Dissertation on

"Speech Emotion Recognition through Federated Learning for Quality Assurance in Call Centres"

Submitted in partial fulfilment of the requirements for the award of degree of

Bachelor of Technology
in
Computer Science & Engineering

UE20CS390A – Capstone Project Phase - 1

Submitted by:

Andre Aditya Roy	**PES2UG20CS048**
Ann Kurian	**PES2UG20CS055**
Arshan Lawrence Rodrigues	**PES2UG20CS067**
Kristal Joanne D'souza	**PES2UG20CS170**

Under the guidance of

Prof. Swathy M
Assistant Professor
PES University

January - May 2023

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
FACULTY OF ENGINEERING
PES UNIVERSITY
(Established under Karnataka Act No. 16 of 2013)
Electronic City, Hosur Road, Bengaluru – 560 100, Karnataka, India

PES UNIVERSITY

(Established under Karnataka Act No. 16 of 2013)
Electronic City, Hosur Road, Bengaluru – 560 100, Karnataka, India

FACULTY OF ENGINEERING

CERTIFICATE

This is to certify that the dissertation entitled

'Speech Emotion Recognition through Federated Learning for Quality Assurance in Call Centres'

is a bonafide work carried out by

Andre Aditya Roy	**PES2UG20CS048**
Ann Kurian	**PES2UG20CS055**
Arshan Lawrence Rodrigues	**PES2UG20CS067**
Kristal Joanne D'souza	**PES2UG20CS170**

In partial fulfilment for the completion of sixth semester Capstone Project Phase - 1 (UE20CS390A) in the Program of Study -Bachelor of Technology in Computer Science and Engineering under rules and regulations of PES University, Bengaluru during the period Jan. 2023 – May. 2023. It is certified that all corrections / suggestions indicated for internal assessment have been incorporated in the report. The dissertation has been approved as it satisfies the 6th semester academic requirements in respect of project work.

Signature **Prof. Swathy M** Assistant Professor	Signature Dr. Sandesh B J Chairperson	Signature Dr. B K Keshavan Dean of Faculty

External Viva

Name of the Examiners	**Signature with Date**
1. _____	_____
2. _____	_____

DECLARATION

We hereby declare that the Capstone Project Phase - 1 entitled **"Speech Emotion Recognition through Federated Learning for Quality Assurance in Call Centres"** has been carried out by us under the guidance of Prof. Swathy M, Assistant Professor and submitted in partial fulfilment of the course requirements for the award of degree of **Bachelor of Technology** in **Computer Science and Engineering** of **PES University, Bengaluru** during the academic semester January – May 2023. The matter embodied in this report has not been submitted to any other university or institution for the award of any degree.

Andre Aditya Roy	**PES2UG20CS048**	**<Signature>**
Ann Kurian	**PES2UG20CS055**	**<Signature>**
Arshan Lawrence Rodrigues	**PES2UG20CS067**	**<Signature>**
Kristal Joanne D'souza	**PES2UG20CS170**	**<Signature>**

ACKNOWLEDGEMENT

I would like to express my gratitude to Prof. Swathy M, Department of Computer Science and Engineering, PES University, for her continuous guidance, assistance, and encouragement throughout the development of this UE20CS390A - Capstone Project Phase – 1.

I am grateful to the Capstone Project Coordinators, Dr. Sarasvathi V, Professor and Dr. Sudeepa Roy Dey, Associate Professor, for organizing, managing, and helping with the entire process.

I take this opportunity to thank Dr. Sandesh B J, Professor & Chairperson, Department of Computer Science and Engineering, PES University, for all the knowledge and support I have received from the department. I would like to thank Dr. B.K. Keshavan, Dean of Faculty, PES University for his help.

I am deeply grateful to Dr. M. R. Doreswamy, Chancellor, PES University, Prof. Jawahar Doreswamy, Pro Chancellor – PES University, Dr. Suryaprasad J, Vice-Chancellor, PES University for providing to me various opportunities and enlightenment every step of the way. Finally, this Capstone Project could not have been completed without the continual support and encouragement I have received from my family and friends.

ABSTRACT

There are numerous applications for emotion recognition in call centres, including the avoidance of customer abuse, efficient matching of customers versus agents, etc., Federated Learning is a promising approach to address these challenges, as it enables the training of a highly accurate model while maintaining the privacy of agents and operational secrets. It can help call centers improve their customer experience and take their operations to the next level

Combining SER with FL can be a powerful tool for call centers to improve their operations. By using FL, call centers can train a highly accurate SER model without compromising the privacy of their agents or their operational secrets. The FL approach ensures that the sensitive data remains on the local devices/servers and is not sent to a central location, lowering the possibility of data breaches while ensuring that data protection laws are followed.

The trained SER model can then be used to classify customer emotions in real-time, providing agents with valuable information that can help them tailor their responses to the customer's emotional state. For example, if the SER model detects that a customer is angry, the agent can use a more empathetic tone and offer a solution that is more likely to satisfy the customer.

In addition to improving customer service, SER through FL can also be used for quality assurance purposes. Call centers can use the SER model to analyze customer interactions and identify areas where agents need further training or support. This can help call centers improve their overall operations and ensure that they are meeting their performance goals. Overall, SER through FL is a promising approach to improving the customer experience in call centers. By leveraging the power of AI and machine learning, call centers can provide more personalized and effective customer service while protecting the privacy of their agents and their operational secrets..

TABLE OF CONTENTS

LIST OF DIAGRAMS

CHAPTER 1

INTRODUCTION

1.1 Federated Learning

Federated Learning (FL) is a decentralised approach to machine learning that facilitates models to be trained across numerous sites without requiring sharing of data. In traditional centralized machine learning, all data is aggregated in a central location for model training. However, in some scenarios, it may not be possible or desirable to centralize the data, such as in the case of sensitive data or distributed data sources.

FL addresses this challenge by training machine learning models across multiple decentralized sites, such as mobile devices, IoT devices, or servers. Instead of sending the data to a central location for training, the models are trained locally on each site, and the updated models are then aggregated to create a global model. This approach allows the training of models on sensitive or distributed data without compromising data privacy or security.

1.2 Quality Assurance in Call Centers

The use of call centers has become ubiquitous in modern-day businesses as they offer a direct and efficient way of addressing customer queries and concerns. However, the quality of these interactions is of utmost importance as it plays a crucial role in maintaining customer satisfaction and loyalty. To ensure the quality of these interactions, call centers typically employ Quality Assurance (QA) analysts to monitor calls and provide feedback to agents.

1.3 Speech Emotion Recognition

Speech Emotion Recognition (SER) is a promising technology that can be used to automatically detect and classify the emotional content of a conversation. However, the development of accurate SER models requires a large amount of data, which may be difficult to obtain in call center environments.

To address this challenge, we propose the use of Federated Learning (FL), a decentralised machine learning technique that permits model training across numerous sites without the necessity for data sharing.

In this project, we aim to develop an SER model using FL that can be deployed in call centers to automate the QA process.

CHAPTER 2

PROBLEM STATEMENT

Call centers receive a large number of customer calls on a daily basis, and it is important to ensure that these calls are of high quality and customer satisfaction is maintained. Assessing the quality of customer interactions with agents can be a challenging task as traditional methods of quality assurance rely on manual evaluation of calls, which can be time-consuming and prone to errors.

Therefore, there is a need for an automated system that can accurately and efficiently evaluate customer interactions with agents. However, building such a system requires a large amount of labeled data, thereby bringing forward three main issues:

a. Difficult to obtain from a single call center

b. Difficulty in obtaining data from different call centers due to privacy concerns

c. Data storage

CHAPTER 3

LITERATURE SURVEY

3.1 Emotion Recognition Method for Call/Contact Centre Systems [1]

3.1.1 Introduction:

The introduction provides an overview of the research on call/contact center systems and the importance of automation in these systems. It discusses the role of virtual assistants and the challenges of correctly recognizing client intentions, particularly in relation to emotional states. The article introduces a new technique for identifying emotions in contact center systems that can be used to improve the effectiveness of virtual assistants and increase customer satisfaction.

3.1.2 Characteristics and Implementation:

This section describes the key characteristics of introduced emotion recognition methods, which includes the capacity to recognize emotional states in two channels namely text and voice channels. The implementation details of the method are also discussed, including the use of ML techniques to train the model and the use of actual conversations from a commercial contact center to validate the method's effectiveness.

3.1.3 Features:

The method is able to recognize emotional states in both clients and agents, which can be used to build behavioral profiles and improve the efficiency of contact center operations. The article also explores the potential for automatic transcriptions of recordings to be used for emotion evaluation, which can improve the accuracy of recognising emotion in the voice channel.

3.1.4 Evaluation:

The article presents an evaluation of the introduced emotion recognition technique, which involved using actual conversations from a commercial contact center to test the efficacy of the method. The evaluation shows that the method proposed is effective in recognizing the emotional state in the text as well as voice channels, with high accuracy and precision.

3.1.5 Advantage:

The benefit of the method proposed is that it provides a more accurate and efficient way of recognizing emotional states in contact center systems, which can be used to improve the effectiveness of virtual assistants and increase customer satisfaction. The method can also be used to build behavioral profiles of agents, which can improve the efficiency of contact center operations.

3.1.6 Limitations:

One limitation of the proposed emotion recognition method is that it relies on accurate transcriptions of conversations, which can be affected by background noise and other factors.

To attain high levels of accuracy, the approach also needs a significant amount of training data. Finally, the proposed method may not be effective in recognizing emotions in clients or agents who are skilled at masking their emotional states.

3.2 German End-to-end Speech Recognition based on DeepSpeech [2]

3.2.1 Introduction:

The paper begins by highlighting the importance of automatic speech recognition and how it is difficult to find freely available models for languages other than English. The authors then describe the process of training German models for automatic speech recognition using the Mozilla DeepSpeech architecture and publicly available data. This is a significant contribution to the field, as it provides freely available German models for automatic speech recognition.

3.2.2 Characteristics and Implementation:

The paper discusses in detail how the training of the German models was carried out, including the use of pre-processing and hyperparameters tuning. The authors highlight the importance of pre-processing and tuning the hyperparameters to obtain better results. They also provide information on the publicly available data used for training. The process of training the models is detailed, providing a good understanding of how the models were created.

3.2.3 Features:

The paper compares the German models with other available speech recognition services for German and finds that the results are comparable. The authors release their trained German models and training configurations, which is a valuable resource for researchers and developers who want to use the models for their own work.

3.2.4 Evaluation:

The paper focuses on the accuracy of the German models as the performance measure. The authors provide detailed information on the performance of the models, including their Word Error Rate (WER) and comparison with other speech recognition services for German. However, the authors note that acceptable performance under noisy conditions would require more training data. The paper does not explore the effect of different recording environments or noise levels on emotion classification.

3.2.5 Advantage:

The paper makes freely available German models for automatic speech recognition using the Mozilla DeepSpeech architecture. This is a significant contribution to the field as it provides researchers and developers with a valuable resource that can be used for their own work. The detailed process of creating the models also provides insights into how to train models for other languages.

3.2.6 Limitations:

The paper does not explore the effect of different recording environments or noise levels on emotion classification. This is an important limitation as recording environments and noise levels can have a significant impact on the performance of automatic speech recognition

models. Additionally, the paper focuses only on accuracy as the performance measure, and it would have been beneficial to explore other performance measures as well.

3.3 Emotion in Speech: Recognition and Application to Call Centers [3]

3.3.1 Introduction:

Two experimental investigations are described in the study that investigate the identification of vocal emotions in speech. The first research used a corpus of 700 brief utterances to show five distinct emotions, while the second used a corpus of 56 telephone messages to express largely normal and furious emotions. The experiments' purpose was to create and evaluate neural network recognizers for various emotional states, which would then be used in a decision support system for contact centres.

3.3.2 Characteristics and Implementation:

In the first research, 30 non-professional actors acted out five emotions: anger, fear, happiness, sadness, and normal mood. Relevant characteristics were recovered from this corpus utilising feature selection techniques, including pitch, first and second formants, energy, and speaking rate statistics. Backpropagation neural network models were trained, and multiple neural network recognisers and recogniser ensembles were constructed. The second research employed a corpus of 56 telephone calls recorded by 18 non-professional actors, and recognisers were produced using the previous study's methods. A group of such recognizers was employed in a decision support system to prioritise voice communications and assign an agent to reply to them.

3.3.3 Features:

The relevant features selected for the first study included speaking rate, energy, statistics of pitch and lastly first & second formants. The recognizers created in both studies used neural network models. Pitch (basic frequency) is the primary vocal cue for emotion identification, according to all studies in the subject. In addition to vocal energy, frequency spectrum properties, formants, and temporal elements (speech pace and pauses), there are other acoustic factors that influence vocal emotion signalling. A different method of extracting features is to add to the list of features by taking into account some derivative characteristics, such as the signal's LPC parameters or the features of the smoothed pitch contour and its derivatives.

3.3.4 Evaluation:

The recognizers in the first study demonstrated accuracy rates ranging from 60-75% for normal state, 60-70% for happiness, 70-80% for anger, 70-85% for sadness, and about 35-55% for the emotion fear. The overall average accuracy was found to be approximately 70%. In the second trial, recognisers had an average accuracy of 77% in distinguishing between "agitation" (including anger, pleasure, and fear) and "calm" (which covers sadness and normal condition).

3.3.5 Advantage:

The paper presents a methodology for recognizing vocal emotions in speech and demonstrates the viability of performing this task using neural network models.

The decision support system using the ensemble of recognizers could be a useful tool for call centers to prioritize voice messages and assign agents to respond to them. The developed recognition models have practical applications, such as in call centers for prioritizing and directing voice messages to appropriate agents based on the caller's emotional state.

3.3.6 Limitations:

The studies involved non-professional actors and a relatively small corpus of utterances, which may limit the generalizability of the results. The recognizers had lower accuracy rates for some emotional states, such as fear, which may limit their effectiveness in real-world applications. The paper also does not explore the effect of different recording environments or noise levels on emotion classification.The accuracy of emotion recognition for the fear emotion was relatively low, at 35-55%, indicating a potential challenge in accurately recognizing this emotion from speech.

3.4 Analysis of Vocal Pattern to Determine Emotions using Machine Learning. [4]

3.4.1 Introduction:

This paper discusses how ML algorithms are used to analyze vocal patterns and accurately detect emotions. The authors compare the performance of different classifiers on a dataset of recorded speech samples labeled with emotional categories. The results show that machine learning can effectively identify emotions from vocal patterns with high accuracy, which has important implications for fields such as psychology, medicine, and entertainment.

3.4.2 Characteristics and Implementation:

Uses a dataset of recorded speech samples labeled with emotional categories. Shows that machine learning can effectively identify emotions from vocal patterns with high accuracy. The authors likely collected a dataset of speech samples labeled with emotional categories and analyzed the results of the model to draw conclusions about the effectiveness of different machine learning techniques for emotion detection from vocal patterns.

3.4.3 Features:

Use of a dataset of recorded speech samples labeled with emotional categories for training and testing machine learning classifiers. Comparison of different classifiers, that consists of k-NNs, SVMs and DTs, for emotion detection from vocal patterns. Detailed analysis and evaluation of the performance of the classifiers to identify the most effective machine learning technique for emotion recognition.

3.4.4 Evaluation:

Classifier performance was more likely tested using measures like precision, F1-score, recall, and accuracy. These metrics quantify the classifier's ability to properly detect emotions based on vocal patterns.. Additionally, the authors may have conducted a more qualitative analysis of the results, comparing the performance of the different classifiers and identifying which ones performed best for each emotional category.

3.4.5 Advantage:

The project provides a way to accurately detect emotions from vocal patterns, which has important implications for fields such as psychology, medicine, and entertainment. It allows for automation of emotion detection from vocal patterns, which can save time and resources compared to manual analysis. The project can be easily scaled to analyze large datasets of vocal patterns, which can provide insights into emotional trends and patterns on a larger scale.

3.4.6 Limitations:

This dataset used may not fully represent the diversity of vocal patterns and emotional expressions across different populations and cultures. Therefore, the results of the project may not be generalizable to other populations and cultures. Additionally, the project's performance may be impacted by factors such as the quality of the recordings or the accuracy of the emotional labels.

3.5 Acoustic and Lexical Sentiment Analysis for Customer Service Calls [5]

3.5.1 Introduction:

The paper presents a study on analyzing the sentiment of customer service calls using both acoustic and lexical features. The authors aim to enhance the sentiment analysis accuracy in this context by combining the two types of features. The paper provides a summary of related work on sentiment analysis and discusses the challenges of analyzing customer service calls, such as the presence of multiple speakers, the use of domain-specific vocabulary, and the need to consider the overall context of the conversation.

3.5.2 Characteristics and Implementation:

The paper proposes a methodology that extracts both acoustic (e.g., pitch, volume) and lexical (e.g., sentiment-bearing words) features from the audio recordings of customer service calls, and then uses machine learning algorithms to classify the sentiment of each segment of the conversation. The implementation of the methodology involves several steps, including data collection, feature extraction, and machine learning-based sentiment classification.

3.5.3 Features:

The features extracted from the audio recordings include pitch, volume, and sentiment-bearing words. ML algorithms are used to classify the sentiment of each segment of the conversation, and the effectiveness of the methodology is evaluated on a dataset of customer service calls.

3.5.4 Evaluation:

The dataset is annotated with ground-truth sentiment labels for each segment of the conversation. The authors assess the performance of the methodology using metrics such as precision, F1-score, recall, and accuracy. The results show that combination of lexical and acoustic features leads to improved accuracy in sentiment analysis compared to using either type of feature alone.

3.5.5 Advantage:

By combining both acoustic and lexical features, the methodology achieves higher accuracy in sentiment analysis of customer service calls compared to using either type of feature alone. The methodology could be applied to large datasets of customer service calls and automated to provide real-time feedback on the sentiment of customer interactions, which could help businesses improve their customer service operations more efficiently.

3.5.6 Limitations:

Paper relies on the quality of the audio recordings and the accuracy of the annotations. If the audio quality is poor or the annotations are inaccurate, it could impact the performance of the sentiment analysis methodology. Additionally, the methodology is designed for English-speaking customer service calls, and may not be as effective in other languages or cultural contexts.

3.6 Temporal Context in Speech Emotion Recognition [6]

3.6.1 Introduction:

The paper explores the role of temporal context in SER and proposes a new approach that takes into account the context of emotional events in speech. The authors argue that emotions are not static but rather evolve over time, and therefore, temporal context is an essential factor in accurately recognizing emotions in speech.

3.6.2 Characteristics and Implementation:

LSTM neural network to model the temporal context of emotional events in speech. The LSTM network is trained to predict emotional labels at each time step, given the previous

time steps' acoustic features and emotional labels. The LSTM network's hidden state is used to capture the temporal context of emotional events in speech. The proposed approach models the temporal context of emotional events in speech by considering a sliding window of input frames. The sliding window size is a hyperparameter that can be adjusted to control the temporal context modeling's granularity.

3.6.3 Features:

By combining Acoustic features, Temporal features and Emotional Labels the authors propose a new approach that takes into account the temporal context of emotional events in speech. The speech signals in both datasets are segmented into 1-second frames and converted into Mel-Frequency Cepstral Coefficients (MFCCs). The MFCCs are then stacked to form a 2D matrix representation of the speech signal.

3.6.4 Evaluation:

The proposed approach uses two publicly available datasets: the IEMOCAP dataset and the EPST dataset.

3.6.5 Advantage:

Improved accuracy: The proposed approach outperforms baseline models that do not consider temporal context, improving the accuracy of SER.

Robustness: The proposed approach is robust to variations in emotional expression and background noise.

Efficiency: The proposed approach is computationally efficient and can be trained on a single GPU.

3.6.6 Limitations:

Model complexity: The proposed approach uses a complex model, which may be difficult to interpret and require significant computational resources.

Generalizability to other languages: The proposed approach is evaluated on English datasets and may not generalize well to other languages.

3.7 Improved speech emotion recognition with Mel frequency magnitude coefficient [7]

3.7.1 Introduction:

The paper "Improved speech emotion recognition with Mel frequency magnitude coefficient" by Ancilin and Milton proposes a new approach for SER using MFMC. The paper describes the implementation and evaluation of this approach and compares its efficacy to various cutting-edge approaches.

3.7.2 Characteristics and Implementation:

The proposed approach in the paper uses MFMC features extracted from speech signals to train a classifier for SER. The MFMC characteristics originate from the magnitude spectrum of the speech signal's short-time Fourier transform. The authors use a machine learning

algorithm (support vector machine) to categorise the voice signals according to their emotional content. The implementation of the approach involves data preprocessing, feature extraction, and model training..

3.7.3 Features:

The key features of the proposed approach in the paper include:

Mel Frequency Magnitude Coefficients (MFMC): The approach uses MFMC features to capture the spectral characteristics of the speech signal, which are related to emotional expression.

SVM: The approach uses SVM to classify the speech signals into different emotional categories.

3.7.4 Evaluation:

The performance of the method proposed in the paper is evaluated on two datasets: the Berlin Emo-DB and EPST dataset. The evaluation involves several steps, including data preprocessing, feature extraction, model training, and performance evaluation. The performance is evaluated using metrics such as precision, recall, F1 score and accuracy.

3.7.5 Advantage:

Improved accuracy: The method outperforms other cutting-edge approaches for SER, achieving high accuracy on the evaluated datasets.

Simplicity: The approach is simple and easy to implement, requiring only feature extraction and classification.

Efficiency: The approach is computationally efficient and can be trained on a single CPU.

3.7.6 Limitations:

Limited generalizability: The approach is evaluated on two datasets and may not generalize well to other datasets or emotional categories.

Limited interpretability: The approach uses a black-box model (SVM) that may be difficult to interpret.

Data preprocessing: The approach requires extensive preprocessing of the speech signals, including feature extraction and normalization, which can be time-consuming.

3.8 Sentiment Analysis of Call Centre Audio Conversations using Text Classification [8]

3.8.1 Introduction:

This paper explores the use of text mining techniques to analyze emotions in transcribed audio recordings. To acquire a better grasp of the content of these recordings, the authors suggest a unique approach that blends voice recognition technology with text categorization techniques. The research is motivated by potential applications in call centers and other areas.

3.8.2 Characteristics and Implementation:

The suggested strategy entails transcribing audio recordings with automatic speech recognition technologies, followed by text-based sentiment analysis methodologies. Authors

explore different feature selection methods and evaluate the accuracy of their approach using various metrics.

3.8.3 Features:

Some of the features explored in this paper include different feature selection methods, like chi-squared feature selection and TF-IDF. Authors also explore different ML algorithms for text classification, like SVMs and DTs.

3.8.4 Evaluation:

The authors evaluate the accuracy of their approach using various metrics, such as F1-score, recall, and precision. They also compare their results to those obtained using other sentiment analysis techniques.

3.8.5 Advantages:

One advantage of the proposed approach is that it can be used to analyze emotions in large volumes of audio recordings quickly and efficiently. It also has potential applications in call centers and other areas where understanding customer emotions is important.

3.8.6 Limitations:

One limitation of this approach is that it relies on accurate transcription of audio recordings, which can be challenging for some types of speech or accents. It also requires significant computational resources for processing large volumes of data.

CHAPTER 4

DATA

4.1 Overview

The dataset required is the call recordings of multiple call centers or processes. There will be a single call recording for each call which consists of both the customer's and agent's voice. The dataset consists of call recordings that are sourced from an Indian call center; this implies we must train our model to work with the Indian language and also learn how to handle Indian slang.To ensure applicability in real-life situations and the authenticity of the call recordings, the dataset is sourced from real-life call centers.

4.2 Dataset

The dataset in use will have to first be denoised to remove any background noise or disturbance or any other inconsistency. This data will then be segmented and fed to the model. The recordings will be broken down into smaller segments and labeled individually for training the data. The segmentation of the training data will ensure that fluctuations in the

emotions of a customer are recorded and also will help the model give the user (the call center manager in this case) an overview of the range of emotions happening throughout the call. In conclusion, The dataset used is sourced from local call centers and it is then fine-tuned for a more precise result.

CHAPTER 5

SYSTEM REQUIREMENTS SPECIFICATION

5.1 Functional Requirements

Audio Pre-processing: This refers to the functional requirement of the SER system to process the audio recordings of call center conversations before analysis. This could involve filtering out background noise, normalization of audio levels, removal of silences, and other such pre-processing techniques to improve the accuracy of emotion recognition.

Accurate Indian English Recognition: This refers to the functional requirement of the system to be able to obtain emotions accurately from speech in the Indian English language. This would involve training the model using Indian English speech data and ensuring that the system can handle variations in accents, intonations, and other speech characteristics specific to the Indian English language.

Speech-to-text conversion: This refers to the functional requirement of the system to convert the audio recordings of call center conversations into text format for further analysis. This would involve using speech recognition techniques to transcribe the speech signal into text.

Multi-modal Analysis: This refers to the functional requirement of the system to use a combination of both acoustic and linguistic features to improve the accuracy of emotion recognition. Acoustic features could include pitch, tone, and energy, while linguistic features could include sentiment and word choice.

Sentiment Analysis Model Selection: This refers to the functional requirement of the system to explore different sentiment analysis models based on the relevant data attributes such as language, domain, and context of the call center conversation recordings. This would involve testing different machine learning models and selecting the one that gives the best results for the given data attributes.

Emotion labelling: This refers to the functional requirement of the system to label the different emotions present in the call center conversation recordings. The system would need to classify the emotions into different categories such as happiness, anger, sadness, fear, and neutral, based on the analysis of the acoustic and linguistic features of the speech signal.

5.2 Non - Functional Requirements

5.2.1 Performance Requirement

Reliability: The dependability of a system refers to its capacity to execute the required function consistently and accurately under varying situations. The system's reliability is determined by the quality of the voice data, the accuracy of the model, and the stability of the computing and communication resources.

Robustness: The system's robustness refers to its ability to perform consistently and accurately under different conditions, including noisy or degraded speech signals. The system should be designed to handle variations in the speech data, such as background noise, accents, and speech patterns, and adapt to them to maintain accurate emotion recognition.

Availability: The system should be designed to handle high volumes of speech data and model updates without affecting its performance or availability.

Performance: The system's performance refers to its ability to perform the desired function efficiently and accurately within a specific time frame. The system's performance depends on the quality of the speech data, the complexity of the model, the computing resources available, and the communication network's speed and reliability.

5.2.2 Safety Requirements

The system should be designed to ensure the privacy and security of the speech data being transferred and processed

The system should be designed to comply with ethical considerations regarding the use of speech data.

Mitigate potential biases in the data and the model that may result in unfair or discriminatory outcomes

5.2.3 Security Requirements

Access control: The system should have a secure login mechanism with authentication and authorization controls to ensure only authorized personnel have access to the system.

Data confidentiality: All voice data recorded during call center conversations must be kept confidential and should only be accessed by authorized personnel who need it for analysis purposes.

Confidentiality of User Information: Private information of the users will be kept confidential so that third parties will not get access to such information.

5.3 External Interface Requirements

5.3.1 User Interfaces

Required screen formats with GUI standards for styles: The interface should follow a standard GUI format with a consistent style that is easy to use and understand by all user classes.

Screen layout and standard functions: The layout of the screens should be organized and intuitive, with standard functions such as help available to guide users through the process.

Relative timing of inputs and outputs: The interface should provide timely responses to user inputs, with appropriate feedback and response times.

Availability of programmable function keys: The interface should provide programmable function keys that can be customized to meet the needs of different user classes.

Error messages: The interface should provide clear and concise error messages that inform users of any issues and guide them through the process of correcting them.

5.3.2 Hardware Requirements

- A server with sufficient processing power, memory, and storage to host the federated learning model and manage communication between nodes.

- Multiple client devices (such as smartphones or laptops) with adequate processing power and memory to perform local training and inference tasks.

- Reliable network connectivity with sufficient bandwidth to transfer data between the server and client devices.

- Adequate backup and recovery mechanisms to ensure the availability and integrity of data in case of hardware failure or data loss.

5.3.3 Software Requirements

- Name and Description : The system is a SER platform that utilizes federated learning to train models on customer-agent conversations in call centers. The system aims to improve quality assurance in call centers by identifying emotional patterns in customer-agent interactions.

- Version / Release Number : The initial release of the system will be version 1.0.

- Operating Systems : The system should be compatible with common operating systems, including Windows, macOS, and Linux. The system should be tested on each of these platforms to ensure compatibility.

- Tools and libraries : The system will require the following tools and libraries:
  Python 3.x: The system will be developed using Python programming language, which is commonly used in machine learning and data processing applications.

TensorFlow: TensorFlow is a ML library that is open source and will be used to train the emotion detection models.

PyTorch: PyTorch is another popular machine learning library that will be used to train emotion recognition models.

Keras: Keras is an API for high-level NNs that will be used to build and train the emotion recognition models.

NumPy and Pandas: NumPy and Pandas are popular libraries used for data manipulation and processing.

Scikit-learn: Scikit-learn is a ML library particularly for analysis of data and modeling.

Librosa: Librosa is a Python package for analyzing and processing audio data.

DeepSpeech, Vosk, and CMU Sphinx: These are all speech recognition frameworks that can be used to convert audio recordings into text.

Flask: Flask is a micro web framework that will be used to build the system's user interface.

TensorFlow Federated (TFF)/OpenMined/PySyft/Flower : These are a few federated learning frameworks that can be chosen to implement the system.

Source : The system's source code will be available on a public GitHub repository for developers to review, contribute to, and provide feedback.

5.3.4 Communication Interfaces

- Speech Data Transfer: The product may require the transfer of large amounts of speech data between the call center system and the Federated Learning server. To minimize the transfer time, the product may require a high-speed network connection

with low latency. The data transfer may use TCP or UDP depending on the reliability and latency requirements.

- Model Update Transfer: The product may require the transfer of model updates between the Federated Learning parties. To minimize the transfer time, the product may require a high-speed network connection with low latency. The data transfer may use TCP or UDP depending on the reliability and latency requirements.

CHAPTER 6

SYSTEM DESIGN

6.1 Design Considerations

6.1.1 Design Goals:

- Accuracy: The system should strive for high accuracy when it comes to recognising speech emotions. Using modern machine learning models and methodologies, as well as training the models on various and representative data sets, the accuracy may be enhanced.

- Privacy: The system should be built to ensure that the voice data being sent and processed is kept private. To prevent data from unauthorized access or disclosure, encryption and security measures can be used to maintain privacy.

- Scalability: The system should be built to handle enormous amounts of speech data and to scale up and down as needed. Scalability may be attained through the use of distributed computing frameworks and cloud-based infrastructures.

- Robustness: The system should be built to be resistant to noise, distortion, and other fluctuations in the voice data. The resilience may be obtained by employing signal processing techniques and training the models on noisy data sets.

- User-centered design: The system should be created with the users' wants and preferences in mind. This may be accomplished through the use of user research, usability testing, and user input to guide design decisions.

- Simplicity: The system should be simple to use, comprehend, and traverse. The design should be maintained basic and straightforward to minimise needless complexity and misunderstanding.

6.1.2 Architecture Choices:

In a decentralized architecture of FL for SER in Call Centers, the audio data is distributed across different call centers or nodes, and the machine learning models are trained locally on the data. The model updates are then periodically shared with a central server for aggregation.

Decentralized architectures have several advantages, including:

- Privacy preservation: In a decentralised architecture, speech data is stored locally on each call centre, and model training is also performed locally. Because the data is not shared with a central server, the privacy of the speech data is preserved.

- Reduced data transfer: Since only the model updates are transferred to the central server, decentralized architectures reduce the amount of data that needs to be transferred, which can save bandwidth and reduce latency.

- Local customization: Decentralized architectures allow for more local customization of the machine learning models since the models are trained on data specific to each call center. This can lead to more accurate and relevant models for each call center.

- Fault tolerance: Decentralized architectures are more fault-tolerant since the system can continue to operate even if some of the nodes fail. In this architecture, if one call center goes down, the other call centers can continue to function, and the overall system can still operate.

However, there are also some disadvantages to decentralized architectures, including:

- Complexity: Decentralized architectures can be more challenging to manage and maintain since there are many nodes to manage, and the system requires more complex communication protocols.

- Computational overhead: Decentralized architectures can be more computationally intensive since the models need to be trained locally on each node.

- Coordination: In a decentralized architecture, coordination is required to ensure that the models trained on each node are consistent with each other. This can be challenging to manage.

6.1.3 Constraints, Assumptions and Dependencies

- Interoperability requirements: The system is assumed to be interoperable with other systems that provide call center recordings and transcripts for emotion recognition. The system may require integration with third-party APIs and libraries to perform various tasks such as speech-to-text conversion, linguistic feature extraction, and emotion recognition.

- Interface/protocol requirements: The system is designed to have a user-friendly interface which helps users have the ability to interact with the system and access the results of the emotion recognition process. The system should support standard protocols such as HTTP, HTTPS, TCP, and UDP.

- Data repository and distribution requirements: The system relies on a data repository that contains call center recordings for the emotion recognition model. The data

repository should be scalable and easily accessible by the system. The system should also be able to distribute the results of the emotion recognition process to other systems or users as required.

- Performance-related issues: The performance of the system may be affected by the complexity of the dataset, the quality of the recordings, and the processing power of the hardware. The system should be designed to handle large datasets and scale horizontally to accommodate an increase in the volume of requests.

- End-user environment: The system is designed to be used in a call center environment where call recordings are collected for the training and testing of the SER model. The system should be able to handle a high volume of calls and provide accurate and timely results.

- Availability of resources: The system requires hardware and software resources to operate, including storage, processing power, and memory. The availability of these resources may impact the performance of the system.

- Hardware or software environment: The system should be designed to run on standard hardware and software environments. The hardware should be capable of running machine learning algorithms and the software environment should support the required programming languages and libraries.

- Deployment, Maintainability, Scalability, and Availability: The system should be designed to be deployed easily in the target environment, including cloud-based platforms such as AWS, Azure, or Google Cloud. The system should be maintainable, scalable, and available to handle a high volume of requests. The system should also be designed to support updates and modifications as needed, without disrupting the overall functionality of the system.

- Limitations and Constraints: The system may be limited by the quality of the call recordings and transcripts available for the emotion recognition model. The system may also be limited by the availability of hardware and software resources to support the system's operations. Finally, the system may be limited by privacy and security concerns, which must be addressed to ensure the system's compliance with applicable laws and regulations.

6.1.4 Risks

There are several risks involved during the development and deployment phases of a speech emotion recognition system using federated learning. Some of the risks that may be related to resource requirements or functionality of the proposed system include:

- Data quality : The accuracy and quality of the voice data used to train machine learning models can have an impact on the system's performance and reliability. Poor-quality data can result in inaccurate predictions, affecting the overall performance of the system.

- Resource requirements : Federated learning involves distributing the training of machine learning models across multiple devices or servers, which can increase the resource requirements for the system. This can lead to issues such as increased processing time, decreased system performance, and increased hardware costs.
- Security and privacy : Federated learning involves sharing sensitive data across multiple devices or servers, which can increase the risk of security breaches and privacy violations. The system must be designed with robust security and privacy measures to prevent unauthorized access or data breaches.
- Compatibility issues : The speech emotion recognition system may need to integrate with other software or hardware components, such as call center software, which can lead to compatibility issues. These issues can result in decreased system performance or functional limitations.
- User adoption : The success of the speech emotion recognition system relies on user adoption and satisfaction. If the system is not user-friendly or does not meet user needs, it may not be adopted, limiting its effectiveness and usefulness.
- Regulatory compliance : The use of speech data for emotion recognition may be subject to regulatory compliance requirements, such as GDPR and CCPA. Failure to comply with these regulations can result in legal and financial consequences.

Overall, the risks involved in the progress and deployment of an SER system using federated learning can have a significant impact on the effectiveness and reliability of the system. It is important to carefully consider these risks and implement appropriate measures to mitigate them throughout the development and deployment phases.

6.1.5 User Classes and Characteristics

- Customer - The customer is the one who initiates the call, they are to be provided with appropriate service

- Call center agent - The agent is the one who interacts with the customer and ensures the call goes smoothly and the customer is satisfied

- Call center Manager - The manager overlooks the functioning of the call center and ensures quality by supervising the agents and planning areas of improvement.

6.2 High Level Design

6.2.1 High Level Design Diagram

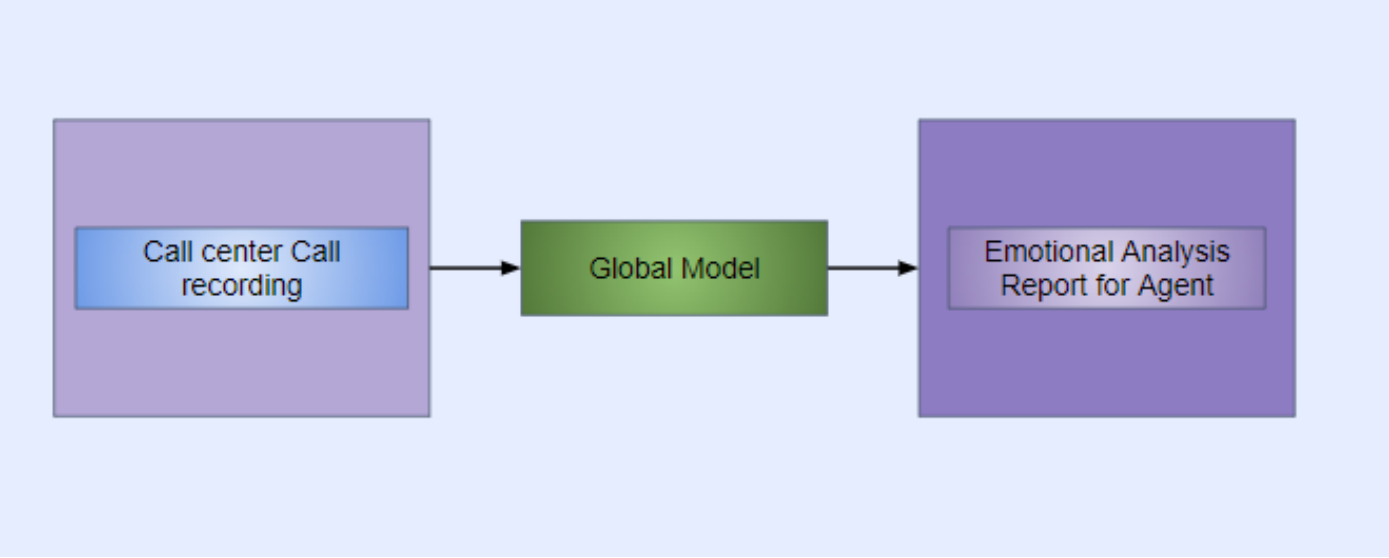

Fig 1.1. a : High Level Design Diagram

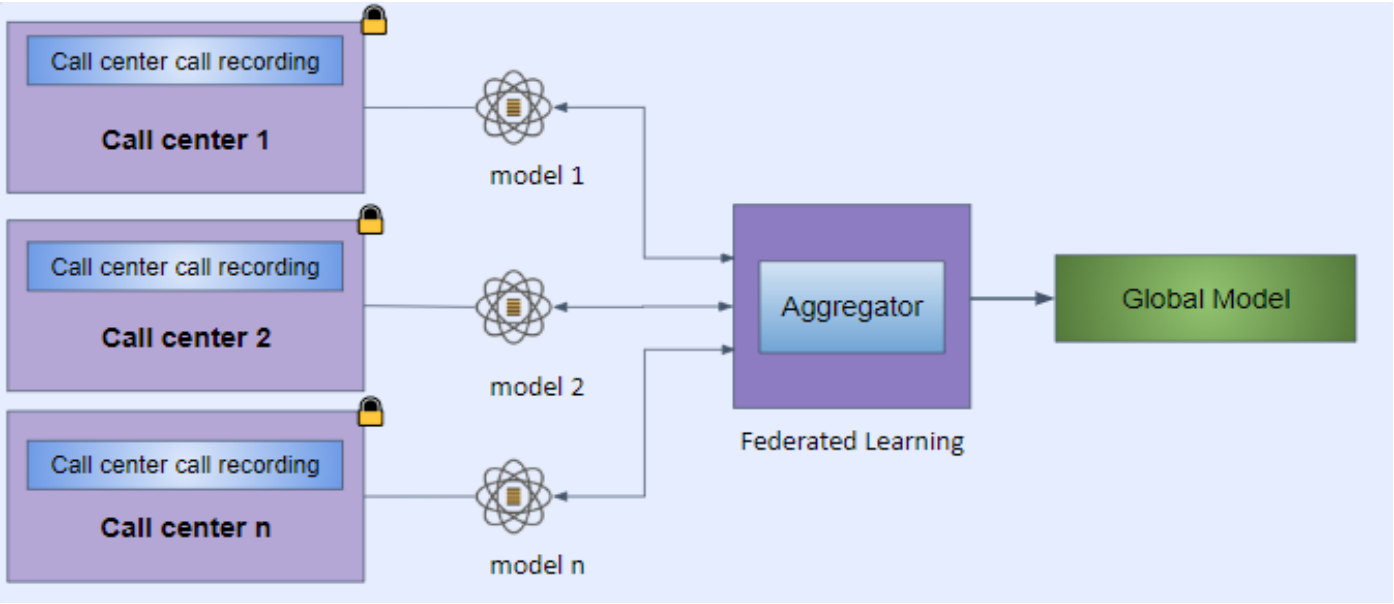

Fig 1.1. b: Model Training

Some of the platforms, systems, and processes that it depends on and may require vital changes are:

- Federated Learning Framework: Framework should support efficient and secure data sharing among multiple parties while preserving data privacy.
- Call Recording System: Captures the audio and transcript of the customer-agent conversations.
- Emotion Recognition API: Should be accurate and compatible with the chosen framework.
- Data Processing: Data processing tools such as TensorFlow and packages like Librosa for analyzing and processing audio data.
- Security and Privacy: Federated learning to ensure adequate security and privacy of customer and agent data.

- Deep learning frameworks: The project may depend on deep learning frameworks like TensorFlow or PyTorch for implementing and training machine learning models.
- Natural language processing libraries: NLP libraries like NLTK, Gensim, or spaCy may be required to preprocess and analyze textual data.
- Audio processing libraries: Libraries like Librosa, PyAudio, or SpeechRecognition may be required for audio preprocessing and analysis.

6.2.2 Architecture Diagram

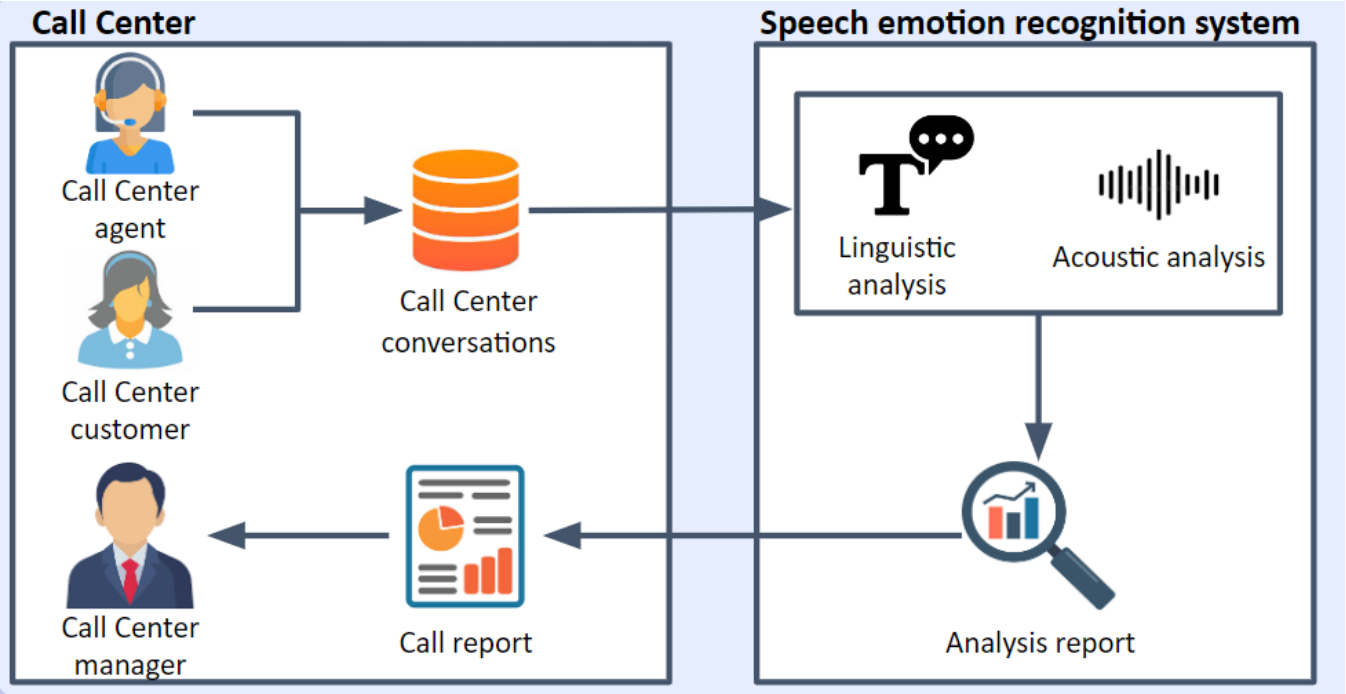

Fig 1.2: Architecture Diagram

6.2.3 System Modules

- Data Preprocessing Module: This module is responsible for cleaning and processing raw speech data collected from the call centers. It includes functions for feature extraction, normalization, and data augmentation.

- Local Model Training Module: This module is responsible for training the machine learning models locally on each call center. It includes functions for model initialization, hyperparameter tuning, and model training using the preprocessed data.

- Model Aggregation Module: This module is responsible for aggregating the local model updates from each call center and updating the global model. It includes functions for model weighting, averaging, and updating.

- Model Evaluation Module: This module is in charge of assessing the global model's performance on test data. It includes functions for model prediction, accuracy calculation, and result visualization.

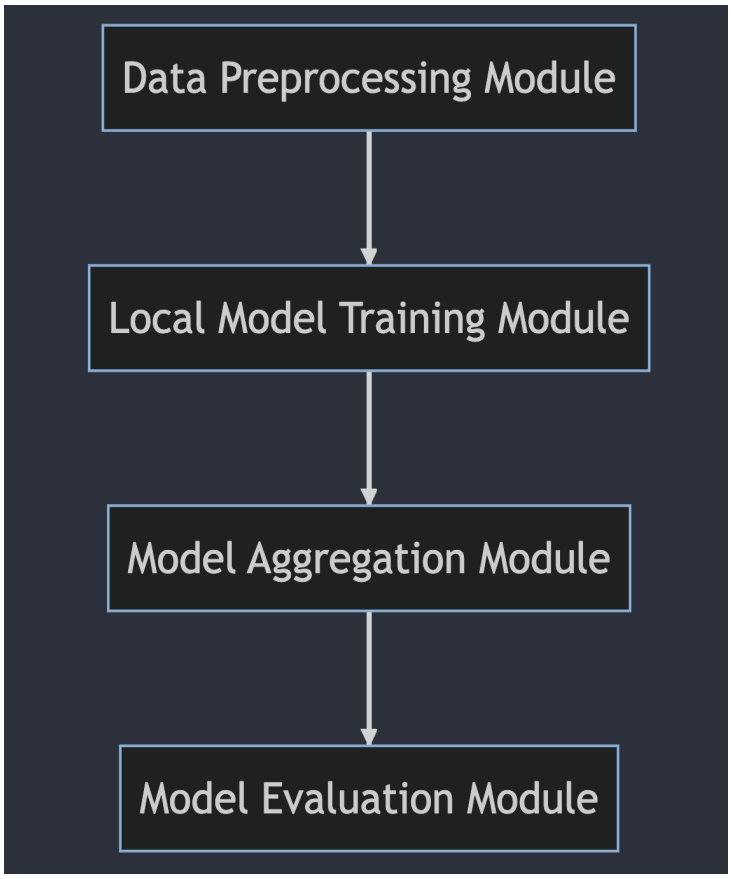

Fig 1.3: System Modules Diagram

6.3 Design Description

6.3.1 Master Class Diagram

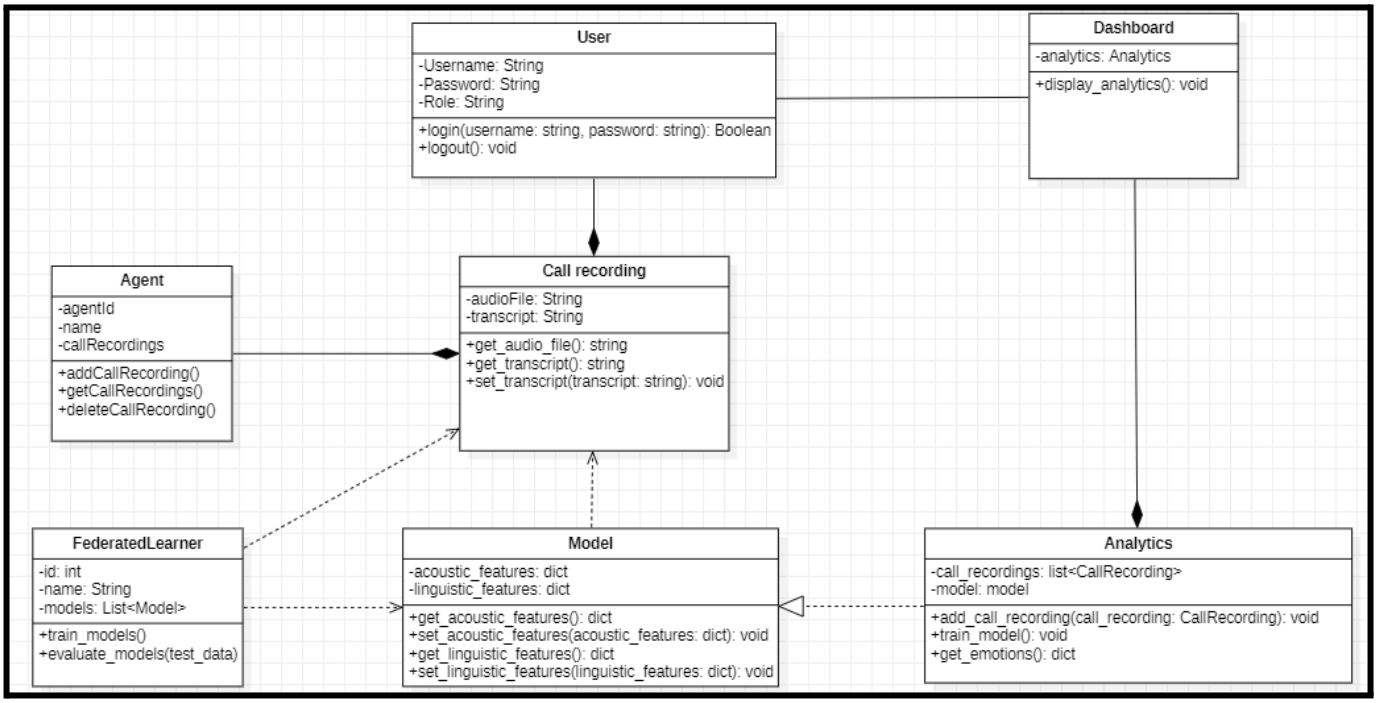

Fig 1.4: Master Class Diagram

- The class diagram consists of 7 classes: User, Dashboard, Call recording, Agent, Model, FederatedLearner, and Analytics.

- User: This class represents the users of the system. The user can be either a manager or an admin.

- Dashboard: This class represents the graphical user interface of the system. It provides a visual representation of the system's features and functionalities.

- Call Recording: This class represents the audio recordings of the calls made in the call center. It stores the audio files and provides access to them for analysis.

- Agent: This class represents the call center agents who handle the calls. It stores the agent's information, such as name and ID.

- Model: This class represents the machine learning models used for emotion recognition. It stores the models and their parameters.

- FederatedLearner: This class represents the FL process used to train the ML models. It manages the communication between the devices used for training the models and ensures that the training process is secure.

- Analytics: This class represents the analytics module of the system. It performs the emotion recognition and sentiment analysis on the audio files and generates reports for the manager and admin.

6.3.2 Use Case Diagram

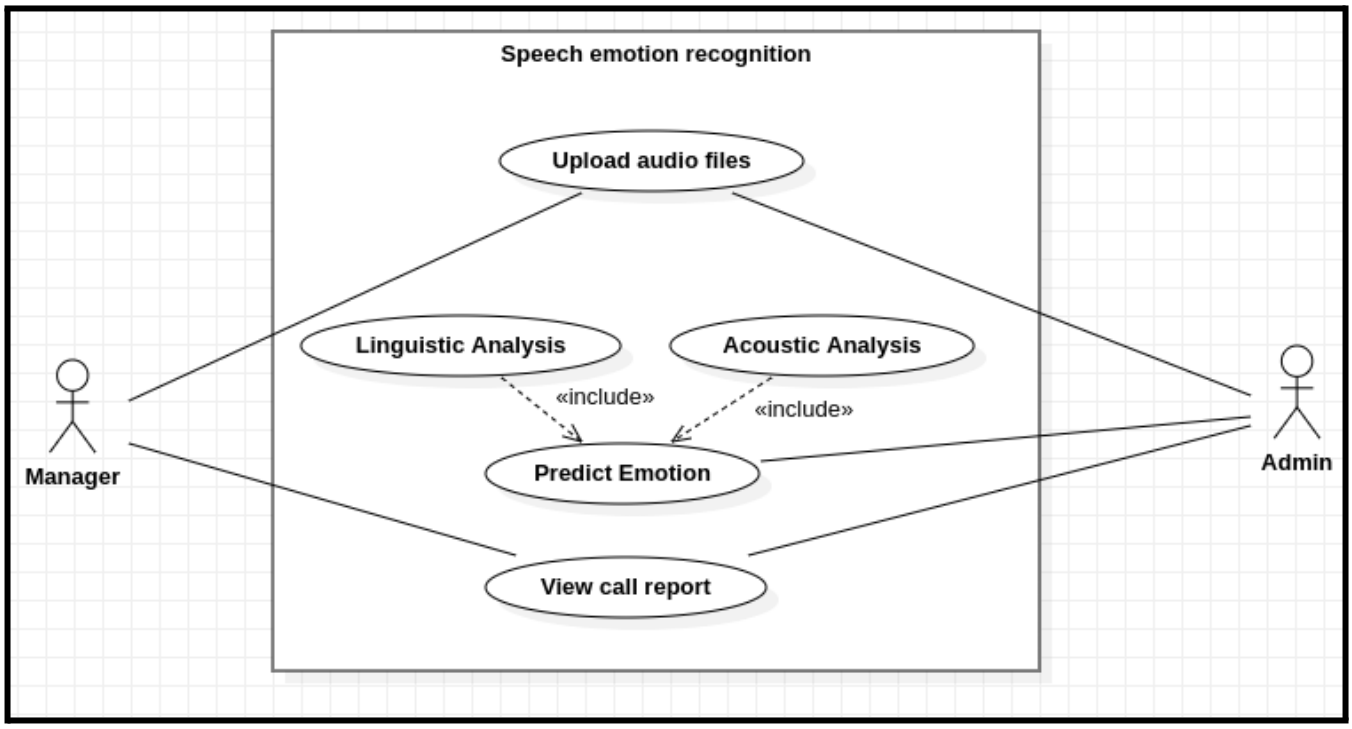

Fig 1.5: Use Case Diagram

- The use case graphic includes two actors, Manager and Admin, as well as three use cases: Upload Audio Files, View Report, and Predict Emotion. Both the Manager and Admin actors have access to the Upload Audio Files and View Report use cases, while only the Admin actor has access to the Predict Emotion use case.

- The Upload Audio Files use case allows both actors to upload audio files to the system. This use case involves selecting the audio files and uploading them to the system. Once uploaded, the files can be used for analysis.

- The View Report use case allows both actors to view reports generated by the system. This use case involves selecting the type of report and viewing the results. The types of reports that can be generated include call reports and emotion reports.

- The Predict Emotion use case is only accessible by the Admin actor. This use case involves predicting emotions in the uploaded audio files using linguistic and acoustic analysis. The analysis is performed by the system and the results are presented to the Admin actor in the form of an emotion report.

- Overall, the use case diagram shows how the Manager and Admin actors can use the system to upload audio files, view reports, and predict emotions. The system allows for easy analysis of call center data, which can help managers and administrators make informed decisions about call center operations.

6.3.3 User Interface

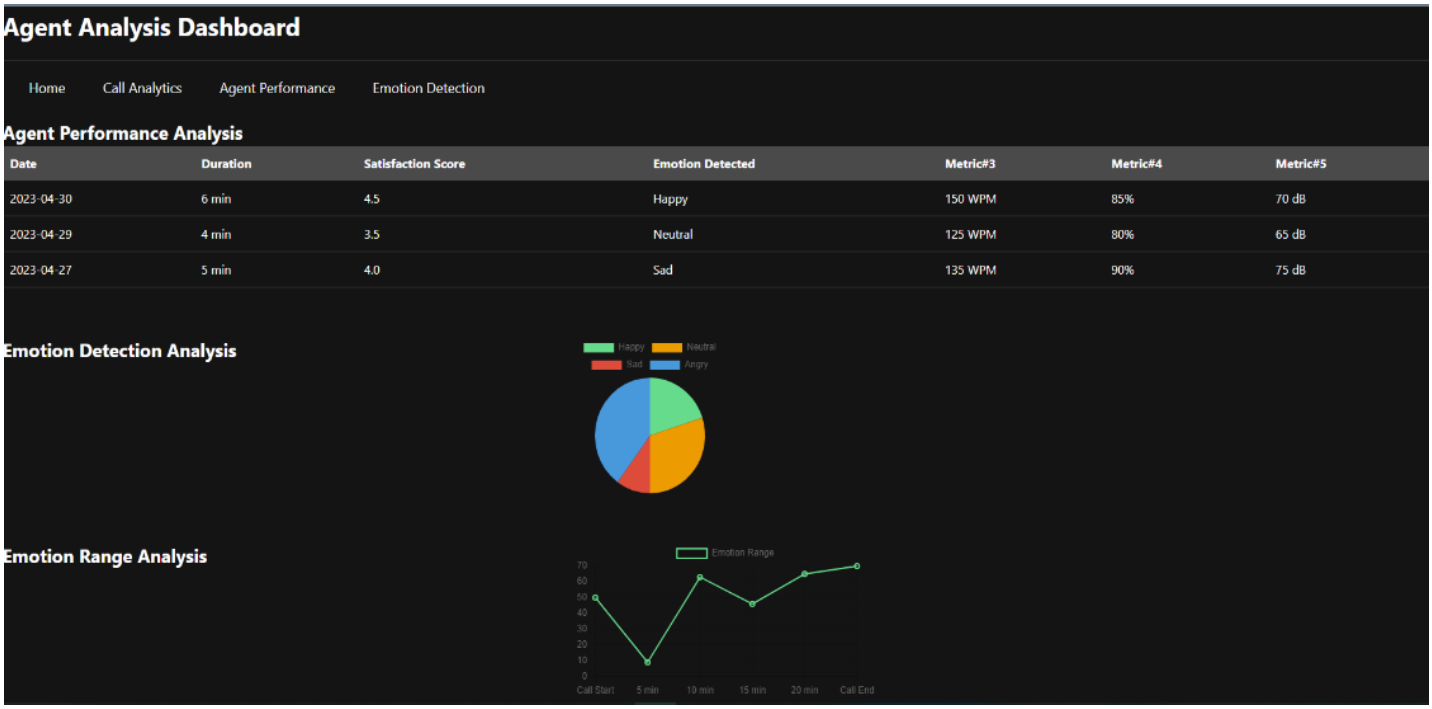

Fig 1.6 User Interface

6.3.4 Reusability Considerations

Reusability is an essential aspect of software development, and our project also considers reusability to improve the software development process's efficiency. The following are the reusability considerations planned for our project:

- Use of existing libraries and frameworks: We plan to leverage existing libraries and frameworks to implement some of the project functionalities. For example, we plan to use the TensorFlow and PyTorch libraries for developing machine learning models.

- Modular design: Our project architecture is designed to be modular, allowing for the reuse of components across different modules. This approach promotes code reusability and ensures that each module can be used independently in other projects.

- Documented code: We plan to document the code, making it easier for other developers to understand and reuse the code. The documentation will include information about how to use the code, its input and output formats, and any dependencies.

- Open-source release: We plan to release our project as open-source software. By doing so, other developers can access our code and reuse it in their projects.

Overall, we aim to make our project code as reusable as possible to promote efficiency, reduce development costs, and encourage collaboration.

CHAPTER 7

IMPLEMENTATION AND PSEUDOCODE

- **Denoising of the audio files**

```
import noisereduce as nr
import os
import soundfile as sf

folder_path = '/content/gdrive/MyDrive/Capstone_Dataset/boston'
# load the audio file into memory
for file_name in os.listdir(folder_path):
    if file_name.endswith(('.wav', '.mp3', '.flac', '.ogg')):
        file_path = os.path.join(folder_path, file_name)
        data, sample_rate = sf.read(file_path)

        # reduce the noise in the audio file
        reduced_noise = nr.reduce_noise(y= data, sr= sample_rate)

        denoised_file_path = os.path.join(folder_path+ '/denoised', 'denoised_' + file_name)
        sf.write(denoised_file_path, reduced_noise, sample_rate)
```

- **Splitting the audio file into segments**

Segments of 10 seconds for each audio file

```
from pydub import AudioSegment
import os


segment_length = 10000

if not os.path.exists("/content/gdrive/MyDrive/Capstone_Dataset/boston/denoised/audio_segments"):
    os.mkdir("/content/gdrive/MyDrive/Capstone_Dataset/boston/denoised/audio_segments")

# iterate over each file in the input directory
denoised_folder_path = "/content/gdrive/MyDrive/Capstone_Dataset/boston/denoised"
for filename in os.listdir(denoised_folder_path):
    # check if the file is an audio file (mp3, wav, etc.)
    if filename.endswith(('.wav', '.mp3', '.flac', '.ogg')):
        # load the audio file
        audio_file = AudioSegment.from_file(os.path.join("/content/gdrive/MyDrive/Capstone_Dataset/boston/denoised/", filename))

        # get the total length of the audio file in milliseconds
        audio_length = len(audio_file)

        # iterate over the audio file, segmenting it into 10-second chunks
        for i, start_time in enumerate(range(0, audio_length, segment_length)):
            # calculate the end time of the segment
            end_time = start_time + segment_length

            # extract the segment from the audio file
            segment = audio_file[start_time:end_time]

            # save the segment as a new audio file
            output_filename = os.path.join("/content/gdrive/MyDrive/Capstone_Dataset/boston/denoised/audio_segments/", f"{filename}
            segment.export(output_filename, format="mp3")
```

- **Acoustic Feature extraction using MFCC**

Features Extracted and K mean clustering to label the data

```
import librosa
import numpy as np
from sklearn.cluster import KMeans


# set the path to the audio files folder
au_folder_path = "/content/gdrive/MyDrive/Capstone_Dataset/boston/denoised/audio_segments"

# loop through each file in the folder
mfcc_features = []
for filename in os.listdir(au_folder_path):
    if filename.endswith(('.wav', '.mp3', '.flac', '.ogg')):
        # load the audio file
        audio_file, sr = librosa.load(os.path.join(au_folder_path, filename), sr=None)

        # extract the features
        mfccs = librosa.feature.mfcc(y=audio_file, sr=sr, n_mfcc=40)

        # concatenate the features into a single feature vector
        features = np.concatenate((mfccs.mean(axis=1), mfccs.var(axis=1)))

        mfcc_features.append(features)

        fe_folder_path = "/content/gdrive/MyDrive/Capstone_Dataset/boston/denoised/audio_segments/features_extracted"

        csv_path = os.path.join(fe_folder_path, f"{filename}.csv")

        # save the features to a CSV file
        #np.savetxt(f"{filename}.csv", features, delimiter=",")
        np.savetxt(csv_path, features, delimiter=",")
```

- **Clustering features for emotion labelling**

```
k = 4  # number of clusters
kmeans = KMeans(n_clusters=k, random_state=0, n_init= 10).fit(mfcc_features)

# loop through each file in the folder and print its emotion label
for i, filename in enumerate(os.listdir(au_folder_path)):
    if filename.endswith(('.wav', '.mp3', '.flac', '.ogg')):
        if len(mfcc_features) > i:

            # get the cluster assignment for the MFCC features of the audio file
            cluster = kmeans.predict([mfcc_features[i]])

            # print the emotion label based on the cluster assignment
            if cluster == 0:
                print(f"{filename} is happy")
            elif cluster == 1:
                print(f"{filename} is sad")
            elif cluster == 2:
                print(f"{filename} is neutral")
            else:
                print(f"{filename} is angry")
```

CHAPTER 8

CONCLUSION OF CAPSTONE PROJECT PHASE 1

In conclusion, Phase 1 of our capstone project concluded successfully and we made a considerable amount of progress in the direction of our objectives. The problem statement was well-defined and a thorough literature review was conducted to understand the various methods for identifying speech emotion from audio recordings.

Two distinct call centres have provided the research with the necessary dataset. To precisely describe the needs and lay the groundwork for implementation, a high-level design document and project requirements specification were drafted.

CHAPTER 9

PLAN OF WORK FOR CAPSTONE PROJECT PHASE 2

Model Development:

Model development involves selecting an appropriate machine learning algorithm, preparing and processing the audio data, training the model on the processed data, and fine-tuning the model for better performance.

Testing and Validation:

The testing step entails assessing the model's performance using a collection of data that the model has never seen before, known as the test set. The validation step entails testing the model's performance on a distinct collection of data, known as the validation set, on which the model has not been trained. The goal of testing and validation is to guarantee that the model can generalise successfully to new data while not being overfit to the training data.

Research Paper:

The creation of a novel algorithm or system for recognising emotions in speech would be described in a research report, containing details on the data utilised, the algorithm or system design, performance metrics, and the results gained. The article would normally contain a review of existing techniques' literature as well as a discussion of how the new approach compares to existing approaches. The article may also provide future research directions in the topic.

REFERENCES

[1] Płaza, Mirosław, et al. *"Emotion Recognition Method for Call/Contact Centre Systems."* Applied Sciences 12.21 (2022): 10951.

[2] Potdar, Veena & Santhosh, Lavanya & Bhatt, Supritha. (2021). "*Analysis of Vocal Pattern to Determine Emotions using Machine Learning.*"

[3] Petrushin, Valery. (2000). *"Emotion in Speech: Recognition and Application to Call Centers."* Proceedings of Artificial Neural Networks in Engineering.

[4] B. Li, D. Dimitriadis and A. Stolcke, *"Acoustic and Lexical Sentiment Analysis for Customer Service Calls," I*CASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019

[5] Daniel Rueckert, Jonathan Passerat-Palmbach *"Federated learning : Opportunities and challenges"* arXiv:2101.05428v1 [cs.LG] 14 Jan 2021.

[6] Aashish Agarwal, Torsten Zesch (2019) *"German End-to-end Speech Recognition based on DeepSpeech"*

[7] Yangyang Xia, Li-Wei Chen, Alexander Rudnicky, Richard M. Stern, INTERSPEECH 2021: *"Temporal Context in Speech Emotion Recognition"*

[8] J. Ancilin, A. Milton, *"Improved speech emotion recognition with Mel frequency magnitude coefficient"*

[9] Souraya Ezzat, Neamat El Gayar, and Moustafa M. Ghanem, *"Sentiment Analysis of Call Centre Audio Conversations using Text Classification",* International Journal of Computer Information Systems and Industrial Management Applications. ISSN 2150-7988 Volume 4 (2012) pp. 619 - 627

[10] Rashid Jahangir, Ying Wah Teh, Faiqa Hanif & Ghulam Mujtaba, *"Deep learning*

approaches for speech emotion recognition: state of the art and research challenges", Springer Science+Business Media, LLC, part of Springer Nature 2021, corrected publication 2021

[11] I. Shafran and M. Mohri, "A comparison of classifiers for detecting emotion from speech," Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., Philadelphia, PA, USA, 2005, pp. I/341-I/344 Vol. 1, doi: 10.1109/ICASSP.2005.1415120.

[12] S. Yoon, S. Byun and K. Jung, *"Multimodal Speech Emotion Recognition Using Audio and Text,"* 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 2018, pp. 112-118, doi: 10.1109/SLT.2018.8639583.

[13] Valery Petrushin, *"Emotion in Speech: Recognition and Application to Call Centers",* Article · January 2000.

[14] Blumentals, Eduards, and Askars Salimbajevs. *"Emotion recognition in real-world support call center data for latvian language."* CEUR Workshop Proceedings. Vol. 3124. 2022.

[15] Li Lia,b, Yuxi Fana, Mike Tsec, Kuo-Yi Lina,b,*"A review of Applications in Federated Learning"* Elsevier - Computers & Industrial Engineering Volume 149, November 2020, 106854.

[16] Abbaschian BJ, Sierra-Sosa D, Elmaghraby A. *" Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models. Sensors."* 2021; 21(4):1249.

[17] Byun S-W, Kim J-H, Lee S-P. *"Multi-Modal Emotion Recognition Using Speech Features and Text-Embedding."* Applied Sciences. 2021; 11(17):7967.

[18] S. Lugović, I. Dunđer and M. Horvat, *"Techniques and Applications of Emotion Recognition in Speech"* ,
MIPRO 2016, May 30 - June 3, 2016, Opatija, Croatia
[19] Anna Bogdanova; Nii Attoh-Okine, F.ASCE; and Tetsuya Sakurai, *"End-to-end speech emotion recognition: challenges of real-life emergency call centers data recordings"* ASCE-ASME Journal of Risk and Uncertainty in Engineering
Systems, Part A: Civil Engineering Vol. 6, Issue 3 (September 2020)

APPENDIX A DEFINITIONS, ACRONYMS AND ABBREVIATIONS

- **DEFINITIONS**

1. Federated Learning - Machine learning technique for training a global shared model using decentralized data sources on edge nodes while preserving data privacy.
2. Acoustic features - They are characteristics of sound that can be measured objectively, such as frequency, intensity, and duration. In speech, acoustic features are used to describe the physical properties of speech sounds, such as pitch, loudness, and duration.
3. Linguistic features - They refer to aspects of language that can be analyzed and measured, such as syntax, semantics, and discourse. In the context of speech analysis, linguistic features may include measures of vocabulary richness, grammatical complexity, and discourse coherence.
4. Feature extraction - the process of selecting and transforming relevant information from raw data into a set of features that can be used in machine learning models.
5. Deep learning - a subfield of machine learning that uses neural networks with multiple layers to learn representations of data.
6. Emotion recognition - The process of identifying and categorizing human emotions from various sources, such as text, speech, and facial expressions.
7. Federated learning - a distributed machine learning approach that allows multiple parties to collaboratively train a model while keeping their data private.

8. Deep emotion model - It is a type of machine learning model that is designed to recognize and classify emotions from various sources such as speech, text, and video. It is a type of deep learning model that typically involves the use of RNNs or CNNs to learn patterns in the input data that correspond to specific emotions.
9. Decentralized architecture - In federated learning, decentralised architecture refers to a model where the learning process is distributed across multiple devices or nodes, rather than being centrally controlled. In this architecture, each device or node has a copy of the machine learning model and contributes to the training process by computing updates on the model based on its local data. These updates are then aggregated by a central server to produce a global model that incorporates the knowledge learned from all the devices.

ACRONYMS / ABBREVIATIONS

1. FL - Federated Learning
2. SER - Speech Emotion Recognition
3. GDPR - General Data Protection Regulation
4. CCPA - Central Consumer Protection Authority
5. MFCC - Mel Frequency Cepstral Coefficients
6. SVM - Support Vector Machine
7. kNN - k Nearest Neighbours
8. API - Application Programming Interface
9. BERT - Bidirectional Encoder Representations from Transformers
10. LSTM - Long Short-Term Memory
11. LPC - Linear Predictive Coding
12. GUI - Graphical User Interface
13. ML - Machine Learning
14. NLP - Natural Language Processing
15. AI - Artificial Intelligence
16. NN - Neural Network
17. CNN - Convolutional Neural Networks
18. RNN - Recurrent Neural Networks
19. TCP - Transmission Control Protocol
20. UDP - User Datagram Protocol
21. AWS - Amazon Web Services
22. QA - Quality Assurance
23. IEMOCAP - Interactive Emotional Dyadic Motion Capture

24. EPST - Emotional Prosody Speech and Transcripts
25. MFMC - Mel Frequency Magnitude Coefficients