conference on digital arts, media and technology with ECTI northern section conference on electrical, electronics, computer and telecommunications engineering (ECTI DAMT & NCON), IEEE, 71−74.

Mirsamadi, S., Barsoum, E., Zhang, C., (2017). Automatic speech emotion recognition using recurrent neural networks with local attention, In IEEE ICASSP.

Mustafa, M.B., Yusoof, M.A., Don, Z.M., Malekzadeh, M., (2018). Speech emotion recognition research: An analysis of research focus International Journal of Speech Technology, 21, 137−156.

Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., Anbarjafari, G., (2017). Audio-visual emotion recognition in video clips. IEEE Transactions on Affective Computing.

Scherer, K.R., (2003). Vocal communication of emotion: a review of research paradigms. Speech Commun.40,227–256.doi:10.1016/S0167-6393(02)00084-5.

Schröder, M., (2001). Emotional speech synthesis: a review, in Seventh European Conference on Speech Communication and Technology (Aalborg), 1–4.

Schuller, B., Rigoll, G., Lang, M., (2004). Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture, In IEEE ICASSP.

Zamil, Adib Ashfaq A., et al. (2019). Emotion Detection from Speech Signals using Voting Mechanism on Classified Frames." 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST). IEEE, 2019.