AI-Based Real-time Voice to Text and Emotion system to enhance communication for Deaf individuals

Submitted By
Farah Ulfat Athoi
ID: 221-35-1040
Department of Software Engineering
Daffodil International University

Supervised By
Tapushe Rabaya Toma
Assistant Professior
Department of Software Engineering
Daffodil International University

Bachelor of Science

DAFFODIL INTERNATIONAL UNIVERSITY

DAFFODIL INTERNATIONAL UNIVERSITY

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name	:	Farah Ulfat Athoi
Date of Birth	:	31-12-2001
Title	:	AI-Based Real-time Voice to Text and Emotion system to enhance communication for Deaf individuals
Academic Session	:	

I declare that this thesis is classified as:

☐	CONFIDENTIAL	(Contains confidential information under the Official Secret Act 1997)*
☐	RESTRICTED	(Contains restricted information as specified by the organization where research was done)*
☐	OPEN ACCESS	I agree that my thesis to be published as online open access (Full Text)

I acknowledge that Daffodil International University reserves the following rights:

1. The Thesis is the Property of Daffodil International University.
2. The Library of Daffodil International University has the right to make copies of the thesis for the purpose of research only.
3. The Library of Daffodil International University has the right to make copies of the thesis for academic exchange.

Certified by:

_____
(Student's Signature)

221-35-1040
Student ID
Date:21-12-25

_____
(Supervisor's Signature)

Tapushe Rabaya Toma
Name of Supervisor
Date: 21-12-25

NOTE : * If the thesis is CONFIDENTIAL or RESTRICTED, please attach a thesis declaration letter.

APPROVAL

This thesis titled on "**AI-Based Real-time Voice to Text recognition and Emotion detection to enhance communication for Deaf individuals**", submitted by **Farah Ulfat Athoi (ID:221-35-1040)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Dr. S M Hasan Mahmud
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Chairman

A.H.M Shahariar Parvez
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1

Tapushe Rabaya Toma
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Khalid Been md. Badruzzaman Biplob
Lecturer (Senior Scale)
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 3

Dr. Md Sazzadur Rahman
Professor
Institute of Information technology
Jahangirnagar University, Bangladesh

External Examiner

SUPERVISOR’s DECLARATION

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Science.

(Supervisor’s Signature)

Full Name : Tapushe Rabaya Toma

Position : Assistant Professior

Date : 21-12-25

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Daffodil International University or any other institution.

(Student's Signature)

Full Name : Farah Ulfat Athoi

ID Number : 221-35-1040

Date : 27-11-2025

AI-Based Real-time Voice to Text and Emotion system to enhance communication for Deaf individuals

Farah Ulfat Athoi
ID:221-35-1040

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Bachelor of Science

Department of Software Engineering (Major in Data Science)

DAFFODIL INTERNATIONAL UNIVERSITY

November 25

ACKNOWLEDGEMENTS

I would like to express my sensor gratitude to my supervising teacher. Her advice, support, and useful suggestions have helped me during this entire study process. Her assistance has enabled me to successfully finish the assignment and improve my thoughts.

I am also thankful of the department and the relevant instructors who gave me the academic assistance I needed and gave a setting that was favorable for my study.

I am deeply grateful to my family as well. Their constant encouragement, tolerance, and support have given me the willpower to finish this study.

Finally, I would like to acknowledge all participants who contributed their speech samples for this study. Without their involvement, this research would not have been possible.

Dedication

This study is dedicated to my beloved family. I am deeply thankful to my family members who encouraged me and motivated me during the period.

Their constant love, support, and real encouragement have been a foundation of all of my successes. Throughout this entire research journey, their constant trust, mental stamina, and dedication to support me have always given me the courage to keep going.

ABSTRACT

In this research, an integrated Automatic Speech Recognition (ASR) and Speech Emotion Recognition (SER) system for the Bangla language has been developed. This system aims to make communication easier for hearing-impaired users. Although ASR and SER technologies are rapidly advancing worldwide, there is a lack of reliable datasets, emotion recognition models, and real-time subtitle systems for the Bangla language. To address this issue, I collected a total of 1,400 audio samples—600 Normal, 400 Angry, and 400 Sad. To clean the audio, I applied Voice Activity Detection (VAD), noise reduction, and trimming.

The ASR component uses the Whisper model. Initially, the Word Error Rate (WER) was 58.8% and the Character Error Rate (CER) was 28.2%. After cleaning and preprocessing the data, both WER and CER decreased significantly, improving the system's transcription quality.

For the SER component, LSTM, Random Forest, and SVC three models were tested. The SVC model showed the highest accuracy at 96.09%. These results indicate that SVC provides comparatively the most stable and effective performance in emotion recognition for the Bangla language.

This research proposes the design and development of an integrated Automatic Speech Recognition (ASR) and Speech Emotion Recognition (SER) system for the Bengali language. The planned system will convert the speaker's voice into written form in real-time.

Additionally, the system will identify the speaker's emotion into three categories—Normal, Angry, or Sad. Although, in practical implementation, noisy environments or multilingual support may pose challenges. However, through initial research and the application of data pre-processing techniques, there is significant potential to enhance the system's performance and accuracy.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

1.1 Background

The ability to convert human speech into text is currently regarded as one of the most important technologies in human-computer interaction. It plays a vital role in various applications such as voice assistant transcription software and customer service bots. Besides enhancing linguistic inclusion in education and health care sectors. Graham and Roll (2024) demonstrated a whisper model. Their research shows that it delivers high accuracy both native and non-native pronunciations. Similarly, Ferraro et al. (2023) compared the performance of various commercial and open-source ASR services and showed that the whisper model is more accurate and effective for multilingual uses. Additionally, Dash et al. (2025) presented a generative AI-based multilingual ASR model in their research.

Though the ASR system visualizes the speech to the deaf, it cannot recognize and show the emotion. Emotions like normal, angry and sad are important to identify the main meaning of the speech. Kuhn et al. (2025) described in their research that those who can't hear most often lose many relevant emotional cues by only seeing the transcript, which makes it difficult to understand the real purpose of the conversation. Samaradivakara et al. (2025), in their research, mentioned that merely displaying text of an utterance cannot provide the necessary contextual relevance and emotional understanding for hearing impaired. Additionally, ZainEldin et al. (2024) discussed the potential of a base communication technology for deaf individuals, demonstrating that emotion and expression recognition makes human communication more effective. Therefore, only the ASR system cannot provide proper communication support to the deaf.

Integrating speech emotion recognition (SER) with an ASR system can create a more relevant experience for hearing-impaired individuals. For example, by adding indicators as ‘angry’ and ’sad’ to the text and detecting emotions from the speaker’s voice users can understand the emotional context of the conversation. Zaineldin et al. (2024) showed in their research that such intelligent systems enhance communication abilities of these individuals and increase their engagement in the society. Furthermore, Devi (2017)

mentioned in their research that speech-based emotion recognition systems for hearing-impaired individuals not only enhance communication but also help in appropriately responding to the emotions of the listeners. Another study by Schuhmann et al. (2025) presented a multimodal emotion recognition system. That uses both vocal features and linguistic characteristics to accurately understand the speaker's expressions.

The inception of Acer technology was in the 1950s with a system called 'Audrey', which can recognize a limited number of words (Davis et al. 1952). Subsequently, hidden Markov model (HMM) and Gaussian mixture model (GMM)-based systems became popular (Rabiner,1989; Jelinek,1997). Since 2010, deep learning-based architecture, specially recurrent neural networks and transformer-based models like OpenAI's whisper, have revolutionized ASR. Dash et al. (2025) showed that modern generative AI-based multi-language models can now effectively transcribe even mixed speech.

Although research on emotion detection or speech recognition started in the 2000s, in recent years it has made significant progress. Initially classification was based on some features like MFCC, pitch, intensity, etc. later the integration of CNN, LSTM and maltimodal approaches greatly enhanced the effectiveness of SER (Pan & Wu, 2023; Fang et al., 2024). Schuhmann et al. (2025) presented a fine-grained and expert validity benchmark called 'EmoNet-Voice' in recent research, which is considered a standard for the development of future SER systems.

1.2 Problem Statements

Although automatic speech recognition (ASR) (Juang et al., 2005) and speech emotion recognition (SER) (Scherer, 2003). Technologies have advanced significantly in recent years; research in the Bengali language remains relatively limited. Most modern models, such as Whisper (Graham & Roll, 2024) or generative AI-based multilingual ASR (Dash et al., 2025), are trained on English or global language datasets. For this, do not accurately reflect the phonetic diversity and dialectical features of the Bengali language. Islam et al. (2021) have shown that the complex structure of Bengali vowels and consonants reduces the accuracy of existing ASR models. Also variations in word formation create this problem. On the other hand, Samin et al. (2024) conducted research on regional

pronunciation and standardization in the “BanglaDialecto" project, but it is not related to emotion recognition. At the international level, Schuhmann et al. (2025) created a benchmark called “EmoNet-Voice” but this type of model is still unavailable for the Bangla language. As a result, the lack of an integrated ASR and emotion detection system in Bangla creates a significant communication barrier for Deaf users, which serves as the main motivation for this research.

1.3 Research Statement

The main goal of this research is to develop an integrated automatic speech recognition and speech emotion recognition system for the Bangla language, which will provide a realistic and relevant communication experience for hearing-impaired individuals. For this purpose, firstly, some Bangla speech was collected, and the samples were categorized into three primary emotion classes: angry, sad, and neutral/normal. During the preprocessing stage voice activity detection (VAD), noise reduction, and audio trimming were applied to make the transcription more accurate. In the ASR system, the Whisper model was used, which demonstrated high-quality performance in Graham and Roll’s (2024) research. For emotion detection, a combination of acoustic features and the learning architecture (LSTM) has been employed, which was used in the research of Schuhmann et al. (2025) and Fang et al. (2024). The system classified teaching into three emotional categories.

1.4 Research Objectives

The main purpose of this research is to develop a real time user friendly and emotion never speech to subtitle system for Bangla language who is will make communication more meaningful and accessible for hearing-impaired users specially the research objects are as follows:

1.Collection of data: there is limited data on Bangla language. The first target is to collect speeches in Bangla language and create relevant data set by classifying them.

2. Real time speech to text:

Develop an AI based system that will convert or transcribe voice into real-time text for the deaf users.

3. Speech to emotion: To create a speech emotion recognition (SER) module that identifies the speaker's emotions in real-time and at appropriate emotional level to the subtitles.

4.Integrating the system: To create a real time integrated Bangla language system by combining ASR and SER models which will present both speech transcription in textual form and detect speaker's emotion with indicators (Angry, Normal, Sad)-particularly serving as a positive tool for communication for hearing-impaired users.

5. Multi model de planning architecture: Design a hybrid deep learning model to simultaneously perform space transcription and emotion classification by analysing the acoustic features of audio.

6.Model Evaluation: To describe the model's performance by analysing their accuracy, error rate etc. and select the best model by comparing the performance.

1.5 Research Scopes

- This study only analyzes the natural pronunciations of the Bangla language. The performance of this model has not been tested on multilingual speech.
- Since the collected speech dataset was completed within a limited time and budget, it was not possible to cover the Bangla pronunciations from all regions.
- The application of the system in a real-time environment was considered only in a controlled environment of this system. Complex real-world environments such as busy markets or noisy places were not considered.
- In the emotion recognition system, only three categories have been considered, such as sadness, anger and normal. The analysis of more subtle emotions is limited within the scope of this research.

- Due to the necessary equipment, computing resources and time required for model training and evolution, the maximum diverse data set could not be used. That may affect the generalizability of the results.

1.6 Thesis Organization

This research paper is organized into six major chapters.

1. The first chapter represents the research background problem statement, research objective, scope, and overall structure.
2. The second chapter discusses the relevant literature review. This includes automatic speech recognition, speech emotion recognition, and comparative analysis. Here, all the works of ASR & SER in English, Bangla, and other languages have been discussed.
3. The third chapter describes the methodological framework of the research, such as data collection, preprocessing, transcription using the Whisper model, and the steps of emotion detection model development. Also describes some classification model.
4. The fourth chapter demonstrated the examination of model training and analyzed the results where performance matrices and evaluation comparisons were included. This chapter describes the thesis result, and insights.
5. The fifth and last chapter concluded with the decision for future research. This arrangement gives the research paper a logical and coherent structure, which will help the reader understand the entire process step by step.
6. The Sixth chapter describes the environment setup.

CHAPTER 2 LITERATURE REVIEW

2.1 Introduction

To maintain the proper structure and conceptual consistency of the thesis, it is important to explain each piece of research in the context of previous studies. This chapter explains the technologies used for real-time speech-to-subtitle conversion. Additionally, methods of emotion recognition are also reviewed. These help to facilitate communication for hearing-impaired users. This covers progress in automatic speech recognition real-time subtitle generation speech-emotion detection and assistive technologies for the deaf and hard-of-hearing community The goal is to pinpoint the strengths, limitations, and research gaps that drive the proposed system.

2.2 Review of Transcription

English Language

In just ten years, the development of STT for English has advanced rapidly. Advanced by moving from outdated mathematical tools to more advanced deep learning and multi-sound techniques. Mel Frequency Cepstral and hidden Markov were used in early research to increase rate. Mel-Frequency Cepstral Coefficients (MFCC) have been widely used by researchers to extract useful features from speech signals in automated speech recognition (ASR) systems. These features have proven effective in early ASR applications. When used with classifiers like SVM, they increase recognition accuracy and make processing more efficient (Garud et al., 2018). Since 2016, deep neural networks (DNN), like CNNs and RNNs, have received the most attention in ASR research. According to Wu (2016), deep models offer additional advantages due to their ability to model time-dependent and context-dependent dependencies. When trained on large and complex datasets, they reduce the error rate and improve recognition performance.

The rapid progress of digital communication technology has significantly accelerated the development of real-time transcription-capable Speech-to-Text (STT) systems (Emerj,

2018). These systems are now being used more reliably across different languages and communication contexts.

Benchmark analysis shows that the performance of STT systems varies depending on the type of dataset and the service used. Ferraro et al. (2023) tested multiple STT services using different audio inputs. They demonstrated that paid systems generally perform better than open-source tools. However, there is still no single STT tool that is equally effective for all types of speech. Reddy et al. (2023) discussed the shift from rule-based to deep learning-based STT and TTS, with a focus on CNNs, RNNs, and transformer architectures. They identified context, noisy conditions, and accent nonuniformity as some of the main issues.

Large models, such as OpenAI's Whisper, have demonstrated outstanding generalization across multilingual and multitask scenarios. One version of Whisper showed near-human-level performance and strong zero-shot transfer capabilities after training on 680,000 hours of internet audio (Radford et al., 2022). Also, research demonstrated enhanced speech recognition for children who are not native English speakers (Jain et al., 2024) and fine-tuning benefits for domain-specific applications such as medical transcription (Roushan et al., 2024). Koenecke et al. (2024) found that the transcripts generated by Whisper showed evidence of a 1% hallucination rate and a higher prevalence among speakers who had speech impairments.

Furthermore, multimodal strategies that combine speech and textual elements have emerged. Fang et al. (2024) enhanced speech-to-speech emotion recognition using GPT-3 by integrating acoustic signals and semantic text features. This evidenced by an overall weighted accuracy of 79.62% across emotion categories. Additionally, Graham and Roll (2024) examined Whisper's performance with various native and non-native English accents. It shows the difference in accuracy caused by speech style and accent.

So far, most STT research has been limited to the English language. However, recent work has shown that efforts to improve STT performance in low-resource and linguistically diverse languages are rapidly increasing by using pre-trained models, fine-tuning strategies, and multilingual benchmarks.

Speech-to-Text (STT) in Other Languages

The advancements in English STT technology make the problem of low-resource languages seem trivial. Still, the work of Pratama et al. (2024) focuses on the contribution of Whisper to low-resource languages in the studies on Javanese speech recognition. By adopting PEFT, they reduced the trainable parameters to under 1% of the Whisper large-v2 model. Model achieved an astonishing reduction in WER from 89.40% all the way to 13.77. This work shows that, for low-resource languages, large performance improvements can be obtained for low computational cost. Liu et al. (2024) and others further confirmed this, pointing out that the appropriate use of fine-tuning methods considerably increases ASR accuracy on many low-resource languages.

The ability to adapt fine-tuning techniques to various situations demonstrates their versatility. Liu et al. (2024) noted the linguistic features of the proposed adaptation strategy while evaluating five techniques for optimizing Whisper models in seven low-resource scenarios. Those affected the varying degree of overall success. In correlation, Polat et al. (2024) on the Turkish ASR was able to use low-rank adaptation (LoRA) to achieve WER (word error rate) improvements of as much as 52.38%. It indicates the simultaneous usability of complexity reduction with morphological adaptation and varying reduction strategies on ASR. This illustrates the greater need for tailored techniques for such languages, which lack well-structured annotated corpora.

Whisper fine-tuning has also worked for specialized speech areas and other low-resource languages. Starzew et al. submitted their paper on modelling dysarthric speech recognition. The adaptation of the model came to a WER of 73.44% (WER meaning word error rate of their model). Wagner et al. (2025) added to this strategy and achieved about a 23% reduction in WER thanks to AdaLoRA and the creation of synthetic dysarthric speech data. Through the data collection techniques, fine-tuning can enhance the ASR performance for atypical speech conditions. This is important, as ASR can be used for accessibility applications.

The use of Whisper has been further expanded through implementation and evaluation in various communication contexts. Zhang et al. (2025) designed an L-oRA-INT8 Whisper

framework. This provides improved performance. Speech recognition is impossible on edge devices. The researchers quantified the Whisper-ti-ny model to attain a 49.5% character error rate (CER). The researchers showed that only using 1.6% of the original weights could reduce the CER to 11.1% and achieve performance near-full fine-tuning with a very low cost. The real-time factor (RTF) of 0.06 on an NVIDIA A10 GPU and 0.20 on a MacBook Pro M1 Max CPU showed that efficient modes could be implemented for mobile and embedded platforms.

In addition, transfer learning techniques have been successfully applied to low-resource languages with linguistically similar relatives. Mengk et al. (2025) fine-tuned learning of the Khalkha Mongolian language using transcription-assisted translation. This included converting Chakhar Mongolian's Uighur script data into the Cyrillic script. Wav2V-BERT, Co-former-Large, and Whisper-large-v3 approach showed an approximate absolute improvement of 19.26% and a maximum relative WER improvement of 32.50%. This demonstrates that in low-resource settings, strategically using data from related languages can effectively reduce the need for annotated corpus collection.

Research on Turkish ASR also highlights the challenges of low-data and morphologically complex languages. It also highlights the potential solutions**.** LoRA was applied by Plot et al. (2024) to significantly lower WER, while Gokcime et al. (2024) found that medium-sized Whisper models performed poorly, and Tasaar et al. (2024) used self-supervised learning to address data scarcity, achieving a noteworthy WER of 0.23. Research shows that fine-tuning, adaptation, and strategic data selection are needed for STT systems of low-resource languages. These methods are applied in a manner that reflects the target language's linguistic features.

Bangla Language

Bangla is one of the low-resource languages and offers unique pros and cons for the STT field. The low-resource nature of Bangla challenges the high-performance ASR systems. Because of the lack of annotated datasets and diverse dialects, the earlier systems had high Word Error Rates (WER). The issues highlighted in literature (Jimerson et al. 2018) were that innovative error-coping strategies were applied to ASR Bangla. Using field-

based experiments and well-planned datasets, they applied domain-specific modifications and reduced the WER by 2%–11%. This makes an important contribution to Bangla ASR research. This has encouraged fine-tuning, the creation of large datasets, and the use of transfer learning approaches.

Later studies began using pre-trained models and language modeling to enhance Bangla speech recognition. In their research, Rakib et al. (2022) applied the wav2vec 2.0 self-supervised framework and later fine-tuned it on the Bengali Common Voice dataset. They fine-tuned the model and applied n-gram language models as a post-processing step by employing n-gram language models. This allowed the model to track some contextual dependencies of the language. Backed with the most extensive open-source Bengali speech corpus available, the research provided remarkable improvements in transcription accuracy. This confirmed that combining pre-trained representations with language models for automatic voice recognition in low-resource environments works well.

To tackle the challenges arising from sparsely labeled data, Nandi et al. (2023) proposed a new approach. This method makes it possible to create a large and domain-independent Bangla ASR dataset using pseudo-labeling. This dataset contained over 20,000 hours of labeled, domain-spanning speech. This included news, telephony conversations, everyday dialogues, various dialects, and recordings made in background noise. The researchers trained a conformer-based ASR system on this large dataset and, to their satisfaction, it performed well across various domains. Furthermore, the researchers constructed a domain-agnostic, human-annotated test set for evaluation purposes. That covered all necessary aspects and domains for evaluation. The work showed that pseudo-labeling and large-scale self-supervised learning are viable candidates for solving the problem of data deficits in Bangla ASR. for Bangla ASR system development, presented a practical, scalable solution. Hossain et al. (2024) collected speech corpora and completed fine-tuning the wav2vec 2.0 models with OpenSLR datasets. Hossain reported a WER of 11.27% and a CER of 6.03%. Improved up to 11% over baseline models. The study demonstrated that transfer learning with a combination of datasets that contains distinct linguistic and phonetic features of Bangla enables accurate transcription even in lower-resource scenarios. The only computational support provided was cluster computing with NVIDIA CUDA-compatible GPUs. It suggests that adjusting state-of-

the-art ASR systems entails considerable support. This also suggests a significant amount of computational power, which can drive such large-scale projects.

The overall progress of Bangla ASR research shows a clear development trend. Initially the main goal of research was to reduce dataset-dependent errors and tackle word- or pronunciation-based challenges (Jimerson et al., 2018). Later research improved word- and sentence-level coherence using pre-trained large models and supervised learning methods (Rakib et al. 2022). After that, the diversity and quantity of data were further enhanced using pseudo-labeling and large-scale data augmentation techniques (Nandini et al., 2023). Recent works have achieved very good results using transfer learning and datasets collected from multiple sources and significantly reducing WER and CER (Hossain et al., 2024). These developments are consistent with global ASR trends. Here, self-supervised learning, pseudo-labeling, and fine-tuning strategies are recognized as effective solutions for low-resource languages.

Conclusion

Deep learning and transformer-based architectures have replaced traditional statistical models in the development of Speech-to-Text (STT) systems over time. Large datasets and extensive research have made high-accuracy transcription in the English language possible. However, despite the lack of data, transfer learning, LoRA-based fine-tuning, and synthetic data augmentation have produced positive results in low-resource languages like Javanese, Turkish, and Khalkha Mongolian. Research on Bangla STT shows a similar pattern. Models are not trained with large amounts of data. Early on, accuracy was low due to the use of small datasets and simple acoustic models (Jimerson et al., 2018). WER and CER have been considerably decreased in recent studies that have used wav2vec 2.0, Whisper, pseudo-labeling, and multi-source fine-tuning (Rakib et al., 2022; Nandi et al., 2023; Hossain et al., 2024). There is limited work on the Whisper model.

Overall, the literature in English and low-resource languages including Bangla emphasizes the value of varied datasets. Also, meticulous fine-tuning, and flexible model design. In the future, the work will focus on multimodal integration, zero-shot and few-shot learning, and deployment on low-resource devices. These developments will

collectively make STT solutions more accessible and inclusive for various language-based communities.

2.3 Review of Emotion

English Language

The early research on speech emotion recognition primarily focused on handcrafted acoustic features (such as MFCC, pitch, and energy). Also focused on conventional machine learning models (like SVM, HMM, etc.). However, recent research has rapidly moved towards deep learning, multitask learning, and graph-based representation learning. These aim to better capture the subtle emotional cues embedded in speech. Cai et al. (2021) were the first to propose a multitask learning framework that trains space-to-text and emotion classification tasks. This achieved state of the earth results in the IMOCAP dataset.

Subsequently, Yue et al. (2022) enhanced performance by 5-7% by incorporating emotion intensity prediction using Wave2Vec 2.0. Kang et al. (2022) demonstrated that multi-task learning allows models to generalize well across different datasets. Inspired by the concept of emotional perception in the human brain, Liu et al. (2023) proposed an emotion-perception-based model. That showed significant improvements in both UA and WA on the IEMOCAP dataset. At the same time, the use of CNN-LSTM-based hybrid models has also become popular.

The CLDNN model by Pan & Wu (2023) demonstrates good performance on RAVDESS, EMO-DB, and IEMOCAP. Since 2023, Graph Neural Network has emerged as important in SER research. Li et al. (2023) and Pentari et al. (2024) showed that graph-based speech representations can be more effective for emotion recognition. Ferreira et al. (2025) and Schuhmann et al. (2025) showed that multi-modal fusion, naturalistic speech, and benchmark standardization are aspects of the future of SER.

Overall, it can be seen that English SER research has progressed:

Handcrafted features -> Deep models -> Multi-tasking and brain-inspired learning -> Graph and multimodal representation.

SER in Other Languages

In the other languages except English, SER research is slowly increasing. Particularly addressing the challenges of dataset scarcity and dialect variation. In the case of Arabic dialects, Messaoudi et al. (2022) created the first TuniSER corpus and demonstrated good results by finding multimodal Wav2Vec. In Mandarin, Kang et al. (2022)'s speechEQ framework learns both emotion categories and intensity, performing consistently well across various datasheets. Samaradivakara et al. (2025) developed an emotion award actioning system that helps hearing-impaired students understand emotions. This is not directly SER, but it is significant from the perspective of 'emotional accessibility.' In Spanish, s e r, Mares et al. (2025) show that using pre-trained speech models significantly improves performance. In Moroccan Arabic SER, Ouali & EI Garouani (2025) demonstrates that handcrafted features can still be effective in low-resource environments.

In summary, SER research in other languages is progressing through transfer learning, multilingual pretraining, and the creation of new speech corpora.

SER in Bangla

Bangla SER is still relatively limited because standard emotional speech datasets are scarce. Nevertheless, recent research has shown significant progress. Ayon et al. (2022) achieved 84.37% accuracy using an ensemble best model. This indicates that a good ensemble strategy can be effective in identifying emotions in Bangla speech. Bangla SER (Das et al., 2022) and BANSPEmo (Hussain et al., 2023) play a major role in constructing Bangla emotional speech datasets. Also provide balanced emotion categories and expert annotations.

Significant progress is observed in deep learning-based models in the research by Alam et al. (2024). Here, using data augmentation, approximately 99% accuracy is achieved on the SUBESCO dataset. Also, good generalization is seen in cross-lingual settings. Shakil et al. (2025) demonstrate that using a multi-stream deep symbol with both spectral and prosodic features together makes Bangla SER more stable.

Overall, Bangla SER is improving in this flow:

Machine learning -> Deep learning -> Dataset expansion -> Cross-lingual validation

2.4 ASR & SER together

In the past, SER mostly depended on auditory elements including prosody, pitch, and spectral qualities. Recent studies have shown that linguistic information collected from ASR transcriptions also aids in the detection of subtle emotions. So, efforts are currently being made to improve emotion recognition by combining ASR with SER.

Li, Bell, and Lai (2021) showed in early studies that SER performance can be enhanced by combining acoustic embeddings with ASR output. A hierarchical co-attention fusion approach that integrates transcription and ASR hidden-layer characteristics was presented in their study. The ASR and SER modules are trained jointly. This joint training approach achieved a weighted accuracy of 63.4 on the IEMOCAP dataset. That is nearly equivalent to the highest results obtained with joint training using ground-truth transcriptions. This shows that linguistic information derived from ASR is effective. However, it was observed that handling speech errors in emotion expression is challenging, as they affect transcription accuracy. To make this issue worse, Li et al. (2023) studied how emotions influence the quality of transcription of ASR at the word level.

In their research, four modern ASR systems (Kaldi, wav2vec2, Conformer, and Whisper) were used to analyze various emotion-based speech corpora (IEMOCAP, MOSI, and MELD). They showed that when speakers speak with very high emotional intensity, the word error rate (WER) of ASR increases significantly. The study found a direct relationship between ASR stability ratio and the robustness of SER. Additionally SER models trained on ASR transcripts showed decreased performance according to the amount of WER.

Inspired by enhancing joint learning even more, some of the previous studies such as Kyung et al. (2023) developed Global Style Tokens (GST) for a Conformer-based ASR and SER integrated system. The GST module generated emotional style embeddings directly from audio. That enabled both tasks to exploit common affective cues. Their jointly trained model on IEMOCAP obtained 15.8% WER for ASR and 75.1 weighted accuracy for SER. Which shows that emotion-aware latent style representations can be

utilized to effectively improve both the transcription and the emotion classification tasks together.

There are still two main obstacles to merely combining acoustic and ASR-based textual modalities:

(1) ASR errors add noisy or invalid showcasing semantic data, and

(2) Using text and audio information together is difficult because their features are located in different spaces. To solve this problem, He and colleagues (2024) proposed a multimodal speech emotion recognition (SER) model called MF-AED-AEC. This model first cleans ASR transcriptions using ASR error detection (AED) and error correction (AEC), and then integrates the audio and text information. Additionally, a modality-shared fusion module is used to combine the feature spaces of text and audio. As a result, this model outperformed previous baselines by 4.1 points in SER accuracy, demonstrating the importance of correctly fixing ASR errors for a robust multimodal SER.

In addition, Li, Bell, and Lai (2024) conducted an extensive study. Here 6 SER fusion strategies and 11 ASR systems were tested on the IEMOCAP, MOSI, and MSP-Podcast datasets. They demonstrated that the performance of text-based SER alone declines rapidly as WER increases, but bimodal fusion models remain relatively stable, especially when they include ASR error-robust gating and correction mechanisms. Their unified framework showed improved results in both WER and SER accuracy compared to a single ASR system, as it integrates modality-gated fusion and ASR correction. This further strongly supports the view that ASR should not be considered merely as an input, but as an adaptive component within the emotional speech processing pipeline.

CHAPTER 3 METHODOLOGY

3.1 Introduction

In this chapter we will see the working process of the proposed real-time speech-to-subtitle and emotion recognition system. To understand each step of the research accurately all process such as- data collection, pre-processing, feature extraction, model training, and evaluation are presented sequentially here.

Due to the characteristics of the Bangla language, pronunciation variations, and limited data, special procedures had to be followed to develop this system. Therefore, cleaning audio data, reducing noise, and selecting relevant features are important parts of this chapter.

Also, the chapter provides a detailed explanation of speech-to-text conversion using Whisper. Explained emotion recognition methods using three models: SVC, LSTM, and Random Forest. The technologies, algorithms, and evaluation metrics used for the implementation of each step are also included in this chapter.

Overall, this chapter gives a clear understanding of how the entire system works and the scientific methods through which the results were achieved.

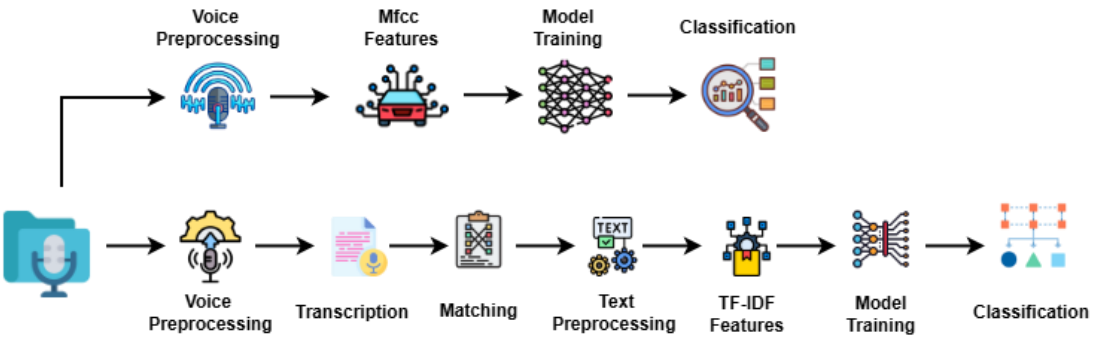

Fig 3.1 Methodology

3.1 Data collection

In the first phase, Bengali voice samples were collected from Bangla-speaking speakers of different ages and genders. Every audio clip contains various emotions such as anger, sadness and neutral expression. Also, the texts corresponding to the collected voice clips were saved. The voice files are saved in .wav format and text files are saved in .csv format and each clip was spoken by a single speaker.

This data collection process ensures that the voice samples are of high quality and sufficiently diverse for research purposes.

3.2 Transcription

This is the process of voice to text generation. After preprocessing data was transcribed using the Whisper model.

3.2.1 Voice Preprocessing

The collected audio data often contains various types of background sounds. Therefore noise reduction has been applied. Also used voice activity detection to isolate the space segment of the speaker by trimming.

i) Noise reduction:

To reduce noise, ‘Spectral Gating Algorithm’ has been used which has been carried out using python’s ‘noisereduce’ library.

How it works:

Creates a noise profile. The parts of the audio where there is no speech are used to understand the pattern of the background noise. then the audio signal is transformed into the frequency domain to show which frequencies contain more noise.

The noise profile is subtracted from the frequency bands of the original audio. As a result, the speaker's voice remains in text and background noise is reduced and then transforms back the signal from the frequency domain to the time domain.

Why selected this: It is highly effective and practical, capable of reducing different types of background noise. The original audio is not damaged.

Alternatives:

- Wiener Filtering
- MMSE Estimation
- Deep Learning-based denoising models

ii) Voice activity detection (VAD)

- VAD is used to separate the speaker’s speech from silent parts.
- The audio clip is streamed so that only the speaker’s part remains.
- This enhances the performance of the transcription model.

The purpose of this phase is to clean audio and improve the accuracy of the transcription and emotion detection.

3.2.2 Transcription Process

This is the process by which audios transcribed using whisper model Bangla ASR. Whisper is a strong deep learning automatic speech recognition model that uses encoder-decoder transformer architecture.

i)Firstly, the input audio signal is divided into small segments to create a mel-spectrogram. This mel-spectrogram represents the time frequency characteristics of the audio signal, which align with human story perception. then the encoder part generates a feature vector from these spectrograms. It extracts the meaningful features or patterns of the audio signal. The encoder consists of multiple transformer layers and each layer employs a self-attention mechanism. This helps the model understand which parts of the audio are related to which words.

ii)Decoder:

This part uses the feature vectors received from the encoder to generate the corresponding text. decoder is an auto regressive transformer, that means it predicts one token at a time and uses the previous output as input when generating the next token. Like the entire speech is gradually converted into text.

Whisper Model Architecture:

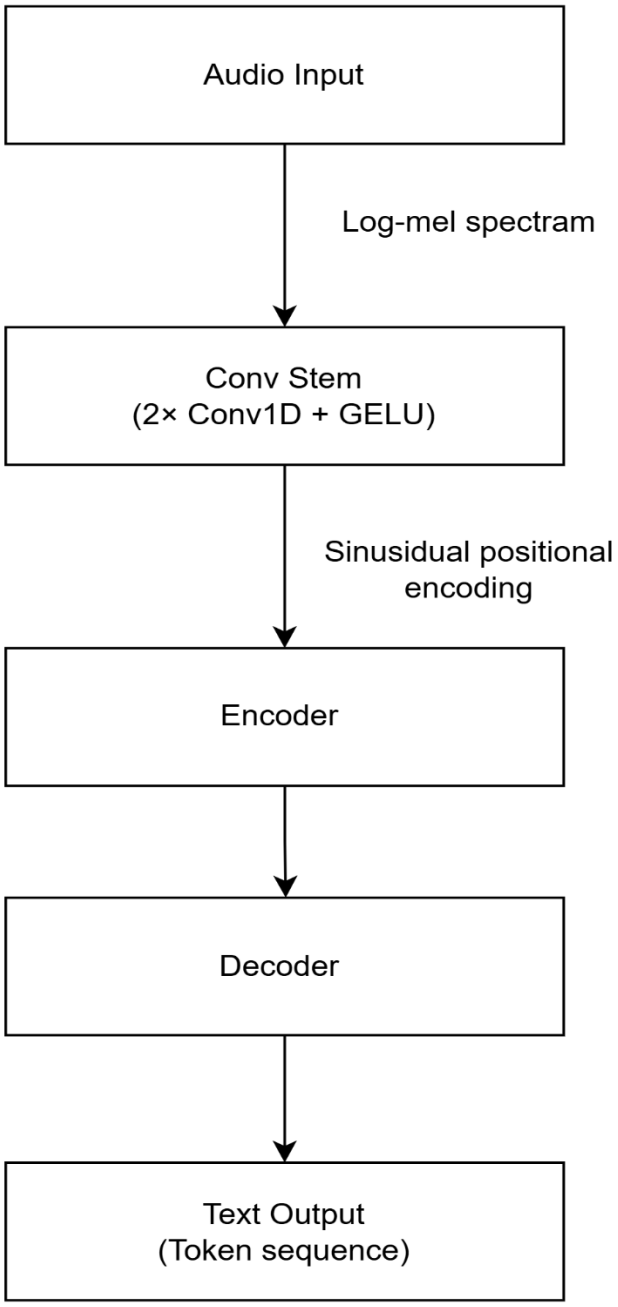

Fig 3.2 Whisper Model

key features of the Whisper model:

- Multilingual automatic space recognition
- Encoder-Decoder transformer-based architecture
- Pretrained on a massive dataset (680000 hours of audio)
- High accuracy even in noising environments
- Capable of both translating and identifying languages from audio

In this research Bangla audio files were converted into text using Whisper Bangla ASR model. After performing noise reduction and voice activity detection (VAD), each refined audio is sent to the Whisper model. The model analyzes the frequency features of the audio, identifies words over time and provides output in the text form.

3.2.3 Matching

After transcription, audio files are completed through Whisper models and transcribed text files are generated for each class. It is extremely important to accurately match these transcription files with the original files because incorrect matching can reduce the accuracy.

File matching: The text data files of the main audio dataset were saved. The transcribed files were subsequently accurately matched with the original text data files so that the text content and features of each audio could be analyzed together. After matching, created corrected csv files.

3.2.4 Text preprocessing

Formatting text into a clear, consistent, and machine-readable form is important in emotion recognition (Li et al., 2022; Latif et al., 2021). For this purpose, various natural language processing techniques have been used.

i) Text normalization

Normalization is the process of converting text into a standard form such as converting to lowercase, removing punctuation and eliminating unnecessary spaces or symbols.

This makes the model understand ‘Happy’ and ‘happy’ as the same word and vocabulary becomes smaller (Latif et al., 2021; Schuller et al., 2022).

ii) Tokenization

Tokenization is the process of breaking text into smaller units (Kumar et al., 2020; Bird et al., 2021).

The transcription obtained from Whisper is usually in the form of subword tokens, which is important for father processing (Radford et al., 2023).

The model understands the meaning of each word and can extract context-based features after tokenization (Zhang & Schuller, 2022; Li et al., 2023).

iii) Stop word removal

Stop words are common words such as "the," "is," and "and" that do not play a significant role in the analysis. Removing them increases focus on the meaningful purpose of the text.

This reduces noise and decreases computation time.

3.2.5 Text vectorization

TF-IDF (Term Frequency-Inverse Document Frequency) and word embedding are used to convert words into a machine-readable form (Mikolov et al. 2013). TF-IDF identifies the important words in the text, and embedding captures the meaning of relationships in the text. This allows the model to understand the semantic meaning and relevance of each word.

TF-IDF gives more weight to words in a text that are more informative. Embeddings represent the contextual meaning and semantic relationships of words. When combined, these methods enable the model to effectively understand the relevance and significance

of each word in the dataset (Mamanning et al., 2008; Jurafsky & Martin, 2023). Word-level TF-IDF is configured with the parameters analyzer='word' and ngram_range=(1,3). This helps the model recognize features based on short word sequences (from unigrams to trigrams) and single words. On the other hand, analyzer='char' and ngram_range=(2,5) are used in character-level TF-IDF. This enables the recognition of subword structures and subtle character patterns. It is also capable of recognizing spelling variations.

3.2.6 label lock

In this section firstly, from the sentence level data set a label table is created and using train test split it is split in and 80:20 ratio at the label level in a stratified manner. The test labels were saved in the test_label.csv file to make them reusable (level lock). then the original sentence-level csv was loaded and the sentence-level train, test and split were made based on the test_label.csv. this preference label-level leakage.

Then duplicate rows were removed from the train set based on level class option. very short sentences were discarded to ensure meaningful examples for training. To address class in balance the normal class was taken as the target class and the minority classes were resampled to achieve balance. Finally, it was checked together there was any level leakage between train and test sets.

This method ensures that the test set is genuinely unseen (level locked) and that the model evolution can be performed reliably while maintaining class balance in the training set.

3.2.7 Model selection

In this study various types of machine learning classification algorithms were tested to detect emotion from the text (Mishra et al. 2021; Zhang & Schuller et al. 2022). Initially three different algorithms were used, such as logistic regression, random forest, and support vector machine (SVM) (Kumar et al. 2020; Latif et al. 2021). The performance of each model was compared, and the most accurate model was ultimately selected.

3.2.7.1 Logistic regression (Hybrid TF-IDF)

Logistic regression is a linear classification algorithm that provides probability through the sigmoid function. Here a hybrid TF-IDF (word-level n-grams + character-level n-grams) has been used for text input and a Logistic regression pipeline has been trained on it.

working method:

1.Each sentence is converted into a numerical vector using the hybrid method.

- Word-level TF-IDF: pictures are extracted from the sentence word sequences
- Character-level TF-IDF: analysis character waste in gram to help capture finar-level context.

2. These two factors are combined using future Union and trained using Logistic regression.

3. The model separates train test data using level-lock splitting.

4. To handle class imbalance minority classes have been balanced through over sampling.

Advantages:

- Logistic regression models are computationally efficient and relatively quick to train.
- The weight of each feature can be easily analyzed, which helps to understand the model's decision.
- Combining word and character-based TF-IDF provides good performance even on data with noise or sampling variations.

Limitations:

- Logistic regression can only create a linear decision boundary and is unable to grass complex patterns.
- The model sometimes over fit to the training data, specially, when using high dimensional TF-IDF features.

3.2.7.2 Random Forest

In this method, text-based sentiment classification combines word-based and character-based TF-IDF features. Truncated SVD is applied to reduce them from high-dimensional sparse vectors to low-dimensional dense space (Halko et al. 2011). Finally, classification is done using a random forest classifier. The training and evolution States follow the level of principle.

working process:

1. Test levels are taken from the created test_labels csv. Best on these labels sentence level data is separated into train and test set.

2. In the training set duplicate rows are removed based on level class and text and very short sentences are discarded.

3. Hybrid TF-IDF featuring:

- Word-level TF-IDF : tokenizer = whitespace tokenization, ngram_range=(1,3), max_features=2000, min_df=2, max_df=0.9 (analyzer='word') |
- Char-level TF-IDF:character n-grams ngram_range=(2,5), max_features=1000, (analyzer='char') |

4. The TF-IDF space is transformed into a 300 dimensional representation using TruncatedSVD(n_components=300) - reducing memory/computation and allowing subsequent RF to run faster.

5. Trained models like tfidf_word,tfidf_char,svd,rf,etc. have been saved in pickle format ; the confusion matrix and classification report have been exported as csv. Model parameters and test scores have been locked in manifest.jason.

Advantages:

- (word + char) TF-IDF together were used. By using both word-level and character level micro-patterns can be captured (spelling variations, sub-word cues).
- Using SVD, the high dimensional sparse TF-IDF space can be reduced to lower-dimensions. As a result, Random Forest trains and predicts faster.
- Balancing at the training level awareness for minority classes is increased.
- Random forest can capture non-linear patterns and it is relatively resistant to overfitting.

Limitations:

- The TF-IDF size was tuned. The number of SVD components and the n_estimators and depth of RF is necessary.
- Due to the symbol and SVD, understanding the results is somewhat difficult.

3.2.7.3 SVC

Linear SVC is a linear support vector machine classifier that separates classes by finding an optimal hyperplane in the feature space. Here, a hybrid TF-IDF approach is used to represent textual data, and the linear SVC model is trained on it.

Working Method:

1. Each sentence is converted into numbers using the hybrid TF-IDF method:

- Word level TF-IDF: Analysis n-grams from the worst in a sentence to retain meaningful context.
- Character level TF-IDF: It identifies a small and large parts of word. Those are prefixes, suffixes, or spelling variations.

2. Word and character TF-IDF features are combined using FeatureUnion.

3. Training is done using the LinearSVC model. This tries to create the maximum margin between classes.

4. Train and test data are separated using a lockback split.

5. During training, smaller classes are oversampled proportionally to handle class imbalance.

Advantages:

- LinearSVC is computationally efficient and scales well with high dimensional data.
- With hybrid TF-IDF features, it also delivers strong performance on noisy or unbalanced data.
- Analyzing feature weights makes it easy to understand the model's decision.

Limitations:

- LinearSVC can only create linear decision boundaries so it may fail to capture complex non-linear patterns.

- Using high dimensional TF-IDF features can lead to the possibility of overfitting.

3.2.8 Transcription Testing

This section presents the process of evaluating the transcription performance of the speech-to-text system. A separate testing phase was conducted to examine how accurately the Whisper model can convert Bangla audio into text.

Here, the model's performance was mainly measured using two metrics. Those are Word Error Rate (WER) and Character Error Rate (CER). Initially, raw audio was used to collect preliminary results. After that, an enhanced analysis of the results was conducted by applying dataset-based matching, text cleaning, and correction processes.

3.2.8.1 Normalization

To ensure accurate error detection, both the Reference (ground truth) and the Hypothesis (Whisper-generated transcription) texts have been standardized. This removes unnecessary differences and allows errors to be correctly identified.

The following were included in normalization:

- Putting all text in lowercase
- Combining many white spaces into one
- Removing or standardizing punctuation
- Normalizing mixed-Bangla and English Unicode characters

Phonetic variations, standard number mapping, and other features are optional. As a result, the hypothesis and reference texts have the same format.

3.2.8.2 Tokenization

After normalization, each text was tokenized at two levels to determine assessment metrics:

Table 3.1 Token of Words

Token Type	Purpose	Metrix
Word Token	To detect word-level errors	WER
Character token	Capture pronunciation	CER

Word tokenization was space-based, while character tokenization was Unicode-level.

3.2.8.3 Hypotheses

To break down performance by emotion, we sorted both the hypothesis and reference texts into classes like angry, normal, and sad.

Here's how we built the hypothesis dataset:

First, we checked each audio file's label. Then, the hypothesis text was cleaned up and matched to the right audio. Next, we pulled both the reference and hypothesis texts together under one index. After that, we organized everything into a tidy data frame. Now ready for analysis.

After this step the dataset was clearly labelled as class and fully lined up for the next step.

3.2.8.4 Evaluation

Word error rate (WER)and character error rate (CER) matrices are used to evaluate transcription accuracy. Using these matrices the effectiveness of transcription was analyzed in every class. Accuracy, Precision, Recall, F1-score are used to evaluate emotion classification (Erickson et al., 2021).

i) Accuracy:

Accuracy is the proportion of correct predictions made by model. It shows the overall correctness of the model.

Formula:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{TP + TN + FP + FN}$$

TP (True Positive): Cases that are correctly identified as positive
TN (True Negative): Cases that are correctly identified as negative
FP (False positive): Cases that are incorrectly identified as positive.
FN (False Negative): Cases that are incorrectly identified as negative.

2. Precision: It shows how many of the predictions made as positive are actually true. It determines the impact of false positives.

Formula:

$$\text{Precision} = \frac{TP}{\text{TP+FP}}$$

3.Recall: It shows how many of the real positive examples are correctly found. It tells how well a model finds correct data.

Formula:

$$\text{Recall} = \frac{\text{TP}}{TP + FN}$$

4. F1-score: It is the combination of precision and recall. It summarizes the model's performance into a single matrix especially when there is class in balance.

Formula:

$$\text{F1\_score} = 2 \times \frac{\text{Precision} + \text{Recall}}{\text{Precision} \times \text{Recall}}$$

5. Word Error rate (WER): Word error rate is mainly used for transcription models. It measures the difference between predicate words and actual words.

Formula:

$$\text{WER} = \frac{\text{S+D+I}}{N}$$

S (Substitutions): Incorrectly replace words.
D (Deletions): Omitted words.
I (Insertion): Extra words
N=Total number of actual words
Purpose: The lower the WER the more accurate the transcription.

6. Character Error Rate (CER): It is the character level version of WER for transcription. It counts how many characters are incorrect, missing or extra.

Formula:

$$\text{CER} = \frac{\text{S} + \text{D} + \text{I}}{N}$$

S (Substitutions): Incorrectly replace characters.

D (Deletions): Omitted characters.

I (Insertion): Extra characters.

N=Total number of characters

Purpose: CER is particularly effective for detecting small words or spelling errors.

Summary: All of these matrices were used to compare the performance of each model. The model that showed the highest accuracy and F1-score for classification was chosen as the final model of emotion classification. The lowest WER/CER describes the highest accuracy of the transcription model.

3.3 Emotion Detection

Here emotion detection methodology has been shown.

3.3.1 Preprocessing

Followed the same process of noise reduction as transcription. Each .wav file was loaded using ‘librosa’ at its sampling rate and converted to mono. The spectral gating algorithm has been performed using Python’s ‘noisereduce’ library; noise reduction was applied to minimize background noise while preserving the speaker’s voice characteristics. This process enhances the quality of the voices.

Working process: In this method, firstly, the parts of the audio where there is no space are identified, and then a noise profile of the background noise is created using these segments. Afterwards, the inter-audio signal is transformed into the frequency domain to determine which frequencies have more noise. This noise profile is subtracted from the frequency bands of the original audio. It leaves the speaker’s voice unchanged.

Why selected this: It is highly effective and practical, capable of reducing different types of background noise. The original audio is not damaged.

Alternatives:

- Wiener Filtering
- MMSE Estimation
- Deep Learning-based denoising models

3.3.2. Audio Feature Extraction

From processed audio samples collected videos' acoustic features. Every file was loaded using the library at a 48-kilohertz sample rate. Then computed various features like MFCC, chroma, spectral, zero crossing rate, RMS energy, pitch, tempo, etc. (Tzanetakis & Cook 2002; Eyben et al. 2010). For each sample, all extracted features were combined into a fixed-length feature vector and stored with its corresponding emotion label in the future emotion_feature.csv file. In total, more than 45 features were extracted, which were later used to train the machine learning model for emotion classification.

i) MFCC (Mel-Frequency Cepstral Coefficients): Thirteen coefficients related to the spectral structure of the audio and human auditorium sensitivity have been collected. MFCC is effective in identifying differences in a speaker's voice timbre and vocal quality.

ii) Chroma Features: The twelve Chroma features capture the harmonic and tunnel characteristics of the audio. It helps to detect subtle changes in the speaker's voice tone and emotion.

iii) Spectral features: Three spectral features have been extracted to determine the frequency distribution of the audio.

- Spectral Centroid: This indicates the frequency region where the energy is most concentrated.
- Spectral Bandwidth: This measures the spread of frequencies from the main frequency to higher and lower frequencies.
- Spectral Rolloff: This indicates the lower limit of frequency in which a certain percentage of the total energy is contained.

iv) Zero Crossing Rate: Determines how many times the signal is the audio a frame process zero. It is an indicator for measuring the percussiveness and sharpness of the audio.

v) Energy (Root Mean Squared): The mean and standard deviation of the energy value of each frame are calculated, which indicate the speaker's vocal strains or intensity.

vi) Prosodic Features: Pitch, jitter, tempo, onset rate, and other prosodic features.

- pitch: the fundamental frequency of the speaker's voice has been calculated. That indicates how high or low the speaker is speaking.
- Jitter: Measures the irregularity in successive pitch changes, which is associated with stress or emotional changes.

- Tempo: Refers to the speed of speech or rhythmic pattern.
- Onset Rate: The rate of voice onset per second has been determined, which indicates speech rate or emotion-rate changes.

After extracting the above features from each audio file, they were safe in the CSV file. This file was used in the next step for training and evaluating the classification model.

Extracted Features

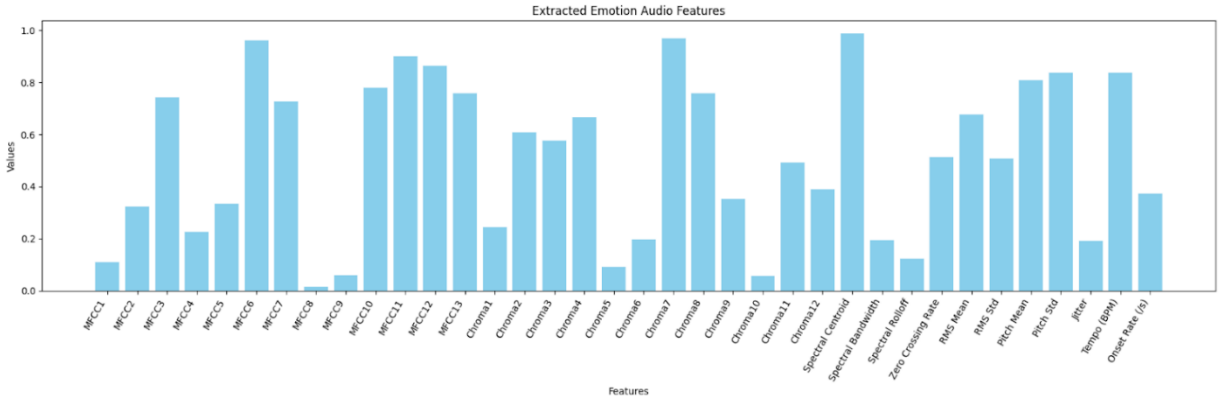

Fig 3.3 Extracted audio features

The figure shows a bar chart representing all types of audio features collected from a speech sample. This feature vector includes 13 Mel-Frequency Cepstral Coefficients (MFCC), 12 Chroma energy components, and several other spectral, prosodic, and temporal features. These are commonly used in Speech Emotion Recognition (SER).

The MFCC values (MFCC1–MFCC13) show significant variation in each coefficient. This indicates fluctuations in the short-term spectral envelope of the speech signal. Some mid-level MFCCs (such as MFCC6–MFCC11) exhibit relatively higher values. These values reflect the presence of formant-related energy and are generally associated with vocal intensity and emotional arousal.

Chroma features (Chroma1–Chroma12) represent the energy distribution across the 12 semitones. They also exhibit moderate fluctuations. These variations suggest that the harmonic content of speech changes with emotion.

In addition to MFCC and Chroma features, the figure includes several other high-level features. Those are spectral centroid, spectral bandwidth, spectral rolloff, and zero-crossing rate. These spectral features indicate the timbre, or tonal quality of the sound. Moreover, RMS energy (mean and standard deviation), pitch (mean and standard deviation), and jitter prosodic measures are also presented. They provide important information about vocal stress, sustain, and intensity.

Finally, two temporal features are shown, such as tempo (beats per minute) and onset rate. These reflect the rhythmic dynamics of speech. The figure shows the wide range of values for these features. This indicates that speech contains multi-dimensional signals for expressing emotions. These include spectral, prosodic, harmonic, and temporal information. Such an extensive feature set helps in detecting subtle differences in emotion classification.

3.3.3 Encoding and Feature Scaling

After feature extraction is completed, each audio sample is associated with an emotion level in the data set. Since these labels are in text form, a machine learning model cannot process them directly. Therefore, the level-including method is used to make the data suitable for model training.

In this step, each emotion is converted into a numerical value using Python's sklearn.preprocessing.LebelEncorder() class.

Angry ->0
Normal->1
Sad->2
This process transforms categorical data into a numerical format and involves the model learning emotion classification. After level encoding is completed, the encoded data set is saved in a new CSV file.

The audio features are on different scales. For example, MFCC versus pitch means a big difference. So, all features have been brought to the same scale using StandardScaler.

3.3.4 Train Test Split

The main data was loaded from a CSV file, which contains features and levels for each sample. After loading the data set, the total number of records was verified using df.shape. First the data frame was checked for any possible unique sample identifier columns such as audio_path, filename, etc. If present, reuse the test split.

When previously saved test indexes or sample IDs were found, they were used to reuse the previous dataset. When retraining or evaluating on the same dataset, it ensures consistency in results.

If there are no previously saved indexes, a new stratified 80:20 split has to be created. It ensures that the proportion of each label in the train and test sets is approximately the same as in the original dataset.

Lastly, distinct train and test datasets were produced:

Table 3.2 Train & Test split

Train	Test
80%	20%

3.3.5 Model Selection

For emotion classification, several machine learning classifiers were tested. Such as Support Vector Machine, LSTM, and Random Forest.

The model that achieved the highest accuracy was selected as the final model. During the training phase, the features were selected using standard scalar. The data set was split into training and testing subsets using the train-test split function.

3.3.6 Model evaluation

Various matrices such as accuracy, precision, recall, and F1-score were used. These evaluated the model's performance. Additionally, a confusion matrix was analyzed to

observe the classification performance for each emotion category. Graphs were also used to present the results more clearly.

3.4 Summary

This chapter presents the complete methodological framework of the study. Initially audio compositions were collected, processed, and transcribed. After that, feature extraction and machine learning classification techniques were applied to identify the speaker's emotion. The proposed system is expected to play a significant role in facilitating real-time communication for individuals with hearing impairment.

3.5 Environment Setup

In this chapter the software tools, programming environment, and hardware setup used for system development. Setting up the right environment is essential for testing, training models, and operating the entire system.

3.5.1 Python

Python was the main programming language used in this study. Python was chosen because it provides a large number of libraries for machine learning, deep learning, and signal processing. Libraries including NumPy, Pandas, Scikit-learn, Librosa, TensorFlow, Matplotlib, and the Whisper API were used for data processing, feature extraction, and model development. Python's great performance and easy syntax allowed for a quicker and more efficient implementation of the research.

3.5.2 Jupyter Notebook

Jupyter Notebook was used for model development and testing. This tool is very useful for generating visualizations, viewing outcomes step-by-step, and interactive coding. The jupyter environment was used for data analysis, graph charting, spectrogram display, and transcription comparison.

Hardware Implementation

This research was conducted on standard consumer-level hardware. The following hardware setup was used:

- Processor: Intel Core i5 (11th Gen)
- RAM: 8GB or 16GB
- Storage: SSD-based drive

This setup was enough for real-time system testing, model training, and processing for this research. In the future, the research can easily be reproduced using the same setup.

CHAPTER 4 RESULT ANALYSIS

In this chapter, we will analyze Whisper and some classification models' performance. Transcription and emotion detection metrics are different.

4.1 Transcription result analysis

Here**,** we'll analyze the results related to speech transcription and text classification. First, the quality of the transcribed speech has been evaluated using WER (Word Error Rate), CER (Character Error Rate), and SER (Sentence Error Rate). Then, the performance of the corrected and uncorrected outputs has been analyzed class-wise. This analysis primarily shows how accurately the speech-to-text conversion has been performed and how reliable it is as input for the emotion recognition model.

Transcription-testing result analysis

In this study, Word Error Rate (WER), Character Error Rate (CER), and Sentence Error Rate (SER) were measured to evaluate the quality of transcribed speech. The transcription evaluation was conducted in two ways—first using the directly transcribed output and then using the class-wise corrected output.

4..1.1. Overall Transcription Result

The total results for the directly transcribed output:

Table 4.3 Overall Transcription Result

WER	CER
0.5879 (~58.8%)	0.2819 (~28.2%)

N_utts: 1399, **N_words:** 12521, **N_chars:** 68456

These values indicate that nearly half of the words were incorrectly identified in the direct transliteration process, and almost every sentence contained at least one word error. In particular, the SER being around 95% shows that sentence-level errors are very high.

4.1.2. Class-wise Transcription Metrics

From the direct transcription analysis by class, it is observed that the word error rate and sentence-level errors are distributed differently across the classes.

Angry:

Table 4.4 Angry class transcription result

WER	CER
0.6359 (63.6%)	0.2945 (29.45%)

In this class, WER is approximately 63.6%, SER around 99.3%. That is, roughly two-thirds of the words in Angry utterances were not correctly recognized. Nearly every sentence contained at least one error. This indicates that the high-intensity and emotional tone caused difficulties in word and sentence recognition during the transcription process.

Normal:

Table 4.5 Normal class transcription result

WER	CER
0.3936 (39.3%)	0.1302 (13.02%)

In the Normal class, WER is about 39.3%, CER around 13%, which is comparatively lower. This shows that for neutral tones, both word errors and sentence-level errors are fewer, making transcription relatively stable.

Sad:

Table 4.6 Sad class transcription result

WER	CER
0.7233 (72.3%)	0.4297 (42.97%)

In the Sad class, WER is approximately 72.3%, SER about 98.8%, which is the highest. This indicates that the soft, low-intensity voice led to the most errors in word and sentence recognition.

Summary: In terms of spelling and sentence-level errors, the Normal class is the most stable. Angry is moderate, and Sad is the most challenging. This proves that the quality of transcription is highly dependent on the type of emotion.

4.1.3. Class-wise Transcription Metrics (Corrected Output)

After correction, the word error rate (WER) and character error rate (CER) in the transcription results have decreased significantly, reflecting the importance of human review or post-processing.

Angry:

Table 4.7 Angry class transcription result (corrected output)

WER	CER
0.0695 (6.95%)	**0.0419 (**4.19%**)**

WER 6.95%, CER 4.19%. This indicates that the errors were primarily minor spelling mistakes. After correcting them, the system is showing almost completely accurate results.

Normal:

Table 4.8 Normal class transcription result (corrected output)

WER	CER
0.0988 (9.88%)	0.0406 **(**4.06%**)**

WER 9.88% and CER 4.06% are relatively stable. Corrections have ensured high accuracy.

Sad:

Table 4.9 Sad class transcription result (corrected output)

WER	CER
0.1338 (13.38%)	0.1296 (12.96%)

WER 13.38% and CER 12.96%, still somewhat high. This indicates that due to the soft and weak vowels of the Bengali language, some words and letters were difficult to identify correctly. However, after correction, it has improved significantly.

Summary: After correction, WER and CER for all categories have decreased significantly. This has improved the quality of the transcription data. Frequent word errors in the Angry and Normal categories have been largely eliminated. The quality has

also improved in the Sad category. However, due to the soft tone in the Sad class, some errors remain. This ensures more reliable input for training and performance of emotion recognition models.

Transcribe classification result analysis:

In this study, three supervised machine learning algorithms such as Logistic Regression, Random Forest Classifier, and Support Vector Machine (SVM) were used to detect emotions (Normal, Sad, Angry) from transcribed speech text. All models were trained and tested on the same dataset after preprocessing and feature scaling. Performance evaluation was conducted using Accuracy, Precision, Recall, F1-score, and confusion matrix analysis, which provide insights into class-wise prediction effectiveness.

4.1.4 Logistic Regression

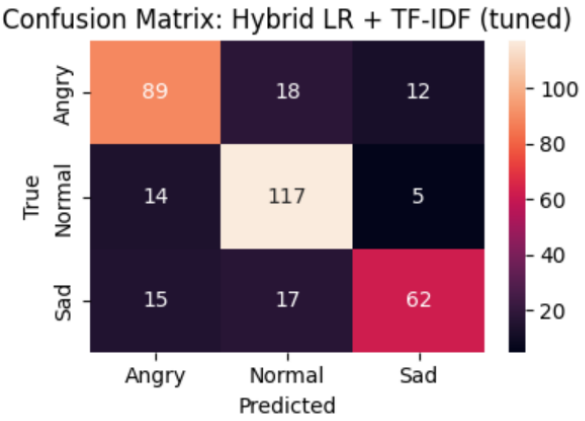

Fig 4.4 Confusion Matrix of Logistic Regression (Transcription)

This model has identified the Angry class relatively well. Here, 89 samples of Angry were correctly identified. 18 were misclassified as Normal and 12 were misclassified as Sad.

The Normal class was identified the best. A total of 117 samples were correctly identified. Only a few samples were misclassified as Angry and Sad. The error rate is comparatively

higher in the Sad class. 62 samples were correctly identified. 15 were misclassified as Angry and 17 were misclassified as Normal.

Overall, this model has shown high performance in the Normal class. It also performed well in the Angry class. The error rate is higher in the Sad class.

4.1.5 Random Forest Classifier

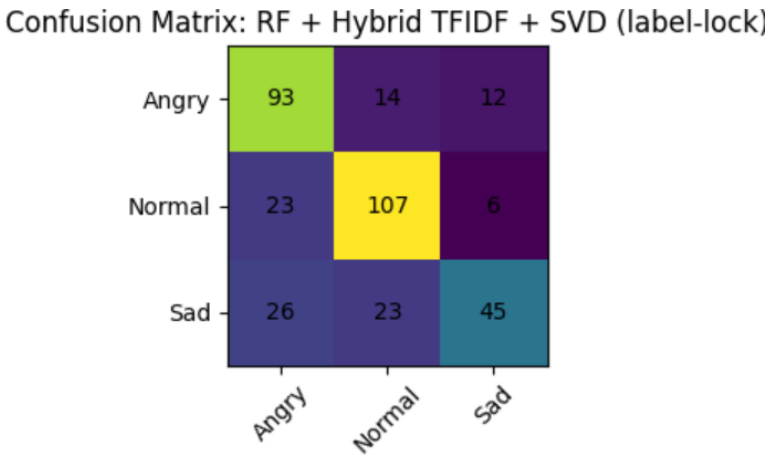

Fig 4.5 Confusion Matrix of Random Forest (Transcription)

In this model, 93 samples from the Angry class were correctly identified. Some samples were misclassified as normal and sad. In the Normal class, 107 samples were correctly identified. However, the number of errors in this class is somewhat higher compared to previous models. The Sad class had the most errors. 45 samples were correctly identified. 26 were incorrectly classified as angry and 23 as normal.

This model performed well on the Angry class. However, the performance is weakest in the Sad class. In the case of Normal, a moderate level of performance was observed.

4.1.6 Support Vector Machine (SVM)

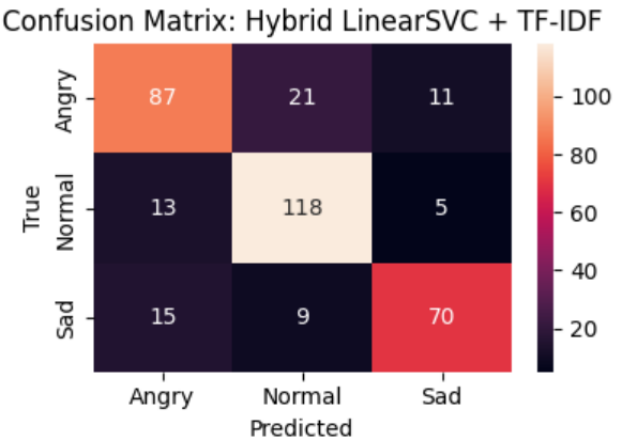

Fig 4.6 Confusion Matrix of Hybrid SVC (Transcription)

This model correctly identified 87 samples of the Angry class. 21 were incorrectly classified as normal, and 11 were incorrectly classified as sad. The Normal class again performed the best. 118 samples were correctly identified, with only a few misclassifications. Improvement is seen in the Sad class. 70 samples were correctly recognized, and the number of errors is lower compared to the previous model.

Overall, this model has shown balanced performance, with notable improvement particularly in the Sad class.

Comparative Discussion

The hybrid LR + TF-IDF tuned model performed best in the Normal class. It also gave good results in the Angry class. However, there were more errors in the Sad class.

The hybrid LinearSVC + TF-IDF model showed balanced performance. Particularly, there is significant improvement in the Sad class compared to the previous model. Strong results were also obtained in the Normal class. The RF + hybrid TF-IDF + SVD model worked well in the Angry class. But the number of errors was the highest in the Sad class. The performance in the Normal class was also moderate.

In summary, the LinearSVC model provided the most stable and balanced performance. The LR model was best in the Normal class. And although the RF model captured Angry well, it was weak in the Sad class.

4.2 Emotion-detection result analysis

Here, MFCC, Spectrogram, Waveform differences have shown of all classes. Then discuss the result of all classification models.

4.2.1 MFCC of all classes

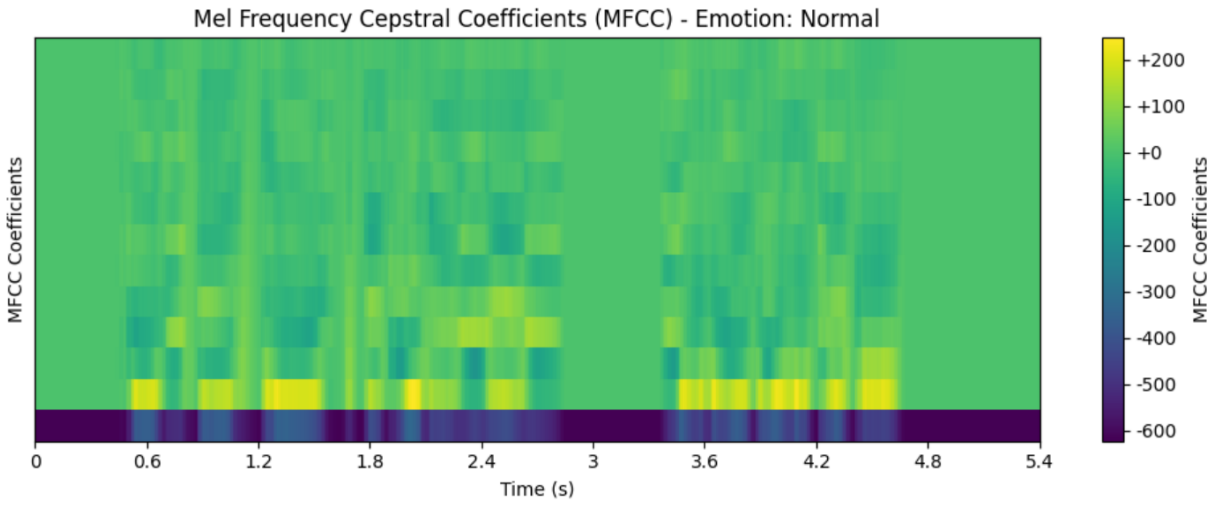

Fig 4.7 MFCC for Normal Emotion

The MFCC graph for normal emotion shows a steady and moderate level of spectral power over time. The coefficients are typically between (–500 and +200). The green-yellow bands in the graph are evenly spread, indicating steady voice intensity and balanced power across frequencies. The changes throughout the graph are smooth. This indicates neutral and controlled speech, similar to normal conversation.

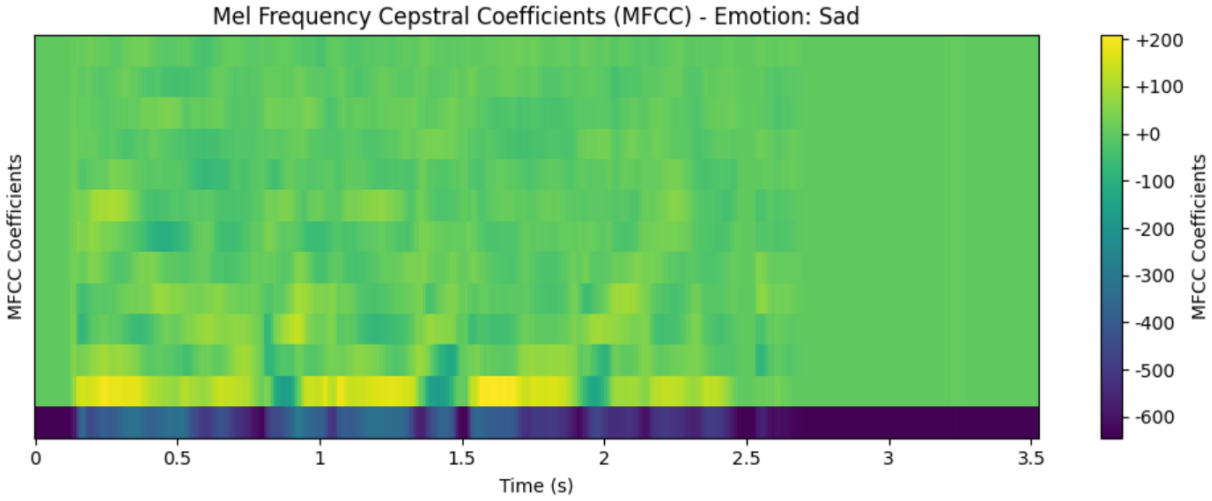

Fig 4.8 MFCC for Sad Emotion

The MFCC graph for the Sad emotion has low power and relatively low spectral variation. The coefficients are typically between (–600 and +150). The yellow part is less prominent and the blue-green part is more prominent. This is a sign of low pitch, low power, and slow vocal variation. Such spectral behaves is typically seen in words with low emotional arousal, slow tempo, even tone, and low power.

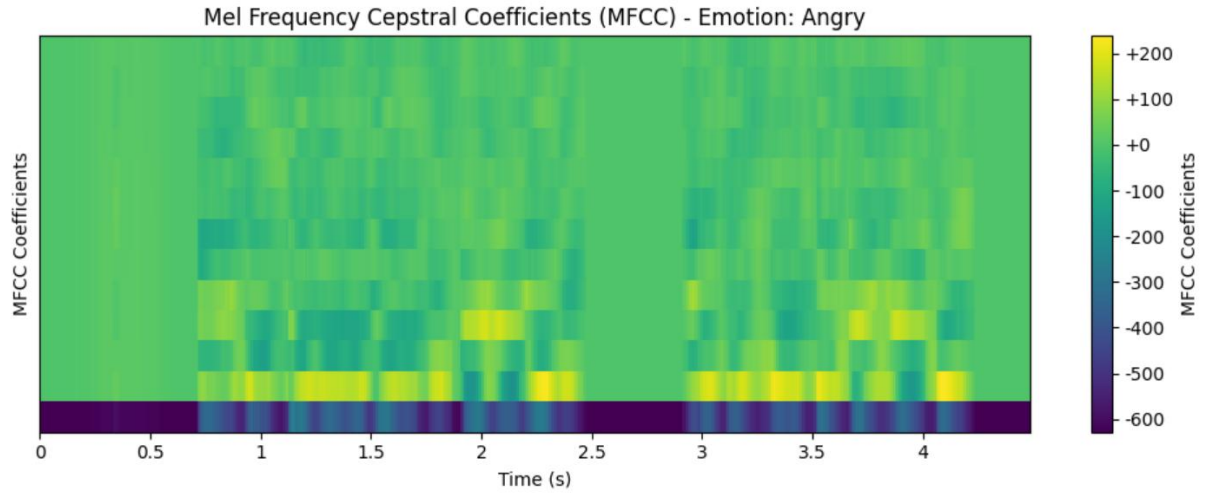

Fig 4.9 MFCC for Angry Emotion

In the MFCC graph of the angry emotion, spectral fluctuations are greater and the energy is higher compared to the other two emotions. Many yellow bands can be seen

on the graph. These indicate sudden surges in energy and rapid pitch changes. Although the coefficients range from (–600 to +200), the variations within this range are more intense and irregular. This type of pattern is associated with high emotional arousal. This increased vocal tension, and a rise in energy flow at higher frequencies, which are characteristic of an angry voice.

4.2.2 Spectrogram

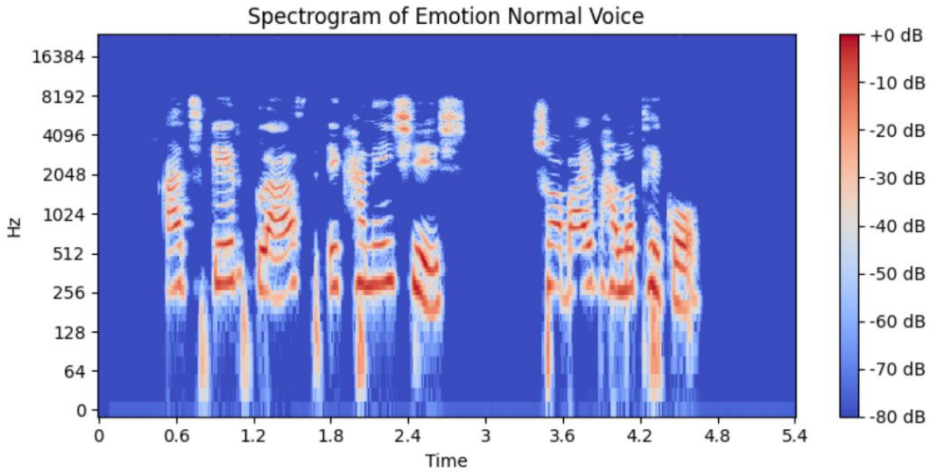

Fig 4.10 Spectrogram of Normal Emotion

In the spectrogram of a normal voice, balanced energy is observed in the low and mid-frequency ranges (approximately 200–2000 Hz). The harmonic bands are uniform and stable. The orange-red areas indicate that the voice has a moderate level of energy. There is no sharp increase at high frequencies, which implies that the speaker was not experiencing excessive tension or stress. The formants are evenly spaced, reflecting normal pronunciation. The entire spectrogram highlights the characteristics of a calm, controlled, and steady voice.

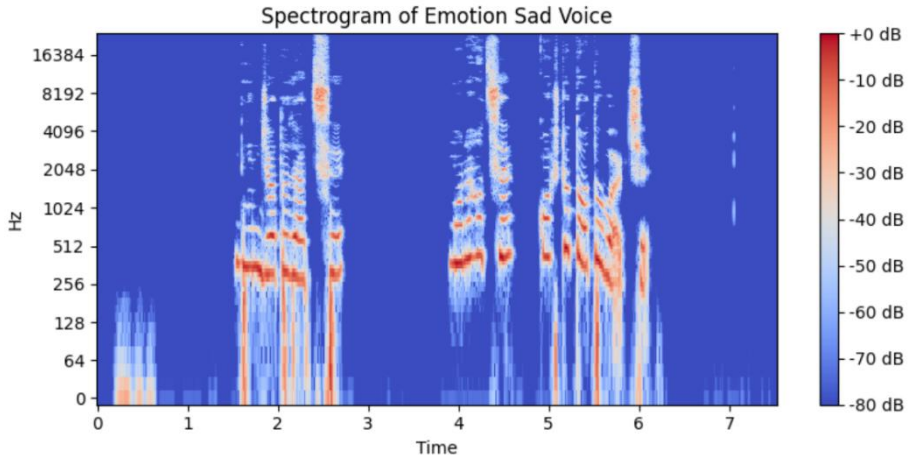

Fig 4.11 Spectrogram of Sad Emotion

In the spectrogram of a sad voice, the energy is mainly concentrated at lower frequencies (below 500 Hz). The high-frequency components appear very weak. The colour intensity usually ranges from –20 to –40 dB, which indicates low energy. The harmonic bands are not very dense and changes occur slowly. This reflects a slow vocal pace and a monotone voice. Activity at high frequencies is very low. This indicates low vocal strength and emotional stress. Overall, the spectrogram represents a soft, slow, and low-pitched voice, which is a common feature of sadness.

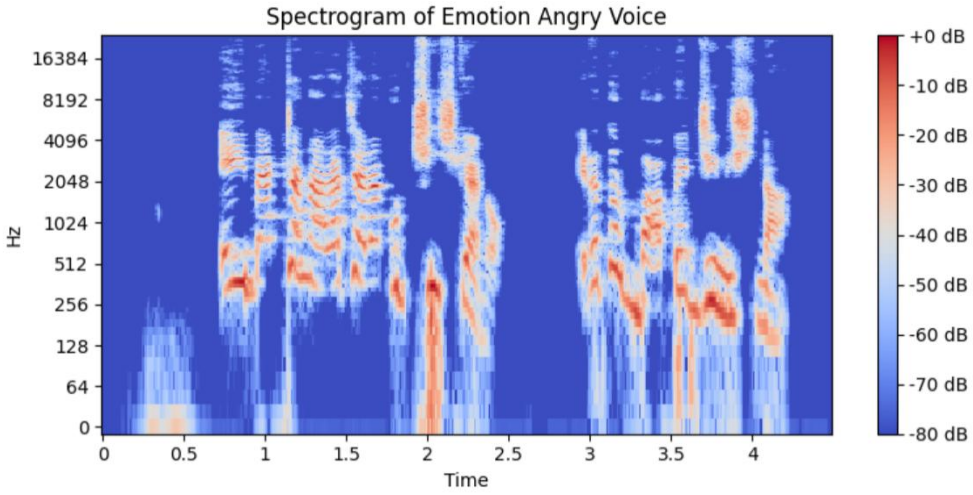

Fig 4.12 Spectrogram of Angry Emotion

In the spectrogram of an angry voice, strong energy is seen across a wide frequency range from low to high (500–4000 Hz). The dense red areas range from 0 to –10 dB, indicating high energy. The harmonic bands are dense, long, and rapidly changing. These indicate intense pitch variations and vocal tension. Compared to normal and sad voices, angry voices show stronger activity at high frequencies. Short pauses and rapid articulation are also signs of anger. Overall, the spectrogram clearly shows high arousal, strong voice, and sharp articulation.

4.2.3 Waveform Before and After Noise Reduction

Here the difference between emotional voices have been shown by waveform pictures. We can see how the waveform of an emotional voice is before noise reduction and after noise reduction.

4.2.3.1 Normal Emotion Waveform

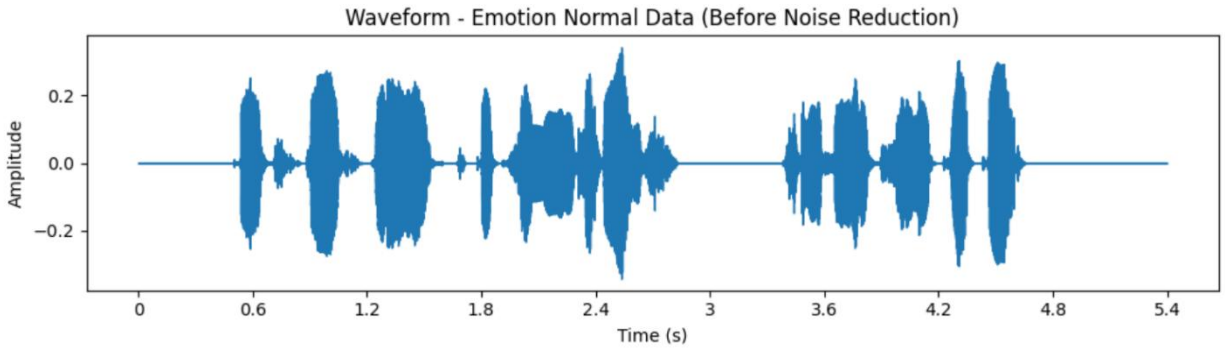

Fig 4.13 Normal Emotion Waveform (Before Noise Reduction)

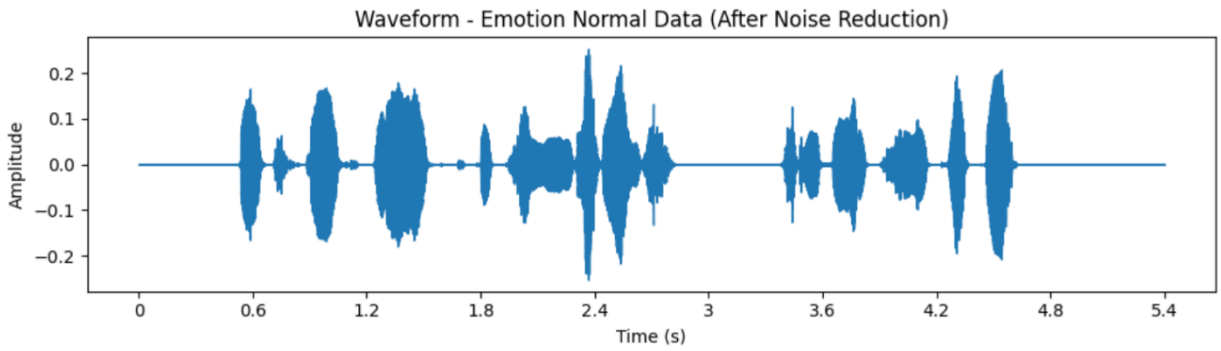

Fig 4.14 Normal Emotion Waveform (After Noise Reduction)

Before noise reduction, the waveform of normal emotion makes the speech clear. However, there is random background noise along with it. Sometimes the amplitude reaches beyond ±0.25. The baseline is also not consistent. This indicates some issues in the recording. While the main pattern of speech can be seen, there is slight distortion in some high-energy parts. This may cause difficulty in extracting accurate features.

After noise reduction, the waveform becomes much smoother. The silent parts between speech are clearly visible. During this time, the amplitude is almost around zero. This indicates that the background noise has been effectively reduced. The overall amplitude range decreases slightly (around ±0.2). However, the natural speech rhythm and voice modulation remain intact.

This improvement shows that noise reduction has successfully removed unnecessary noise and maintained the quality of speech. As a result, the processed signal becomes clear. This makes it more suitable for feature extraction, speech segmentation, and emotion classification.

4.2.3.2 Sad Emotion Waveform

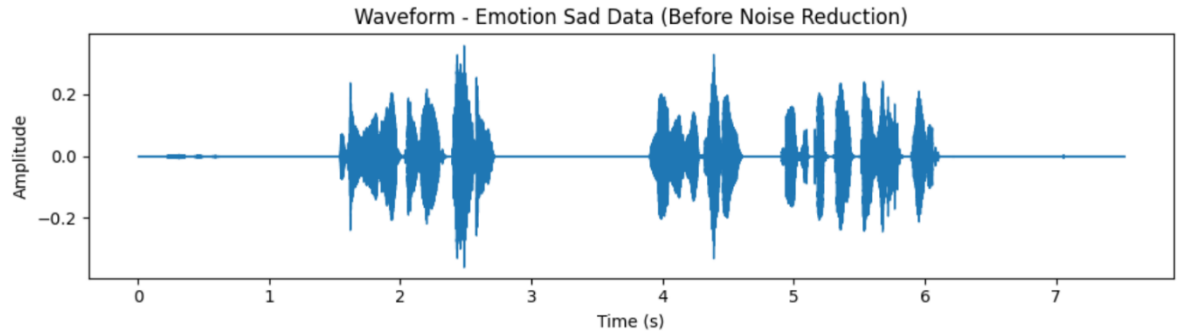

Fig 4.15 Sad Emotion Waveform (Before Noise Reduction)

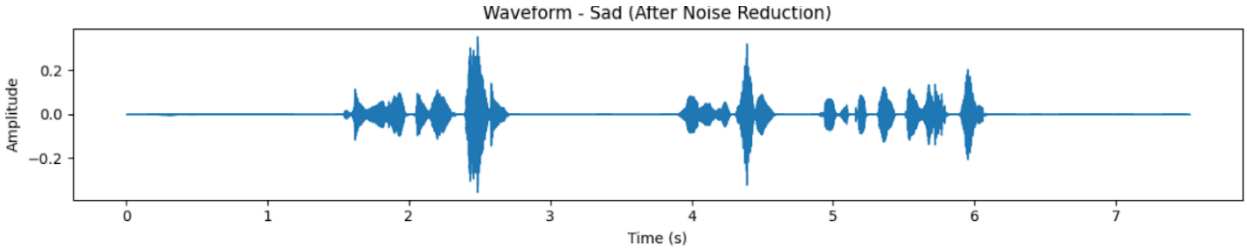

Fig 4.16 Sad Emotion Waveform (After Noise Reduction)

Before noise reduction, the waveform of Sad emotion was very unstable. There were random changes in amplitude. The baseline of the waveform was not stable. As a result, it was difficult to distinguish between silent and voiced parts. Between 2–6 seconds, excessive amplitude was observed. Much of this was not speech, but rather caused by background noise.

After noise reduction, the waveform became much smoother. The silent parts are now clearly visible at nearly zero amplitude. The amplitude is within a moderate range (around ±0.25). This matches the low-energy, slow speech characteristics of Sad emotion. The clear waveform accurately shows the main speech pattern. This allows emotion-based feature extraction to be done more precisely.

Overall, noise reduction has made the signal much cleaner, while also preserving the natural characteristics of Sad emotion.

4.2.3.3 Angry Emotion Waveform

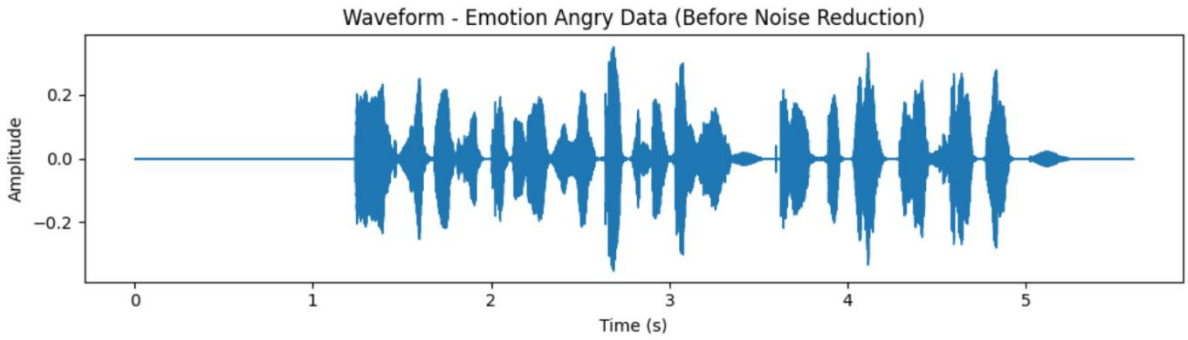

Fig 4.17 Angry Emotion Waveform (Before Noise Reduction)

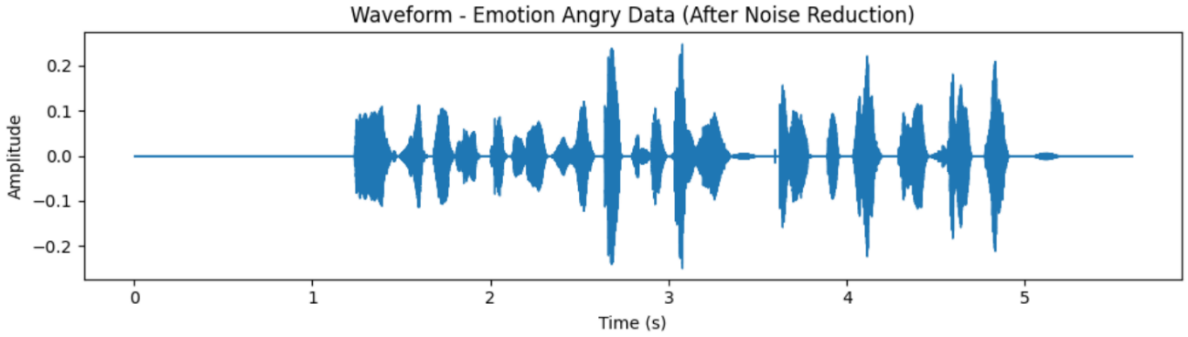

Fig 4.18 Angry Emotion Waveform (After Noise Reduction)

Before noise reduction, the waveform of Angry emotion shows high energy. The amplitude often fluctuates up to ±0.25. There are also irregular changes in the baseline. One reason for this is background noise. Due to this noise, the distinction between speech and silent parts is not clear. As a result, the waveform appears quite dense and disorganized. In anger, the emotion naturally involves strong and rapid amplitude changes. However, random noise masks the fine details of speech.

After noise reduction, the waveform becomes much clearer. Silent parts are more distinctly visible. The overall amplitude decreases slightly (around ±0.2). However, the naturally high-energy peaks of anger remain intact. The reduction of low-level noise smooths the baseline, making the speech segments more noticeable.

These results demonstrate that noise reduction has effectively reduced background noise while preserving the intensity and sharp variations of the Angry emotion. Consequently, the processed signal expresses Angry emotion more clearly and accurately. This improves the reliability of feature extraction and emotion classification.

4.2.4 LSTM

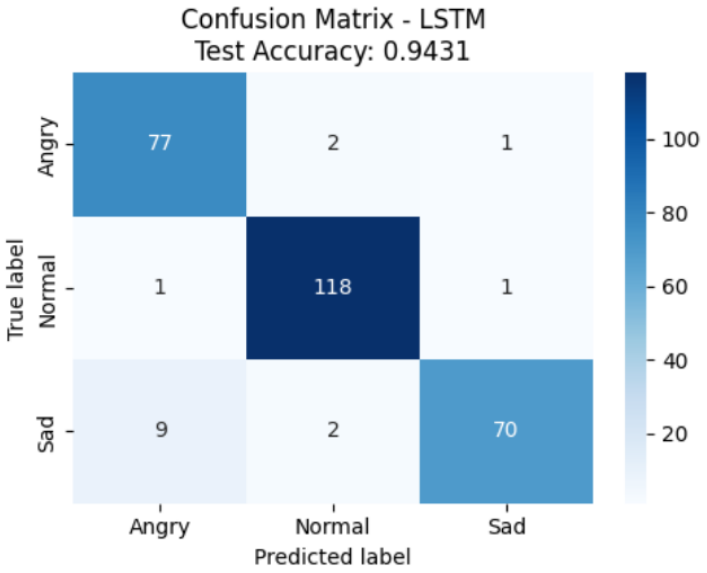

Fig 4.19 Confusion Matrix of LSTM (Emotion Detection)

The overall accuracy of this model was 94.31%. The model correctly identified a total of 77 examples in the Angry class. Here, 2 samples were misclassified as Normal and 1 sample as Sad. The model performed even better in the Normal class. Here, 118 samples were correctly detected. Only 1 misclassified as angry and 1 as sad. On the other hand, in the Sad class, 70 predictions were correct. However, 9 samples were misclassified as angry and 2 as normal. Overall, it is observed that the LSTM model was able to identify the Normal class the best. But there were relatively more errors in the Sad class.

4.2.5 SVC

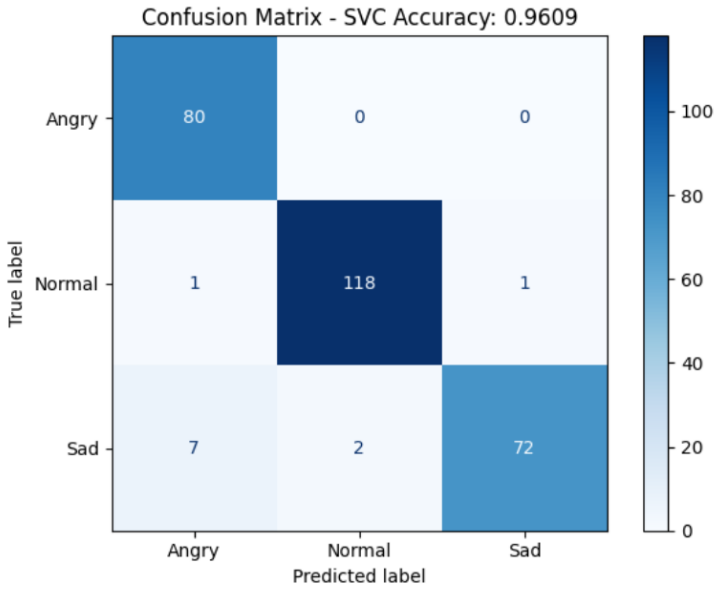

Fig 4.20 Confusion Matrix of SVC (Emotion Detection)

The overall accuracy of the SVC model was 96.09%. This model correctly identified 80 samples in the Angry class and did not misclassify any samples as normal or sad. In the Normal class, 118 samples were correctly identified, but only 1 sample was misclassified as angry and 1 as sad. In the Sad class, the model made 72 correct predictions. But 7 samples were misclassified as angry and 2 as normal. Overall, the SVC model demonstrated stable and reliable performance across all categories. Its high accuracy is particularly clear in the Angry and Normal categories.

4.2.6 Random Forest

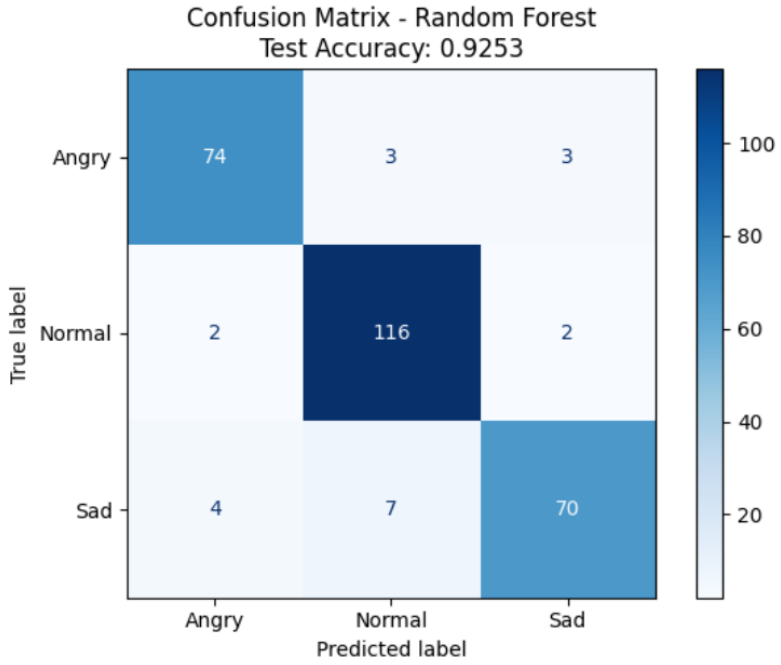

Fig 4.21 Confusion Matrix of Random Forest (Emotion Detection)

The overall accuracy of the Random Forest model was 92.53%. The model correctly identified 74 examples in the Angry class, but misclassified 3 samples as normal and 3 samples as sad. In the Normal class, 116 were correctly predicted. Only 2 misclassified as angry and 2 as sad. On the other hand, in the Sad class, 70 were correctly identified. But 4 samples were misclassified as angry and 7 as normal. Overall, the Random Forest model performed well in the Normal class. But comparatively more errors were seen in the Angry and Sad classes.

Comparing the performance of the three models, it is observed that the SVC model showed the highest accuracy. It consistently performed well across all classes and had the lowest error rate. The LSTM model showed the second-highest performance. Especially in the Normal class. However, there were slightly more errors in the Sad class. On the other hand, the accuracy of the Random Forest model is comparatively lower. In particular, there were more mistakes in the Angry and Sad classes. Overall, while all three models performed well in the Normal class, the SVC model provided the best results overall.

4.2.7 Precision, Recall and F1-Score

LSTM

Table 4.10 Class wise LSTM model's result

Class	Precision	Recall	F1-Score
Angry	0.8851	0.9625	0.9222
Normal	0.9672	0.9833	0.9752
Sad	0.9722	0.8642	0.9150
Accuracy	—	—	0.9431
Macro Avg	0.9415	0.9367	0.9375
Weighted Avg	0.9453	0.9431	0.9428

Random Forest:

Table 4.11 Class wise Random Forest model's result

Class	Precision	Recall	F1-Score
Angry	0.9250	0.9250	0.9250
Normal	0.9206	0.9667	0.9431
Sad	0.9333	0.8642	0.8974
Accuracy	—	—	0.9253
Macro Avg	0.9263	0.9186	0.9218
Weighted Avg	0.9255	0.9253	0.9248

SVC:

Table 4.12 Class wise SVC model's result

Emotion	Precision	Recall	F1-Score
Angry	0.9091	1.0000	0.9524
Normal	0.9833	0.9833	0.9833
Sad	0.9863	0.8889	0.9351

Accuracy			**0.9609**
Macro Avg	0.9596	0.9574	0.9569
Weighted Avg	0.9631	0.9609	0.9606

CHSPTER 5 CONCLUSION AND FUTURE WORK

5.1 Conclusion

The main objective of this research was to develop a Bangla language-based real-time speech-to-subtitle and emotion recognition system. Research in both ASR and SER in the Bangla language is limited. As a result, creating effective communication for hearing-impaired users becomes challenging. After addressing this issue, an integrated system has been developed here that simultaneously converts the speaker's speech into text and identifies their emotions.

For this work, a total of 1,400 audio samples were collected, including 600 Normal, 400 Angry, and 400 Sad samples. The Whisper model was used for the speech-to-text part. Initially, in normal conditions, its WER was about 58.8% and CER about 28.2%, which was expected due to the variation in Bangla pronunciation. However, after dataset-based matching and custom pre-processing, the WER dropped to 6.95% and CER to 4.19%. This demonstrates that the Whisper model can perform well for the Bangla language if it is properly processed.

LSTM, Random Forest, and SVC models were tested for emotion detection. Among the three models, SVC showed the highest performance, achieving 96.09% accuracy. This shows that SVC is an effective method for small-sized Bengali audio data. This integrated system displays the speaker's emotion as 'Normal', 'Angry', or 'Sad' along with their speech. As a result, hearing-impaired users can understand the true intent of the speech more clearly. Therefore, it can play an effective role as an assistive communication technology.

Ultimately, this research develops an integrated ASR-SER system for the Bengali language. This opened an effective way to make communication more meaningful and accessible for hearing-impaired users.

5.2 Future Work

Several aspects can be considered to further improve this system in the future. Firstly, it is essential to increase the model's stability in noisy environments. For this, noise-robust ASR and SER methods can be used. Secondly, the system currently supports only the Bengali language. Its application areas would grow in the future if multilingual support were added. Thirdly, it would be useful under low internet settings if it included offline or batch transcribing capabilities.

Moreover, using larger datasets will make the model's performance more accurate. In the future, multimodal information, such as facial expressions or linguistic features of text, could be added. This would make emotion recognition more reliable. Finally, the system's overall performance can be further enhanced by using a transformer-based SER model.

REFERENCES

Graham, C., & Roll, N. (2024). Evaluating OpenAI's Whisper ASR: Performance analysis across diverse native and non-native English accents. *Journal of the Acoustical Society of America Express Letters*. https://doi.org/10.1121/10.0024876

Schuhmann, C., Kaczmarczyk, R., Rabby, G., Friedrich, F., Kraus, M., Nadi, K., Nguyen, H., Kersting, K., & Auer, S. (2025). *EmoNet-Voice: A fine-grained, expert-verified benchmark for speech emotion detection* (arXiv:2506.09827). arXiv. https://doi.org/10.48550/arXiv.2506.09827

ZainEldin, H., Gamel, S. A., Talaat, F. M., Aljohani, M., Baghdadi, N. A., Malki, A., Badawy, M., & Elhosseini, M. A. (2024). Silent no more: A comprehensive review of artificial intelligence, deep learning, and machine learning in facilitating deaf and mute communication. *Artificial Intelligence Review, 57*(7), 188. https://doi.org/10.1007/s10462-024-10768-2

Samin, M. N. S., Ahad, J. I., Medha, T. A., Rahman, F., Amin, M. R., Mohammed, N., & Rahman, S. (2024). BanglaDialecto: An end-to-end AI-powered regional speech standardization. In *2024 IEEE International Conference on Big Data (BigData)* (pp. 1635–1644). IEEE. https://doi.org/10.1109/BigData58756.2024.10824472

Devi, V. A. (2017). Conversion of speech to Braille: Interaction device for visual and hearing impaired. In *2017 Fourth International Conference on Signal Processing, Communication and Networking (ICSCN)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICSCN.2017.8085703

Islam, S. Z., Akteruzzaman, M., & Abujar, S. (2021). Semantics exploration for automatic Bangla speech recognition. In *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2020, Volume 2* (pp. 171–179). Springer Nature Singapore. https://doi.org/10.1007/978-981-16-3342-3_15

Kuhn, K., Kersken, V., & Zimmermann, G. (2025, April). Communication access real-time translation through collaborative correction of automatic speech recognition. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1–8). https://doi.org/10.1145/3613905.3650771

Samaradivakara, Y., Pathirage, A., Ushan, T., Sasikumar, P., Karunanayaka, K., Keppitiyagama, C., & Nanayakkara, S. (2025). *Tailored real-time AR captioning interface for enhancing learning experience of deaf and hard-of-hearing (DHH) students* (arXiv:2501.02233). arXiv. https://doi.org/10.48550/arXiv.2501.02233

Liu, G., Cai, S., & Wang, C. (2023). Speech emotion recognition based on emotion

perception. *EURASIP Journal on Audio, Speech, and Music Processing, 2023*(22). https://doi.org/10.1186/s13636-023-00289-4

Pan, S.-T., & Wu, H.-J. (2023). Performance improvement of speech emotion recognition systems by combining 1D CNN and LSTM with data augmentation. *Electronics, 12*(11), 2436. https://doi.org/10.3390/electronics12112436

Fang, C., Jin, Y., Chen, G., Zhang, Y., Li, S., Ma, Y., & Xie, Y. (2024). Multimodal speech emotion recognition based on large language model. *IEICE Transactions on Information and Systems, E107.D*(11), 1463–1467. https://doi.org/10.1587/transinf.2024EDL8034

Ferraro, A., Galli, A., La Gatta, V., & Postiglione, M. (2023). Benchmarking open source and paid services for speech to text: An analysis of quality and input variety. *Frontiers in Big Data, 6*, 1210559. https://doi.org/10.3389/fdata.2023.1210559

Yin, C. (2024). An optimized approach to speech transcription using blind source separation and speech-to-text machine learning models. *Applied and Computational Engineering, 30*, 131–138. https://doi.org/10.54254/2755-2721/30/20230085

Gómez-Sirvent, J. L., López de la Rosa, F., Sánchez-Reolid, D., Sánchez-Reolid, R., & Fernández-Caballero, A. (2025). Small language models for speech emotion recognition in text and audio modalities. *Applied Sciences, 15*(14), 7730. https://doi.org/10.3390/app15147730

Dash, P., Babu, S., Singaravel, L., et al. (2025). Generative AI-powered multilingual ASR for seamless language-mixing transcriptions. *Journal of Electrical Systems and Information Technology, 12*, 42. https://doi.org/10.1186/s43067-025-00204-1

Pulatov, I., Oteniyazov, R., Makhmudov, F., & Cho, Y.-I. (2023). Enhancing speech emotion recognition using dual feature extraction encoders. *Sensors (Basel), 23*(14), 6640. https://doi.org/10.3390/s23146640

Madhusudhana Reddy, V., Vaishnavi, T., & Pavan Kumar, K. (2023). Speech-to-text and text-to-speech recognition using deep learning. In *2023 2nd International Conference on Edge Computing and Applications (ICECAA)*. IEEE. https://doi.org/10.1109/ICECAA58104.2023.10212222

Dhakad, A., & Singh, S. (2023). Python-powered speech-to-text: A comprehensive survey and performance analysis. *International Journal of Engineering Research & Technology (IJERT), 12*(9). https://www.ijert.org/python-powered-speech-to-text-a-comprehensive-survey-and-performance-analysis

Fang, C., Jin, Y., Chen, G., Zhang, Y., Li, S., Ma, Y., & Xie, Y. (2024). Multimodal speech emotion recognition based on large language model. *IEICE Transactions on Information and Systems, E107.D*(11), 1463–1467. https://doi.org/10.1587/transinf.2024EDL8034

Ferraro, A., Galli, A., La Gatta, V., & Postiglione, M. (2023). Benchmarking open-source and paid services for speech to text. *Frontiers in Big Data, 6*, 1210559. https://doi.org/10.3389/fdata.2023.1210559

Radford, A., et al. (2022). Whisper: Large-scale weak supervision for speech recognition. *(OpenAI Technical Report)* https://doi.org/10.48550/arXiv.2212.04356

Reddy, V. M., Vaishnavi, T., & Kumar, P. (2023). Speech-to-text and text-to-speech recognition using deep learning. In *2023 2nd International Conference on Edge Computing and Applications (ICECAA).* IEEE. https://doi.org/10.1109/ICECAA58104.2023.10212222

Roushan, R., Mishra, H., et al. (2024). *Fine-tuning Whisper and developing a web application.* In 2024 IEEE Conference on Engineering Informatics (ICEI). https://doi.org/10.1109/ICEI64305.2024.10912421

Zhang, L., Wu, S., & Wang, Z. (2025). LoRA-INT8 Whisper: A low-cost Cantonese speech recognition framework for edge devices. *Sensors, 25*(17), 5404. https://doi.org/10.3390/s25175404

Jain, R., et al. (2024). *Exploring native and non-native English child speech recognition with Whisper.*Preprint.https://arxiv.org/abs/2405.12345

Koenecke, A., et al. (2024). *Assessing hallucinations in Whisper ASR.* arXiv preprint.https://arxiv.org/abs/2404.09912

Liu, Y., et al. (2024). *Whisper fine-tuning strategies for low-resource languages.* arXiv preprint.https://arxiv.org/abs/2403.03472

Mengke, D., et al. (2025). *Transliteration-aided transfer learning for Khalkha Mongolian ASR.*arXiv preprint.https://arxiv.org/abs/2501.00456

Polat, H., et al. (2024). *Turkish automatic speech recognition using LoRA adaptation.* arXiv preprint.https://arxiv.org/abs/2407.08119

Pratama, R. S. A., et al. (2024). *Analysis of Whisper automatic speech recognition performance on low-resource language. Jurnal Pilar Nusa Mandiri.* https://ejournal.nusamandiri.ac.id/index.php/pilar/article/view/4269

Gokcimen, T., Das, B., & Das, R. (2024). Evaluating the performance of Turkish automatic speech recognition using the generative AI-based Whisper model. *Proceedings of …* IEEE. https://doi.org/10.1109/ubmk63289.2024.10773523

Gómez-Sirvent, J. L., López de la Rosa, F., Sánchez-Reolid, D., Sánchez-Reolid, R., & Fernández-Caballero, A. (2025). Small language models for speech emotion recognition in text and audio modalities. *Applied Sciences, 15*(14), 7730. https://doi.org/10.3390/app15147730

Hossain, S. M., Rihan, M. R., Imtiaz, A., Boni, P. K., & Gomes, D. (2024). Enhancing

Bangla local speech-to-text conversion using fine-tuning Wav2vec 2.0 with OpenSLR and self-compiled datasets through transfer learning. In *Proceedings of the 7th IEOM Bangladesh International Conference on Industrial Engineering and Operations Management (2024)*. https://doi.org/10.46254/BA07.20240161

Rakib, M., Ismail Hossain, M., Mohammed, N., & Rahman, F. (2022). Bangla-Wave: Improving Bangla automatic speech recognition utilizing n-gram language models. arXiv preprint. https://arxiv.org/abs/2209.12650

Kang, Z., Peng, J., Wang, J., & Xiao, J. (2022). *SpeechEQ: Speech emotion recognition based on multi-scale unified datasets and multitask learning*. arXiv preprint.https://arxiv.org/abs/2206.13101

Li, Y., Wang, Y., Yang, X., & Im, S. (2023). *Speech emotion recognition based on Graph-LSTM neural network.* EURASIP Journal on Audio, Speech, and Music Processing, 40 (2023). https://doi.org/10.1186/s13636-023-00303-9

Pentari, A., Kafentzis, G., & Tsiknakis, M. (2024). *Speech emotion recognition via graph-based representations.* Scientific Reports, 14, 4484. https://doi.org/10.1038/s41598-024-52989-2

Jain, R., Barcovschi, A., Yiwere, M. Y., Corcoran, P., & Cucu, H. (2024). Exploring native and non-native English child speech recognition with Whisper. IEEE Access, 12, 41601–41610. https://doi.org/10.1109/ACCESS.2024.3378738

Polat, H., Turan, A. K., Koçak, C., & Ulaş, H. B. (2024). Implementation of a Whisper architecture-based Turkish automatic speech recognition (ASR) system and evaluation of the effect of fine-tuning with a low-rank adaptation (LoRA) adapter on its performance. *Electronics, 13*(21), 4227. https://doi.org/10.3390/electronics13214227

Li, Y., Bell, P., & Lai, C. (2022). Fusing ASR outputs in joint training for speech emotion recognition. *ICASSP 2022 – IEEE International Conference on Acoustics, Speech and Signal Processing*, 7362–7366. https://doi.org/10.1109/ICASSP43922.2022.9746289

Li, Y., Zhao, Z., Klejch, O., Bell, P., & Lai, C. (2023). ASR and emotional speech: A word-level investigation of the mutual impact of speech and emotion recognition. *Interspeech 2023*, 1449–1453. https://doi.org/10.21437/Interspeech.2023-2078

Li, Y., Bell, P., & Lai, C. (2024). Speech emotion recognition with ASR transcripts: A comprehensive study on word error rate and fusion techniques. *Spoken Language Technology Workshop (SLT) 2024*. https://arxiv.org/abs/2406.08353