**SPEECH EMOTION RECOGNITION USING DEEP LEARNING**

A Minor Project Report Submitted

in partial fulfilment of the requirements for III year II semester

**Bachelor of Technology**

**In**

**Computer Science and Engineering**

By

**Prathap Tejaswi (2103A52029)**

**Bolla Harini Sri (2103A52006)**

**Bollena Anvitha (2103A52007)**

**Duggempudi Bala Varshitha (2103A52012)**

**Manthurthi Vyshnavi (2103A52023)**

Under the Guidance of

**Dr. P. Chandra Shaker Reddy**

**Professor Dept. of CS&AI**

****

**Department of Computer Science and Artificial Intelligence**

**Warangal Urban-506371, Telangana, India.**



**Department of Computer Science and Artificial Intelligence**

**CERTIFICATE**

This is to certify that project entitled **“Speech emotion cognition using deep learning”** is the bonafide work carried out by **P. Tejaswi, B. Harinisri, B. Anvitha, D. Bala Varshitha, M.Vyshnavi** as a Minor Project for the partial fulfillment to award the degree **BACHELOR OF TECHNOLOGY** in **COMPUTER SCIENCE &** **ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING** during the academic year 2023-2024 under our guidance and Supervision.

### **Dr. P. Chandrashaker Reddy** **Dr. M. Sheshikala**

Asst. Professor Prof. & HOD

S R University S R University,

Ananthasagar, Warangal. Ananthasagar, Warangal.

**DECLARATION**

We declare that the minor project work entitled “**SPEECH EMOTION RECOGNITION USING DEEP LEARNING**” recorded in this project work does not form part of any other project work. We further declare that the minor project report is based on our work carried-out at “**SR University, Anantasagar Mandal, Hanamkonda District – 506371**” in the Third year of our B-Tech course.

**Roll. No:** 2103A52029 **Name:** Prathap Tejaswi

**Roll. No:** 2103A52006 **Name:** Bolla Harini sri

**Roll. No:** 2103A52007 **Name:** Bollena Anvitha

**Roll. No:** 2103A52012 **Name:** Duggempudi Bala Varshitha

**Roll. No:** 2103A52023 **Name:** Manthurthi Vyshnavi

**Date:** **21-04-2024**  **Place:** **SR University**

# ACKNOWLEDGEMENT

We express our thanks to Project Guide **Dr. P. Chandrashaker Reddy, Professor** for guiding us from the beginning through the end of the Minor Project. We express our gratitude to Head of the department CS&AI, **Dr. M.Sheshikala, Professor** for encouragement, support and insightful suggestions. We truly value their consistent feedback on our progress, which was always constructive and encouraging and ultimately drove us to the right direction.

We wish to take this opportunity to express our sincere gratitude and deep sense of respect to our beloved Dean, School of Computer Science and Artificial Intelligence, **Dr C. V. Guru Rao**, for his continuous support and guidance to complete this project in the institute.

Finally, we express our thanks to all the teaching and non-teaching staff of the department for their suggestions and timely support.

**ABSTRACT**

A distinctive and well-structured dataset consisting of 2800 WAV audio recordings with two actresses, 26 and 64 years old, is presented in this study. The dataset is put up to look at how 200 target phrases portray seven different emotions: anger, contempt, fear, happiness, pleasant surprise, sadness, and neutrality. The audio recordings offer a wealth of resources for Speech Emotion Recognition (SER) model training and evaluation. They are contained within the corresponding emotion and actress folders. This dataset can be used by practitioners and researchers in emotional computing, emotion analysis, and artificial intelligence to build and evaluate reliable SER systems. The WAV format is used to guarantee high-quality audio data, which makes it possible to precisely analyse minute acoustic characteristics associated with emotional expression. The structure of the dataset makes it easier to conduct focused studies on particular feelings or age-related trends, which increases its adaptability to a range of research goals. This dataset, which provides a well-organized and varied collection of emotionally annotated audio recordings, is a significant asset for furthering research in speech emotion analysis. Its possible uses range from enhancing our comprehension of the dynamics of emotional speech to creating artificial intelligence systems that are more sophisticated and context-sensitive.

# ABOUT THE ORGANIZATION

The 45-year-old Sri Rajeshwara Educational Society, the parent organization of SR University, is a conglomerate of educational institutions with 10,000 staff members who are not teachers and over 90,000 students. 95 educational institutions in Telangana and Andhra Pradesh are under the management of SR Educational Academy. The mission of SR University is to establish a cutting-edge learning environment that produces graduates who will have a major impact on the development of Telangana and India. We intend to use three crucial differentiators to completely overhaul the educational system. Everyone has the opportunity to participate, flourish, and leave a lasting impression through the co-curricular, extracurricular, and curricular options offered by the collaborative entrepreneurial environment. The system has close connections to both global academic institutions and business.

# Table of Contents

## Chapter No. Title Page No.

1. Introduction
   1. Overview 2
   2. Problem Statement 5
   3. Existing System 5
   4. Proposed system 6
   5. Objectives 7
   6. Architecture 8
2. Literature survey 10
3. Requirement Analysis 14
4. Data set 15
5. Implementation 16
6. Results and discussion 17
7. Challenges 20
8. Conclusion and future scope 22
9. References 23

**INTRODUCTION**

Human voices are much more than just words; via minute changes in rhythm, tone, and pitch, they reveal secrets about our inner selves. By automatically recognizing the emotions expressed in spoken language, speech emotion recognition (SER) seeks to decipher these hidden messages. This area of study has the potential to completely transform human-computer interaction by allowing computers to recognize and react to human emotions. Deep learning has become a potent technique for improving SER because of its capacity to extract intricate patterns from massive volumes of data. Deep learning algorithms surpass conventional techniques by capturing the nuances of human emotion through the analysis of acoustic properties collected from speech data. Imagine having a virtual assistant that adjusts its voice to suit your mood, providing a soothing voice when you're stressed or an energizing one when you're happy. Imagine learning environments that adjust to each student's level of emotional involvement. Imagine customer support bots who are able to understand your complaints and provide appropriate answers. These are but a few examples of the profoundly positive effects that deep learning-driven speech emotion recognition can have. Nevertheless, there are obstacles in the way of fully comprehending emotions through words. Challenges related to data collection, annotation, and cultural differences must be addressed. As this technology advances, ethical issues pertaining to bias and privacy must also be carefully considered.

The potential of SER based on deep learning is still enormous despite these obstacles. We should anticipate the emergence of progressively more complex and subtle systems as research advances and ethical issues are resolved. This will pave the way for a time when technology is able to comprehend not just our spoken words but also the feelings they evoke. The level of accuracy and adaptability of traditional approaches were limited by their dependence on hand-crafted components. Presenting deep learning, an effective instrument that deduces intricate patterns straight from data. Models for deep learning such as Convolutional Neural Networks Even though SER has a promising future, difficulties still exist. This technology has the ability to usher in a new era of emotional communication and understanding when developed responsibly and ethically.

* 1. Overview:

Speech Emotion Recognition (SER) is a field within the broader domain of affective computing that aims to detect and classify emotions conveyed through speech signals. Here's an overview of the SER process:

1. Introduction to SER:

- Explain the significance of understanding emotions in speech for various applications such as human-computer interaction, mental health monitoring, and sentiment analysis.

- Discuss the challenges associated with interpreting emotional cues in speech due to variations in linguistic, cultural, and individual factors.

2. Data Collection:

- Gather a diverse dataset comprising audio recordings of speech samples annotated with corresponding emotion labels.

- Ensure the dataset includes variability in speakers' demographics (age, gender, accent), emotional intensity, and context (dialogues, monologues, scripted vs. spontaneous speech).

3. Preprocessing:

- Convert raw audio files into a format suitable for analysis (e.g., WAV, MP3).

- Extract relevant features from the speech signals, such as Mel-Frequency Cepstral Coefficients (MFCCs), prosodic features (pitch, energy), and spectral features.

- Normalize and standardize the feature vectors to account for variations in recording conditions and speaker characteristics.

4. Feature Representation:

- Represent the extracted features as high-dimensional vectors or matrices suitable for input into machine learning or deep learning models.

- Explore different feature representations and dimensionality reduction techniques to capture discriminative information while reducing computational complexity.

5. Model Selection:

- Choose appropriate models for SER, ranging from traditional machine learning algorithms (e.g., Support Vector Machines, Random Forests) to deep learning architectures (e.g., Convolutional Neural Networks, Recurrent Neural Networks).

- Experiment with different model architectures, regularization techniques, and hyperparameters to optimize performance.

6. Training:

- Split the dataset into training, validation, and test sets to train and evaluate the SER model.

- Train the selected model using the training data, employing optimization algorithms (e.g., gradient descent) to minimize a chosen loss function.

- Monitor the model's performance on the validation set to prevent overfitting and adjust model parameters accordingly.

7. Evaluation:

- Evaluate the trained SER model's performance on the test set using metrics such as accuracy, precision, recall, F1-score, and confusion matrix.

- Analyze the model's performance across different emotion categories and identify areas for improvement.

8. Deployment:

- Deploy the trained SER model into real-world applications, integrating it with relevant systems or platforms.

- Optimize the model for efficiency, scalability, and real-time processing, considering factors like computational resources and latency constraints.

- Implement mechanisms for monitoring the SER system's performance in production and collecting user feedback for iterative improvements.

9. Challenges and Future Directions:

- Discuss ongoing challenges in SER, such as handling noisy data, addressing cultural and linguistic variations, and improving cross-domain generalization.

- Explore potential avenues for advancing SER technology, including multi-modal emotion recognition combining speech with other modalities (e.g., facial expressions, text), and personalized emotion detection tailored to individual users.

10. Conclusion:

- Summarize the achievements and contributions of the SER project, highlighting its implications for various domains and its potential for further research and development.

1.2 Problem Statement:

Speech emotion recognition (SER) deciphers emotional cues in spoken language, revolutionizing human-computer interaction. Deep learning, adept at extracting intricate patterns, empowers SER by capturing nuances in tone and pitch from raw audio data. This advancement transcends traditional methods, offering diverse applications in healthcare, education, customer service, and entertainment.

1.3 Existing System:

The speech emotion detection system is implemented as a Machine Learning (ML) model. The steps of implementation are comparable to any other ML project, with additional fine-tuning procedures to make the model function better. The flowchart represents a pictorial overview of the process (see Figure 1). The first step is data collection, which is of prime importance. The model being developed will learn from the data provided to it and all the decisions and results that a developed model will produce is guided by the data. The second step, called feature engineering, is a collection of several machine learning tasks that are executed over the collected data. These procedures addressthe several data representation and data quality issues. The third step is often considered the core of an ML project where an algorithmic based model is developed. This model uses an ML algorithm to learn about the data and train itself to respond to any new data it is exposed to. The final step is to evaluate the functioning of the built model. Very often, developers repeat the steps of developing a model and evaluating it to compare the performance of different algorithms. Comparison results help to choose the appropriate ML algorithm most relevant to the problem.

1.4 Proposed System:

**Data Preprocessing:**

Prepare your speech dataset by extracting features like Mel-frequency cepstral coefficients (MFCCs), pitch, energy, and other relevant features.

Divide the dataset into training, validation, and test sets.

**Feature Representation:**

Convert the extracted features into a suitable format for input into the LSTM model. This could involve reshaping the data into sequences or using techniques like time-frequency representations.

**Model Architecture:**

Design an LSTM-based architecture for emotion recognition. A common approach is to stack multiple LSTM layers followed by fully connected layers.

Experiment with different LSTM configurations, such as varying the number of LSTM layers, hidden units, dropout rates, and activation functions.

**Model Training:**

Train the LSTM model using the prepared training data. Use appropriate optimization algorithms like Adam or RMSprop.

Monitor the model's performance on the validation set and adjust hyperparameters accordingly to prevent overfitting.

Utilize techniques like early stopping to prevent overfitting and save the best-performing model.

**Model Evaluation:**

Evaluate the trained LSTM model on the test set using appropriate evaluation metrics such as accuracy, F1-score, or confusion matrix.

Analyze the model's performance across different emotion classes to identify any biases or areas for improvement.

**Fine-tuning and Optimization:**Fine-tune the LSTM model by experimenting with different hyperparameters and architectures to further improve performance.

Consider techniques like transfer learning or domain adaptation if applicable, especially if you have access to pre-trained models on similar tasks or domains.

**Deployment and Integration:**

Once satisfied with the model's performance, deploy it into your desired application or system.

Develop integration methods to incorporate the LSTM model into real-world scenarios, such as through APIs or embedded systems.

Ensure that the deployed model meets any performance and computational constraints, especially if deploying on resource-constrained devices like smartphones or IoT devices.

**Continuous Monitoring and Improvement:**

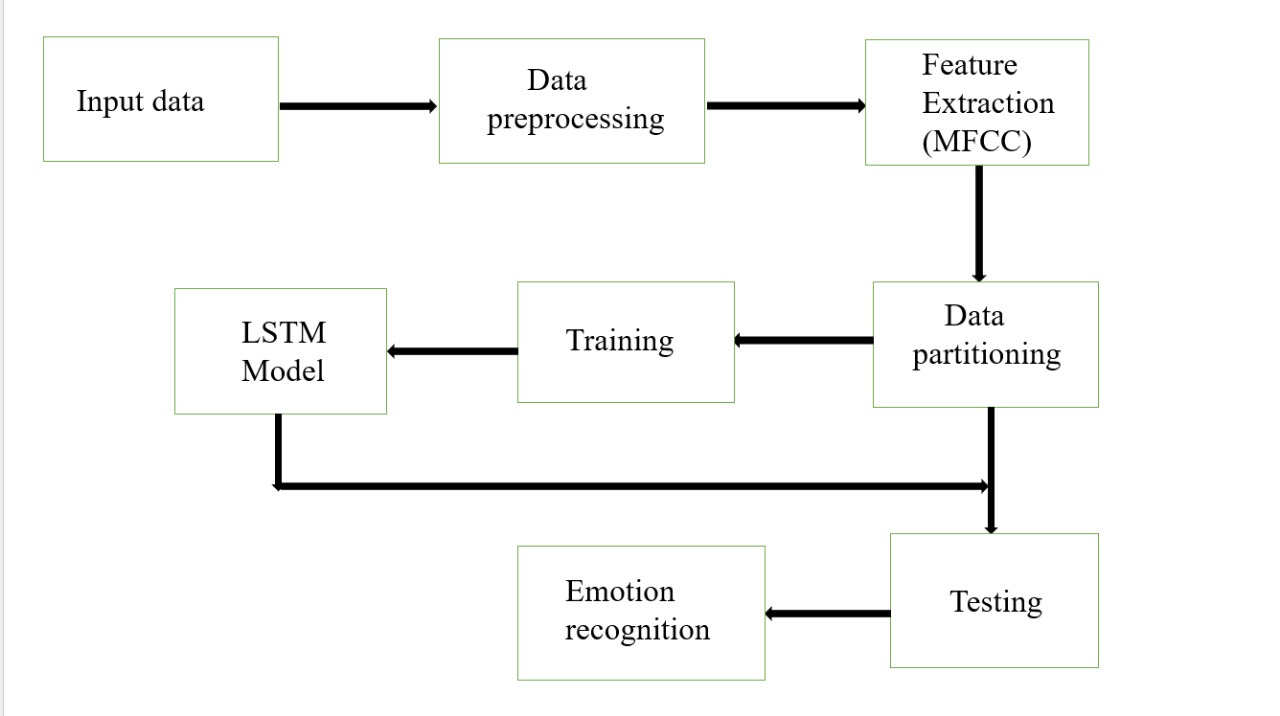
Continuously monitor the performance of the deployed model in real-world scenarios and gather feedback from users.

Use this feedback to iteratively improve the model's accuracy, robustness, and user experience over time.

1.5 OBJECTIVES:

* Several goals are crucial in a speech emotion recognition that uses LSTM models and deep learning. First and foremost, the system has to reliably categorize spoken words emotional content into pre-established groups like happiness, sorrow, anger, etc. Second, to guarantee its usefulness in real-world situations, it should demonstrate robustness when tested with different speakers, languages, and recording settings.
* For training, LSTM models need a lot of processing power, which might be problematic for real-time applications. The model should also aim for efficiency in terms of compute resources and inference time, so that it may be used in situations with limited resources, such as embedded systems or mobile devices.
* LSTMs, while effective at capturing temporal dependencies in speech, can struggle with complex, nonlinear emotional nuances that are embedded in the prosodic features like tone, pitch, and rhythm. This results in challenges with generalizing across different speakers, accents, and languages due to variations in speech patterns and also scarcity and imbalance of labeled emotion datasets.

1.6 Architecture:



**1.Input data:**

The input speech refers to the audio data received by the system for processing. This could be any spoken words, phrases, or sentences captured by a microphone or any audio input device.

**2.Data Preprocessing:**

Data preprocessing involves cleaning, transforming, and organizing raw data to prepare it for analysis or modeling. This includes handling missing values, scaling numerical features, encoding categorical variables, and detecting outliers.

**3.Feature Extraction:**

Feature extraction involves identifying and extracting relevant attributes or characteristics from the input speech data. Common features include spectral features like MFCCs (Mel-Frequency Cepstral Coefficients), and other acoustic properties that are essential for speech recognition and analysis.

**4. Data partitioning:**

Data partitioning involves dividing a dataset into subsets for training, validation, and testing purposes. Typically, the majority of the data is used for training the model, a smaller portion for validation to tune hyperparameters, and a separate portion for testing to evaluate model performance on unseen data. Proper data partitioning helps assess model generalization and prevents overfitting.

**5. Training:**

Training the data involves feeding it into a machine learning algorithm to learn patterns and relationships within the dataset. During training, the algorithm adjusts its internal parameters iteratively to minimize a defined loss function, optimizing its ability to make accurate predictions.

**6.Testing:**

Testing the data involves evaluating the performance of a trained machine learning model on unseen data to assess its generalization ability. Testing data is typically kept separate from the training data and is used to measure various performance metrics such as accuracy, precision, recall, and F1-score.

**7.LSTM Model:**

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture capable of learning long-term dependencies in sequential data. It includes specialized memory cells that can retain information over long sequences, making it effective for tasks such as time series prediction, natural language processing, and speech recognition.

**8.Emotion recognition:**

Emotion recognition involves identifying the emotional state conveyed through the input speech. This could include emotions such as happiness, sadness, anger, or neutral sentiments. Emotion recognition algorithms analyze various acoustic features, intonation patterns, and linguistic cues to infer the underlying emotional state accurately.

**LITERATURE SURVEY**

Komal Bharti, Prachi Shrivastava, (2023) [1]- In this paper, the focus is on Audio-Visual Automatic Speaker Recognition (AV-ASR), an emerging field aimed at enhancing person authentication by integrating both audio and visual modalities. Recognizing the susceptibility of solely relying on audio or video features to interference, Through experimental classification approaches, the proposed AV-ASR model achieves an impressive recognition accuracy of 80.0%, surpassing results obtained by various classifiers. This highlights the effectiveness of integrating hybrid visual features in AV-ASR systems for improved accuracy in person authentication tasks.

Alluhaidan, A.S. (2023) [2]-This paper focuses on enhancing Speech Emotion Recognition (SER) systems by combining Mel frequency cepstral coefficients (MFCCs) with time-domain features (MFCCT) and leveraging convolutional neural networks (CNNs). Traditional SER approaches often rely solely on MFCCs, but this study proposes that incorporating time-domain features alongside MFCCs can improve emotion representation accuracy. By feeding the hybrid MFCCT features into a CNN model, the researchers achieved remarkable results, outperforming both MFCCs and time-domain features on multiple datasets, including Emo-DB, SAVEE, and RAVDESS, with accuracies of 97%, 93%, and 92%respectively.

Leila Kerkeni, Youssef Serrestou (2019) [3]-This chapter presents a comprehensive investigation into Speech Emotion Recognition (SER) systems, focusing on theoretical definitions, affective state categorization, and emotion expression modalities. The study develops an SER system utilizing various classifiers and feature extraction methods, particularly Mel-frequency cepstrum coefficients (MFCC) and modulation spectral (MS) features. Feature selection techniques are employed to identify the most relevant feature subset, and several machine learning paradigms, including recurrent neural networks (RNN), multivariate linear regression (MLR), and support vector machines (SVM), are utilized for emotion classification.

Ramesh, S., Gomathi (2023) [4] -The paper presents over the past decade, automatic speech emotion detection has presented a significant challenge in the realm of human-computer interaction. This study addresses this challenge by focusing on detecting seven primary emotions – neutrality, happiness, sadness, fear, surprise, disgust, and anger – through speech signals. Unlike previous approaches that utilized separate databases, this study merges the SAVEE and TESS datasets to create a comprehensive database for emotion identification. The main objective is to characterize emotions using this robust dataset.

Abolfazl Younesi, Mohsen Ansari. (2024) [5] - In this paper, In today's digital landscape, Convolutional Neural Networks (CNNs) stand at the forefront of Deep Learning (DL), serving as invaluable tools for a wide array of computer vision tasks. This survey paper delves into the diverse spectrum of CNN architectures, ranging from 1D to 3D CNNs, as well as specialized variants like dilated, grouped, and attention-based convolutions. By providing a comparative analysis of these CNN types, the paper elucidates their unique structures, strengths, weaknesses, and practical applications across different domains.

L. Yunxiang and Z. Kexin, (2023) [6] - In this study, this paper addresses the challenges of reduced efficiency in speech emotion recognition (SER) caused by noise interference and gender differences. To tackle this issue, the authors propose two multi-task learning models based on adversarial multi-task learning (ASP-MTL). The first model treats emotion recognition as the main task and noise recognition as the auxiliary task, removing identified noise to enhance recognition accuracy. Subsequently, the second model focuses on emotion recognition as the main task and gender classification as the auxiliary task. By leveraging shared information and identifying specific tasks, these models improve recognition performance.

Mehmet Berkehan Akçay, Kaya Oğuz, (2020) [7] - This paper underscores the significance of speech emotion recognition (SER) systems in extending human-like communication to computer applications. Despite its longstanding history spanning over two decades, SER has experienced a resurgence of interest due to recent advancements in computing and technology. The authors highlight the necessity for an updated understanding of SER methodologies and techniques, given the evolving landscape of research in this field. Through a detailed survey of current literature, the paper delineates distinct areas of SER, offering insights into ongoing research endeavors and listing current challenges. By synthesizing a comprehensive overview, the paper serves as a valuable resource for researchers and practitioners seeking to navigate the complexities of SER and contribute to its continued advancement.

Francesco Ardan Dal Rí, Fabio Cifariello Ciardi, (2023) [8] - This paper delves into the realm of Speech Emotion Recognition (SER) leveraging the advancements of deep learning, particularly focusing on Convolutional Neural Network (CNN) models enhanced with Convolutional Attention Blocks. Addressing the challenges within SER, such as the lack of standardized practices and high-quality datasets, the study implements and evaluates the proposed model across four prominent English datasets: RAVDESS, TESS, CREMA-D, and IEMOCAP. Through rigorous experimentation, the pipeline achieves notable accuracies across individual datasets, with means ranging from 63% to 100%.

Chi-Chun Lee, Theodora Chaspari (2023) [9]- This paper highlights the burgeoning interest in Speech Emotion Recognition (SER) technologies driven by the proliferation of Internet-of-Things (IoT) and smartphone devices. However, it underscores the myriad challenges hindering widespread adoption, spanning conceptual, technical, and societal domains. These challenges are particularly pronounced in "in-the-wild" scenarios due to the complex nature of human emotion, labeling difficulties, and contextual variability. Moreover, societal and ethical concerns regarding privacy, fairness, and explainability pose additional barriers to acceptance.

Changzeng Fu, Chaoran Liu, (2023) [10] - This paper addresses the challenge of speaker individual bias in Speech Emotion Recognition (SER), which can lead to irregular clusters of emotion-related features and overfitting of in-domain datasets. To mitigate this issue, the authors propose an adversarial training-based classifier that regularizes the distribution of latent representations, smoothing boundaries among different emotion categories. This regularization involves mapping representations into Gaussian distributions in an unsupervised manner, with a novel adoption of a mixture of isolated Gaussian distributions.

Zhen-Tao Liu, (2018) [11] - The paper proposes a new method for speech emotion recognition that combines Brain Emotional Learning (BEL) model with Genetic Algorithm (GA) to improve accuracy. BEL model is inspired by how human brains process emotions, but it has limitations. GA helps address these limitations by optimizing the weights in the BEL model. The researchers tested the method on different datasets and feature sets, achieving good results for speaker-dependent recognition (average accuracy above 70%). While speaker-independent recognition accuracy was lower (around 40%), it still demonstrates the method's potential for speech emotion recognition.

Oh-Wook Kwon, Kwokleung Chan, (2023) [12] - This research investigated features for speech emotion recognition. They analyzed pitch, energy, formants, mel-band energy, and MFCCs, along with their changes over time. Statistical features were extracted and analyzed using different classifiers like SVM and QDA. Results showed pitch and energy to be most important. They achieved high accuracy (over 96%) for classifying stressed vs neutral speech using a specific database and classifier setup. However, accuracy dropped for recognizing more emotions (4 or 5 classes) in different databases, suggesting the need for further improvement.

Yue Xie, Ruiyu Liang, Zhenlin Liang, (2019) [13] - This research tackles limitations in current speech emotion recognition methods by focusing on the flow of emotions over time within speech. They propose a new method that uses frame-by-frame analysis of speech features along with a special Long Short-Term Memory (LSTM) network with an attention mechanism. This allows the model to capture the subtle changes in emotions throughout the speech. To improve efficiency, they also modified the LSTM's forgetting gate. Experiments showed significant improvements in accuracy compared to existing methods, demonstrating the effectiveness of their approach in recognizing emotions from speech.

Dariusz Czerwinski, Pawel Powroznik, (2018) [14] - This research explores using artificial neural networks (ANNs) to recognize emotions in spoken Polish. Existing methods often rely on techniques like k-nearest neighbors or decision trees. Here, the researchers use spectrograms, visual representations of speech, as input for the ANN. This approach achieved good classification accuracy, exceeding 75% for some emotions. They also analyzed how different parts of the spectrogram contribute to the recognition process, providing valuable insights for future improvements.

Cheng Lu, Yuan Zong, (2022) [15] - This research introduces a new method, Domain Invariant Feature Learning (DIFL), for recognizing emotions in speech regardless of the speaker (speaker-independent). Traditional methods struggle with this because different speakers have different voice characteristics. DIFL tackles this by learning features that focus on the emotions themselves, rather than the speaker's voice. It achieves this through a two-step process: first, it aligns features from different speakers to minimize discrepancies, and then it uses multiple "discriminators" to confuse the model into ignoring speaker information. Experiments on various datasets show that DIFL outperforms existing methods in recognizing emotions from speech.

**REQUIREMENT ANALYSIS**

**1.** **Project Scope**: Define the scope of the project, including the target audience, the types of emotions to be recognized, and the platform on which the system will operate (e.g., desktop, mobile).

**2. Data Collection and Preprocessing**: Specify the sources from which speech data will be collected, such as databases, recordings, or real-time capture. Outline preprocessing steps like noise reduction, feature extraction, and data augmentation.

**3.** **Feature Selection**: Determine which features will be extracted from the speech signals for emotion recognition, such as MFCCs (Mel-Frequency Cepstral Coefficients), pitch, energy, and spectral features.

**4.** **Machine Learning Models**: Decide on the machine learning models to be used for classification, such as Support Vector Machines (SVM), Random Forests, or deep learning architectures like Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs).

**5.** **Training and Testing** : Define the methodology for splitting the dataset into training, validation, and testing sets. Specify evaluation metrics, such as accuracy, F1-score, or confusion matrix, to assess model performance.

**6.** **Model Deployment** : Determine how the trained model will be deployed, whether as a standalone application, a web service, or integrated into existing software systems.

**7.** **Performance Optimization** : Identify strategies for optimizing model performance, such as hyperparameter tuning, model ensemble methods, or model compression techniques.

**8.** **Ethical Considerations** : Consider ethical implications related to data privacy, bias mitigation, and potential misuse of the system, and implement measures to address these concerns.

**9.** **Documentation and Maintenance** : Document the project thoroughly, including data sources, preprocessing steps, model architectures, and deployment instructions. Establish a plan for ongoing maintenance and updates to ensure the system remains effective over time.

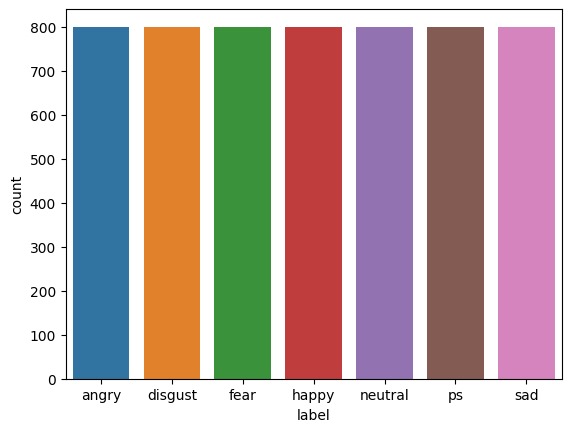
**DATASET**

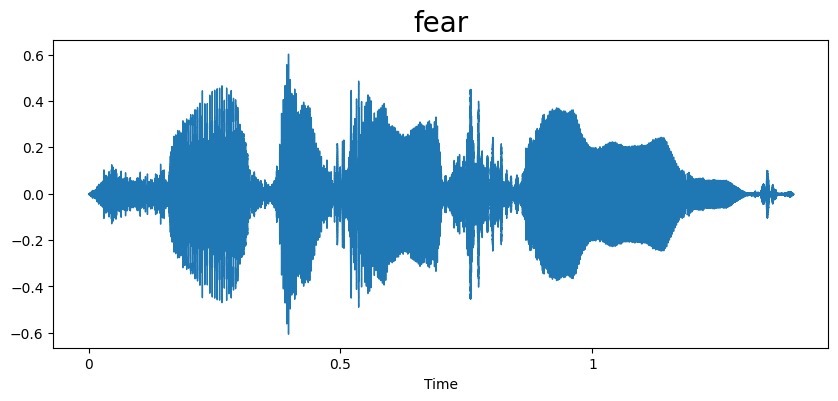
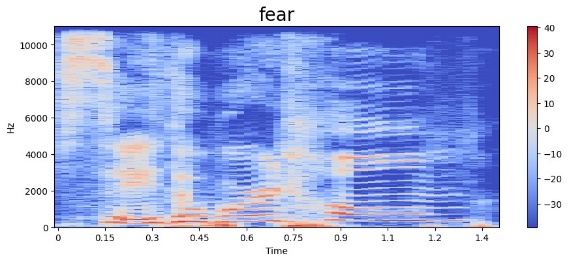
This dataset contains audio files of two actresses, ages 26 and 64, uttered a set of 200 target words in the carrier phrase "Say the word \_." The set was recorded depicting each of the seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). In total, there are 2800 data points (audio files). The two female actors and their feelings are contained within separate folders in the dataset due to its organizational structure. And all 200 target words audio file can be found within that. The audio file is saved in WAV format. It is made up of several audio recordings of performers representing various emotional states. The Toronto University generated the dataset, which is frequently used to examine human emotion in speech signals. The TESS dataset includes recordings of seven different emotional states: neutral, pleasant surprise, anger, contempt, fear, and happiness. Every recording has the appropriate emotional category listed on it. Two female and two male actors recorded 2,743 samples for the dataset. Every performer plays every emotional category several times, for a total of 280 samples per performer. The goal is for the expressions to accurately and naturally convey the relevant emotions. The audio recordings found in the TESS dataset were made under consistent recording settings in a controlled setting with little outside noise. This guarantees that the dataset's main focus is the emotional expressions and makes correct annotation and analysis easier. The WAV format, 16 kHz sampling rate, and 16-bit resolution of the audio files in the TESS dataset are provided. With a period of roughly 3 to 5 seconds, each audio file captures the salient elements of the emotional expression. Research projects on emotional computing, speech emotion recognition, and related topics have made substantial use of the TESS dataset. The dataset is used by researchers to create and assess deep learning architectures, machine learning models, and signal processing methods for precisely identifying and interpreting the emotional states expressed in voice signals. All things considered, the TESS dataset is essential to improving our knowledge of how emotions are expressed in speech and creating computer models for autonomous emotion detection. For researchers in the discipline, its accessibility and consistent annotation make it an invaluable resource.

**IMPLEMENTATION**

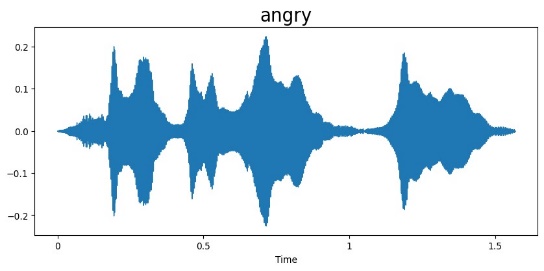
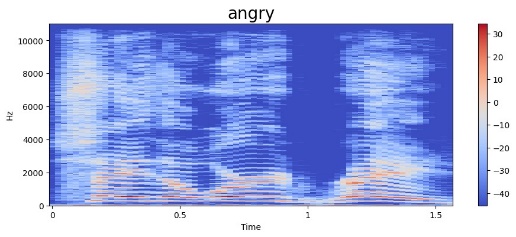
Deep learning-based Speech Emotion Recognition (SER) implementation requires a number of important phases, from data preprocessing to model training and assessment. First, the audio data must be loaded and prepared, usually by normalizing and extracting features. Mel-frequency cepstral coefficients (MFCCs), which represent the speech signal's spectral properties, and pitch or intensity features are examples of common features. Following that, these attributes are fed into a deep learning model, which is frequently built using architectures such as Recurrent Neural Networks (RNNs), or more precisely, Long Short-Term Memory (LSTM) networks, which are renowned for their efficacious modelling of sequential data. The architecture, which includes the number of layers, hidden units, and activation functions, is defined before the model is built. Typically, the architecture of LSTM-based models consists of LSTM layers, which are followed by fully linked layers for classification. The model is assembled with the proper optimizers and loss functions during training. It is then trained using labelled data, iteratively modifying its parameters to minimize the loss function. To evaluate the performance of the model, training entails dividing the dataset into training, validation, and testing sets. To maximize training and avoid overfitting, strategies like early halting and mini-batch gradient descent are commonly used. The model is tested on the test set after training to measure how well it performs at identifying emotions from unobserved data. Long Short-Term Memory (LSTM) models have become highly effective tools in the field of deep learning-based Speech Emotion Recognition (SER) for capturing subtleties and temporal dependencies in speech signals. An LSTM model's implementation in SER usually entails a few crucial steps. In order to represent the speech signals, the audio data is first pre-processed by extracting essential characteristics such spectrograms or Mel-frequency cepstral coefficients (MFCCs). The LSTM network, which is made up of several LSTM layers intended to identify sequential patterns in the data, is then fed these characteristics. By altering its parameters through gradient descent and backpropagation, the LSTM model gains the ability to identify emotional cues embedded in the speech signals during training. Lastly, the trained model's ability to correctly classify emotions is tested using a different test dataset. By capturing long-range relationships and temporal dynamics in voice data, the LSTM model offers a promising way to push the boundaries of speech recognition technology.

**RESULT AND DISCUSSIONS**

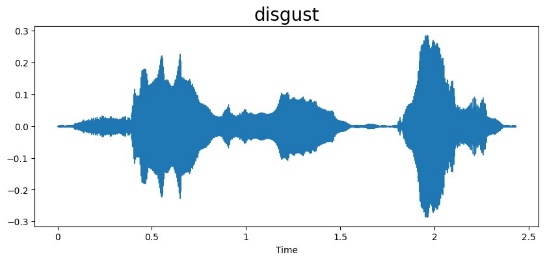
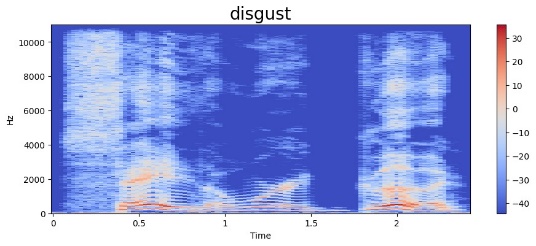


**** 

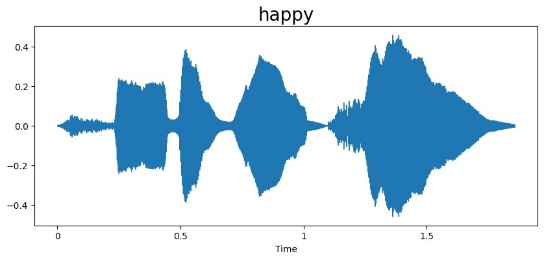
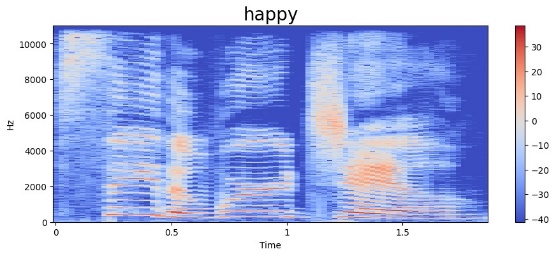
**FEAR**

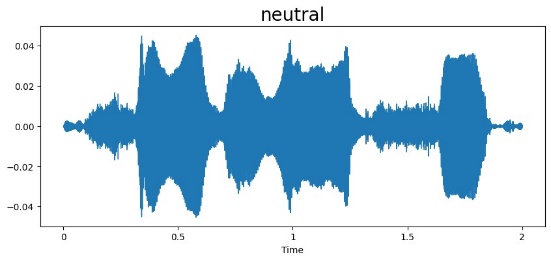
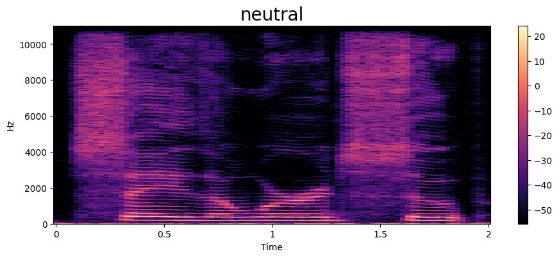
**ANGRY**

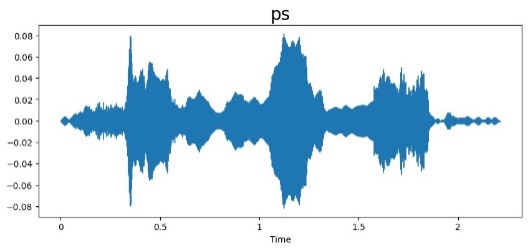
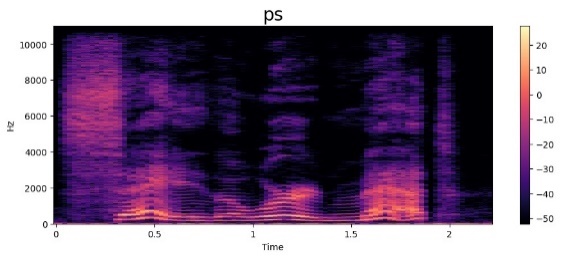
**DISGUST**

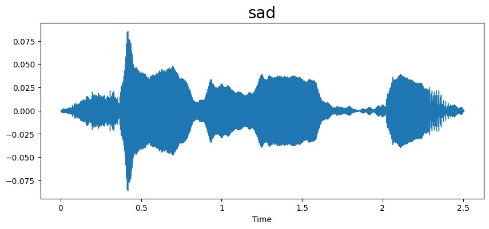
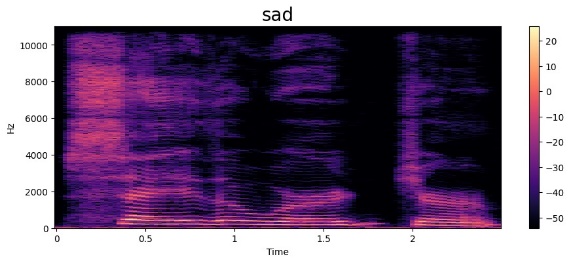
**HAPPY**

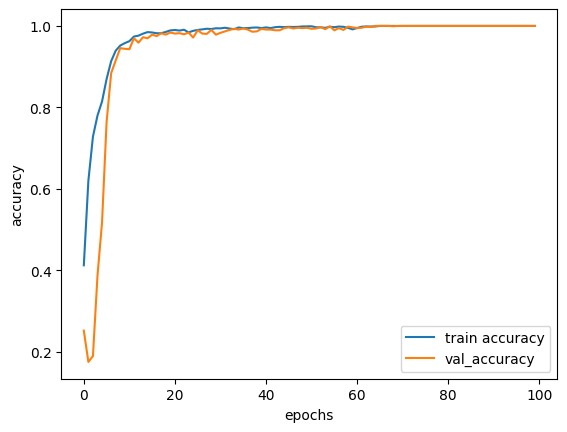
**NEUTRAL**

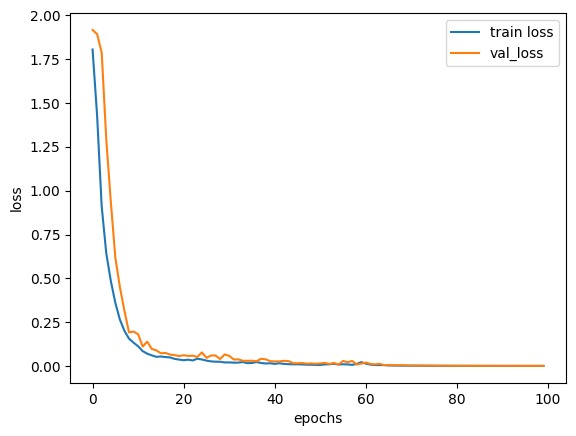
**PLEASANT SURPRISE**

**SAD**



**ACCURACY GRAPH**



**LOSS GRAPH**

**CHALLENGES**

**1. Variability in Emotion Expression**: Emotions can be expressed differently based on cultural, linguistic, and individual factors, making it challenging to develop models that generalize across diverse populations.

**2. Data Availability and Quality**: Acquiring large, diverse, and accurately labeled datasets for training emotion recognition models can be difficult, leading to potential biases and limited model performance.

**3. Noise and Environmental Factors**: Background noise, varying recording conditions, and acoustic interference can degrade the quality of speech signals, affecting the performance of emotion recognition algorithms.

**4. Feature Extraction and Selection**: Selecting informative and discriminative features from speech signals while minimizing dimensionality and computational complexity is a non-trivial task.

**5. Model Complexity and Generalization**: Deep learning models for speech emotion recognition often require large amounts of data and computational resources for training, and there's a risk of overfitting to the training dataset, leading to poor generalization on unseen data.

**6. Inter-class Confusion**: Some emotions may share similar acoustic characteristics, leading to confusion between classes during classification (e.g., distinguishing between happiness and excitement).

**7. Real-time Processing**: Achieving real-time performance for speech emotion recognition applications, especially on resource-constrained devices like smartphones, poses additional challenges in terms of computational efficiency and latency.

**8. Subjectivity and Ground Truth Annotation**: Emotions are subjective experiences, and labeling speech data with ground truth emotion labels can be inherently subjective and prone to annotation errors or inconsistencies.

**9. Ethical Considerations**: Ensuring fairness, transparency, and privacy protection in the collection, processing, and use of speech data is essential to mitigate potential ethical risks and concerns.

**10. Integration with User Interfaces**: Integrating emotion recognition capabilities into user interfaces while maintaining usability and user experience requires careful design and consideration of user feedback and interaction patterns.

**CONCLUSION AND FUTURE SCOPE**

In conclusion, speech emotion recognition represents a significant advancement in human-computer interaction and emotional intelligence applications. Through the extraction of acoustic features and the utilization of sophisticated machine learning or deep learning models, we can effectively classify and understand the emotional content conveyed in speech signals. As technology continues to evolve, the integration of speech emotion recognition systems into various domains, including virtual assistants, healthcare, and entertainment, holds immense potential to enhance user experiences and enable more empathetic interactions. However, ongoing research and development are necessary to improve the accuracy, robustness, and real-time capabilities of these systems, ultimately bringing us closer to a future where machines can comprehend and respond to human emotions with nuance and sensitivity.

The future scope for speech emotion recognition, considering its current limitations, offers several avenues for improvement and innovation. Addressing the challenge of cross-cultural and multilingual adaptation will be crucial to developing more inclusive and globally applicable emotion recognition systems. Additionally, advancements in real-time processing capabilities and the integration of contextual understanding can enhance the responsiveness and accuracy of these systems, enabling more natural and nuanced interactions. Moreover, research efforts focused on mitigating bias and ensuring fairness in emotion recognition algorithms will be essential to building trustworthy and ethically sound systems. By overcoming these limitations, the future of speech emotion recognition holds the potential to revolutionize human-computer interaction, ushering in an era of more empathetic and personalized technology experiences.

**REFERENCES**

[1] Komal Bharti, Prachi Shrivastava, Pradip K. Das, "Comparison of Multiple Classifiers for Audio-Visual Speaker Recognition System", 2023 International Conference on Computational Intelligence, Networks and Security (ICCINS), pp.1-6, 2023.

[2] Alluhaidan, A.S.; Saidani, O.; Jahangir, R.; Nauman, M.A.; Neffati, O.S. Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network. Appl. Sci. 2023, 13, 4750. https://doi.org/10.3390/app13084750.

[3] Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, Mohamed Ali Mahjoub, et al.. Automatic Speech Emotion Recognition Using Machine Learning. Social media and Machine Learning [Working Title], Intech Open, 2019, ⟨10.5772/intechopen.84856⟩. ⟨hal-02432557⟩.

[4] Ramesh, S., Gomathi, S., Sasikala, S. et al. Automatic speech emotion detection using hybrid of gray wolf optimizer and naïve Bayes. Int J Speech Technol 26, 571–578 (2023). https://doi.org/10.1007/s10772-021-09870-8.

[5] Abolfazl Younesi, Mohsen Ansari, Mohammadamin Fazli, Alireza Ejlali, Muhammad Shafique, Jörg Henkel, "A Comprehensive Survey of Convolutions in Deep Learning: Applications, Challenges, and Future Trends", IEEE Access, vol.12, pp.41180-41218, 2024.

[6] L. Yunxiang and Z. Kexin, "Design of Efficient Speech Emotion Recognition Based on Multi Task Learning," in IEEE Access, vol. 11, pp. 5528-5537, 2023, doi: 10.1109/ACCESS.2023.3237268.

[7] Mehmet Berkehan Akçay, Kaya Oğuz, Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers, Speech Communication, Volume116,2020, ISSN0167-6393.

[8] Francesco Ardan Dal Rí, Fabio Cifariello Ciardi, Nicola Conci, "Speech Emotion Recognition and Deep Learning: An Extensive Validation Using Convolutional Neural Networks", IEEE Access, vol.11, pp.116638-116649, 2023.

[9] Chi-Chun Lee, Theodora Chaspari, Emily Mower Provost, Shrikanth S. Narayanan, "An Engineering View on Emotions and Speech: From Analysis and Predictive Models to Responsible Human-Centered Applications", Proceedings of the IEEE, vol.111, no.10, pp.1142-1158, 2023.

[10] Changzeng Fu, Chaoran Liu, Carlos Toshinori Ishi, Hiroshi Ishiguro, "An Adversarial Training Based Speech Emotion Classifier with Isolated Gaussian Regularization", IEEE Transactions on Affective Computing, vol.14, no.3, pp.2361-2374, 2023.

[11] Zhen-Tao Liu, Qiao Xie, Min Wu, Wei-Hua Cao, Ying Mei, Jun-Wei Mao, Speech emotion recognition based on an improved brain emotion learning model, Neurocomputing, Volume 309,2018, ISSN 0925-2312, <https://doi.org/10.1016/j.neucom.2018.05.005>.

[12] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, Te-Won Lee, Emotion Recognition by SpeechSignals,2023,https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=e9a0cade5b8ce466429fc033f8e75b6a735ec761.

[13] Yue Xie, Ruiyu Liang, Zhenlin Liang, Chengwei Huang, Cairong Zou, Björn Schuller, "Speech Emotion Classification Using Attention-Based LSTM", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.27, no.11, pp.1675-1685, 2019.

[14] Dariusz Czerwinski, Pawel Powroznik, "Human Emotions Recognition with the Use of Speech Signal of Polish Language", 2018 Conference on Electrotechnology: Processes, Models, Control and Computer Science (EPMCCS), pp.1-6, 2018.

[15] Cheng Lu, Yuan Zong, Wenming Zheng, Yang Li, Chuangao Tang, Björn W. Schuller, "Domain Invariant Feature Learning for Speaker-Independent Speech Emotion Recognition", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.30, pp.2217-2230, 2022.