**VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY**
**(An Autonomous Institute Affiliated to University of Mumbai)**
**Department of Computer Engineering**


Project Report on

# EmoSpeak: An Emotionally Intelligent TTS System for Visually Impaired

Submitted in partial fulfillment of the requirements of the degree

**BACHELOR OF ENGINEERING** IN **COMPUTER ENGINEERING**

By

**Shamal Dhekale /13**
**Chandni Gangwani /16**
**Bhagyashree Vaswani /66**

**Project Mentor**

Prof. Mrs. Yugchhaya Galphat

**University of Mumbai (AY 2023-24)**

**VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY**

**(An Autonomous Institute Affiliated to University of Mumbai)**

**Department of Computer Engineering**


# CERTIFICATE

This is to certify that the Mini Project entitled "**EmoSpeak: An Emotionally Intelligent TTS System for Visually Impaired**" is a bonafide work of **Shamal Dhekale(13), Chandni Gangwani(16) and Bhagyashree Vaswani(66)** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of "**Bachelor of Engineering**" in "**Computer Engineering**".

**(Prof. Mrs. Yugchhaya Galphat)**
Mentor

**(Prof.Dr. Nupur Giri)**
Head of Department

**(Prof.Dr. J.M. Nair)**
Principal

# Mini Project Approval

This Mini Project entitled "**EmoSpeak: An emotionally intelligent TTS system for the visually impaired"** by **Shamal Dhekale(13), Chandni Gangwani(16) and Bhagyashree Vaswani(66)** is approved for the degree of **Bachelor of Engineering** in **Computer Engineering.**

**Examiners**

**1……**

(Internal Examiner Name & Sign)

**2……**

(External Examiner name & Sign)

Date:

Place:

# Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

| | |
|---|---|
| --- | --- |
| (Signature) | (Signature) |
| --- | --- |
| (Name of student and Roll No.) | (Name of student and Roll No.) |
| --- | --- |
| (Signature) | (Signature) |
| --- | --- |
| (Name of student and Roll No.) | (Name of student and Roll No.) |

Date:

# ACKNOWLEDGEMENT

We are thankful to our college Vivekanand Education Society's Institute of Technology for considering our project and extending help at all stages needed during our work of collecting information regarding the project.

It gives us immense pleasure to express our deep and sincere gratitude to Assistant Professor **Mrs. Yugchhaya Galphat** (Project Guide) for her kind help and valuable advice during the development of project synopsis and for her guidance and suggestions.

We are deeply indebted to Head of the Computer Department **Dr.(Mrs.) Nupur Giri** and our Principal **Dr. (Mrs.) J.M. Nair** , for giving us this valuable opportunity to do this project.

We express our hearty thanks to them for their assistance without which it would have been difficult in finishing this project synopsis and project review successfully.

We convey our deep sense of gratitude to all teaching and non-teaching staff for their constant encouragement, support and selfless help throughout the project work. It is a great pleasure to acknowledge the help and suggestion, which we received from the Department of Computer Engineering.

We wish to express our profound thanks to all those who helped us in gathering information about the project. Our families too have provided moral support and encouragement several times.

# Index

**Chapter 5: Implementation of the Proposed System**

5.1. Methodology employed for development

5.2 Algorithms and flowcharts for the respective modules developed

5.3 Datasets source and utilization

**Chapter 6: Results and Discussion**

6.1. Screenshots of User Interface (UI) for the respective module

6.2. Performance Evaluation measures

6.3. Input Parameters / Features considered

6.4. Graphical and statistical output

6.5. Comparison of results with existing systems

6.6. Inference drawn

**Chapter 7: Conclusion**

7.1 Limitations

7.2 Conclusion

7.3 Future Scope

**References**

**Published Papers**

# Abstract

EmoSpeak stands at the forefront of innovation in assistive technology, not only as a powerful tool for the visually impaired but as a symbol of the ongoing progress in the field of human-computer interaction. It takes a giant leap forward by addressing a critical aspect of communication that is often taken for granted by those with full sensory capabilities - the ability to perceive and convey emotions.

For individuals who are visually impaired, EmoSpeak represents a newfound sense of empowerment. It offers them the key to unlock a world where spoken words are not just a means of conveying information but a channel for understanding the complex tapestry of human emotions. Through its advanced emotion recognition algorithms, EmoSpeak can detect and infuse synthesized speech with a rich emotional context, providing users with the invaluable ability to comprehend the feelings and intentions of those they interact with.

This innovative system goes beyond mere words; it delves into the heart of what makes human interaction so profound - emotions. By translating emotional cues into audible signals, EmoSpeak enhances the user's experience of the world. It allows them to engage more deeply in conversations, fostering a stronger sense of connection with others, whether it's with loved ones, colleagues, or strangers. This newfound depth of communication brings not only comprehension but also empathy, compassion, and shared understanding.

EmoSpeak is not just a technological achievement but a stepping stone toward creating a more inclusive society. By bridging the emotional communication gap, it dismantles barriers that may have previously hindered the visually impaired from fully participating in social, educational, and professional settings. It doesn't just provide a tool; it offers a pathway to enriched interactions and heightened inclusivity, fostering a world where everyone, regardless of their visual abilities, can engage meaningfully with the people and environments around them.

# List of Abbreviations

| SR. NO. | ABBREVIATED FORM | FULL FORM |
|---|---|---|
| 1 | TTS | Text-To-Speech |
| 2 | SST | Speech-To-Text |
| 3 | NLP | Natural Language Processing |
| 4 | CNN | Convolutional neural networks |
| 5 | RNN | Recurrent neural networks |
| 6 | GPU | Graphics Processing Unit |
| 7 | E2E-TTS | End-to-end text-to-speech |
| 8 | ASR | Automated Speech Recognition |
| 9 | MOS | Mean opinion score |
| 10 | GRU | Gated Recurrent Unit |
| 11 | MAML | Model Agnostic Meta-Learning |
| 12 | VC | Voice conversion |
| 13 | LSTM | Long Short Term Memory |
| 14 | seq2seq | Sequence-to-Sequence |
| 15 | VCTK | Voice Cloning Toolkit |
| 16 | SSRN | Social Science Research Network |
| 17 | API | Application Programming Interface |
| 18 | UTF-8 | Unicode Transformation Format |
| 19 | CART | Classification and Regression Tree |
| 20 | MFCCs | Mel-frequency cepstral coefficients |

| | | |
|---|---|---|
| 21 | RAM | Random Access Memory |
| 22 | Ghz | GigaHertz |
| 23 | GB | GigaBytes |
| 24 | JDK | Java Development Kit |
| 25 | LSTM | Long Short Term Memory |
| 26 | CSV | Comma-separated values |
| 27 | UTF-8 | Unicode Transformation Format-8 |
| 28 | OCR | Optical Character Recognition |
| 29 | ML | Machine Learning |
| 30 | LR | Learning Rate |
| 31 | TP | True Positive |
| 32 | TN | True Negative |
| 33 | FP | False Positive |
| 34 | FN | False Negative |

# List of Figures

# List of Tables

# Chapter 1: Introduction

## 1.1 Introduction

Blindness is a severe visual impairment that significantly reduces one's ability to see. It is characterized by total blindness or the inability to see forms, colors, or light. In order to go around, blind persons usually rely on support from other senses such as touch, hearing, smell, and taste and adaptive techniques. These senses are essential for people to comprehend emotions in their daily lives.

Worldwide, a minimum of 2.2 billion individuals experience either near or distance vision challenges. While vision loss can impact individuals across all age groups, the majority of those affected by vision impairment and blindness are typically aged 50 years and older. Types of low vision are Central vision impairment (difficulty seeing objects in the center of the field of vision), Peripheral vision impairment (difficulty seeing objects in the peripheral vision), Nocturnal vision impairment (difficulty seeing in low-light conditions), Blurred or hazy vision.

The project primarily focuses on individuals with low vision, elderly individuals facing difficulty in reading or seeing clearly, and also offers insights applicable to dyslexic individuals struggling with fluent word reading and spelling. They often experience feelings of exclusion from the outside world. They desire to be treated equally and wish to communicate with those considered "normal". However, they encounter difficulties when using technology, especially when dealing with text-based content. For instance, navigating social media or chat applications can be challenging for them, ultimately resulting in their isolation from society. These challenges include difficulties in accessing and navigating digital platforms due to inadequate accessibility features, compatibility issues with screen reader software, and struggles with reading small text sizes or complex fonts. Moreover, they encounter challenges in interpreting emotional nuances conveyed through written language. These challenges encompass difficulties in discerning subtle emotional cues such as nuances or humor, which are commonly conveyed visually through font styles, emojis, or punctuation marks. The lack of alternative text descriptions for visual elements like emoticons or images further impedes their ability to grasp the emotional context of written content. Consequently, these obstacles contribute to potential misunderstandings and communication barriers for individuals with visual impairments. This highlights the importance of accessible and inclusive technology to bridge the gap and foster connections within the community. This project presents "EmoSpeak," a chat application tailored for individuals with visual impairments. "EmoSpeak" is crafted to facilitate robust interaction, empowering visually

impaired users to engage confidently with their environment. The system primarily facilitates text-to-speech and speech-to-text conversions, empowering users to communicate seamlessly.

Furthermore, the EmoSpeak application is integrated with an Emotion Detection from text Model for enhancing communication and emotional understanding among visually impaired individuals. The technology guarantees smooth interaction, which enables visually impaired people and sighted users to communicate effectively. A speech message sent by a user is transformed to text for the recipient. On the other hand, when a user receives a text message from the sender, it is converted into voice and identified emotions are also communicated. The model categorizes the sentiment into 6 emotions- joy , sadness , fear , anger , love and surprise. This method improves the overall communication experience of visually impaired users by guaranteeing that they keep a true link with the outside world, especially in text-based communication.

## 1.2 Motivation

The motivation behind the EmoSpeak project is deeply rooted in our commitment to empowering visually challenged individuals and enhancing their quality of life. Visual impairment creates unique challenges in daily life, particularly in the realm of human communication. Emotions serve as a bridge to connect people on a profound level, conveying not only the words but also the underlying sentiment, intent, and nuances. By developing an Emotion-Aware Text-to-Speech system, we aim to equip visually impaired individuals with the tools they need to access these vital emotional cues in conversations and written content. The EmoSpeak project aligns with our vision of creating a more inclusive and accessible society. We believe that technology should be a force for good, breaking down barriers and promoting equality. This system embodies our commitment to ensuring that no individual is left behind in the ever-evolving digital landscape. Technology is often seen as a means to an end, a tool that serves a practical purpose. However, we firmly believe that technology can do more than that; it can embody the essence of human connection. Emotions are at the core of human interaction, enriching conversations, and deepening relationships. The EmoSpeak project is a testament to our commitment to bring the human touch into technology.

## 1.3 Problem Definition

The problem is twofold: first, communication apps often lack features and design considerations that accommodate the unique needs of visually impaired users. Second, the emotional nuances that form an essential part of human interaction are inadequately represented in the synthesized

speech that these users rely on. The absence of emotional expression in speech impedes their ability to perceive and convey emotions effectively, thereby limiting the depth and richness of their interactions.

The primary objective of the EmoSpeak project is to address these challenges by creating a state-of-the-art Emotion-Aware Text-to-Speech (TTS) system tailored specifically for visually impaired individuals.

In summary, the problem at hand is the digital divide that restricts visually impaired individuals from fully accessing and participating in the digital world, and the limited emotional expressiveness in communication tools designed for them. The EmoSpeak project aims to bridge this gap by developing an innovative solution that empowers visually impaired users to connect more deeply, communicate more effectively, and enrich their interactions across various aspects of life.

## 1.4 Existing Systems

There are several existing systems designed to assist visually challenged individuals, spanning various areas including mobility, reading, communication, and daily tasks. Following are the systems:

1. VoiceOver (iOS): VoiceOver is a built-in screen reader feature on iOS devices that provides spoken feedback to visually impaired users, enabling them to navigate and interact with apps and websites.
2. TalkBack (Android): Similar to VoiceOver, TalkBack is a screen reader feature on Android devices that provides spoken feedback and touch exploration to assist visually impaired users in accessing and interacting with apps and websites.
3. Voice Dream Reader: Voice Dream Reader is an accessible reading app that supports various formats, including text-to-speech, for reading digital content such as eBooks, PDFs, and web articles. It offers customization options for font, color, and reading speed to suit individual preferences.
4. IBM Watson Natural Language Understanding: IBM Watson offers a Natural Language Understanding service that includes sentiment analysis capabilities. It can analyze text to identify the overall sentiment as positive, negative, or neutral, along with detecting specific emotions such as joy, anger, sadness, and fear.
5. Google Cloud Natural Language API: Google Cloud's Natural Language API provides sentiment analysis functionality, allowing developers to extract sentiment and emotion

information from text. It can detect emotions such as joy, sadness, anger, and fear, along with sentiment polarity (positive, negative, or neutral).

6. Clarifai: Clarifai offers a text analysis API that includes sentiment analysis and emotion detection features. It can analyze text to determine sentiment polarity and detect emotions such as joy, sadness, anger, and fear, providing insights into the emotional content of text data.

## 1.5 Lacuna of the existing systems

While the existing systems mentioned have made significant strides in improving accessibility for visually impaired individuals, there are still areas where they may have limitations or room for improvement:

1. Language Support: Some apps may have limited language support, which can be a barrier for users who speak languages other than the primary languages supported by the app.
2. Accessibility of User Interface: While the primary purpose of these apps is to improve accessibility, the user interface itself may not always be fully accessible to all users, particularly those with additional disabilities or impairments.
3. Accuracy and Reliability: The accuracy and reliability of features such as text recognition (OCR) and object identification can vary depending on factors like lighting conditions, image quality, and the complexity of the content being processed.
4. Context Understanding: Emotion detection systems may struggle to accurately interpret emotions in contextually complex or ambiguous text. Understanding sarcasm, irony, or cultural nuances remains a challenge.
5. Bias and Fairness: Emotion detection models can inherit biases present in training data, leading to unfair or inaccurate assessments, especially for underrepresented groups. Addressing biases and ensuring fairness in emotion detection algorithms is crucial for ethical deployment.
6. Integration with Third-Party Apps and Services: Some apps may have limited integration with other apps and services, which could restrict their usability and functionality for certain tasks or workflows.

## 1.6 Relevance of the Project

We are currently witnessing a transformative era where technological innovations, particularly in the realms of Text-to-Speech (TTS), Speech-to-Text (STT), and emotion detection from text, are rapidly advancing. In this project, we harness the capabilities of these integrated technologies to develop a groundbreaking system tailored to support visually challenged individuals in their daily communication and emotional well-being.

A significant challenge faced by visually impaired individuals is accessing and comprehending textual information, especially in real-time interactions. By amalgamating TTS, STT, and emotion detection within a chat application, we aim to bridge this gap and empower visually challenged individuals to engage more seamlessly in conversations and social interactions.

One of the primary hurdles in conventional communication aids for the visually impaired is the lack of emotional context conveyed through text-based interactions. Through the integration of emotion detection, our system goes beyond mere text conversion, enabling users to discern emotional cues in conversations, thereby fostering more meaningful and empathetic communication exchanges.

Furthermore, the ability to convert spoken dialogue to text and vice versa in real-time greatly enhances the fluidity and spontaneity of conversations for visually challenged users. With TTS, users can listen to incoming messages, while STT allows them to respond vocally, ensuring a dynamic and interactive chat experience.

In essence, our integrated chat application not only facilitates efficient text-based communication but also enriches the emotional depth of interactions, empowering visually challenged individuals to connect more deeply with others and navigate the nuances of human expression in the digital realm.

# Chapter 2: Literature Survey

## A. Brief Overview of Literature Survey

The challenges faced by visually challenged individuals in web communications are significant and often overlooked. With the increasing reliance on digital platforms for information and communication, the accessibility barriers they encounter can lead to exclusion and isolation. Machine Learning (ML) has emerged as a powerful tool to address these challenges and transform the way visually impaired individuals interact with the web and to detect emotions.

Various ML algorithms and techniques have been proposed to enhance web accessibility for visually impaired users, ranging from screen reader compatibility and text-to-speech synthesis to emotion detection and sentiment analysis. By leveraging these advancements, we aim to develop an end-to-end solution that addresses the specific needs of visually impaired individuals in navigating and interacting with web content seamlessly.

## B. Related Works

Table 1 Literature Survey of Existing System

| Sr. No. | Title | Dataset Used | Method | Result/Comments |
|---|---|---|---|---|
| 1 | Adaspeech | LibriTTS datasets, VCTK and LJSpeech datasets | FastSpeech2 | The MOS and SMOS scores with 95% confidence intervals when adapting the source AdaSpeech model (trained on LibriTTS) to LJSpeech, VCTK and LibriTTS datasets. |
| 2 | Efficiently Trainable Text-To-Speech System Based On Deep | LJ Speech Dataset | Convolutional Neural Networks (CNN), Deep Learning | The training throughput was ~3.8 minibatch/s (Text2Mel) and ~6.4 minibatch/s (SSRN). This implies that they could iterate the updating formulae of |

| | | | | |
|---|---|---|---|---|
| | Convolutional Networks With Guided Attention | | | Text2Mel 200K times in 15 hours. It shows that the method can almost correctly focus on the correct characters, and synthesize quite clear spectrograms. The MOS (95% confidence interval) was 2.71±0.66 (15 hours training) while the Tacotron's was 2.07±0.62. |
| 3 | Espnet-Tts: Unified, Reproducible, And Integratable Open Source End-To-End Text-To-Speech Toolkit | LJ Speech Dataset | Tacotron 2, Transformer TTS and FastSpeech | From the results shown in the paper, all of the models can generate the features in less than RTF = 1.0 even on CPU. Transformer TTS is slower than Tacotron 2 but FastSpeech is much faster than the other models. Especially on GPU, FastSpeech is 30 times faster than Tacotron 2 and 200 times faster than Transformer TTS. Since FastSpeech is an nonautoregressive model, it can fully utilize the GPU without the bottleneck of the loop processing. Therefore, the improvement rate on GPU is higher than the other models. |
| 4 | Meta-TTS: | LibriTTS | Model | The results in the paper |

| | | | | |
|---|---|---|---|---|
| | Meta-Learning for Few-Shot Speaker Adaptive Text-to-Speech | dataset, VCTK dataset. | Agnostic Meta-Learning (MAML), FastSpeech2 | indicate that when it comes to speaker encoding methods, the most effective option is using d-vector, where the speaker encoder is pre-trained and kept fixed. Training the speaker encoder alongside the TTS model leads to worse performance, whether it is trained from scratch or pre-trained, as it can overfit to the TTS training data. |
| 5 | Sentence-Level Emotion Detection from Text Based on Semantic Rules | ISER emotion dataset | Sentiment Analysis | The results shown in few of the statements were grammatically incorrect; instead, they were more in line with the speaker's idiom. Each synonym in the database of emotions is currently searched. However, The methods used to choose the placement of such words should be appropriate. Idioms that describe an emotion that could not be constructed in this experiment. |
| 6 | An effective approach for emotion detection in multimedia text data using | ISER dataset, blog posts and annotated headlines. | Deep Learning Models | The collected dataset’s performance has been evaluated on the proposed CNN model and LSTM model, and the performance of the models has been |

| | | | | |
|---|---|---|---|---|
| | sequence based convolutional neural network | | | analyzed accordingly. The special effects of the hyper-parameters are evaluated in the experiments. |
| 7 | Emotional Voice Conversion Using Multitask Learning With Text-To-Speech | male-Korean-Emotional-Text-to-speech (mKETTS) dataset. | Emotional Voice Converter | To investigate the emotional voice conversion, the VCTTS model mentioned above was used for inference. After train-ing, we randomly chose 20 samples per each emotion, and those samples were fed into the model. The hs was obtained per each sample, and the cosine similarity between each sample was measured.The mean values of the cosine similarity between emotion pairs are also shown. |

## 2.1 Research Papers

**1) Adaspeech**

**a) Abstract**: AdaSpeech is a novel custom voice Text-to-Speech (TTS) system designed for commercial platforms, with two primary challenges: accommodating diverse acoustic conditions and supporting a large number of customers efficiently. To address these challenges, AdaSpeech employs innovative techniques. It utilizes dual acoustic encoders during pre-training and fine-tuning to capture both utterance-level and phoneme-level acoustic information. It also introduces conditional layer normalization in the mel-spectrogram decoder to balance adaptation parameters and voice quality. With limited adaptation data, AdaSpeech significantly outperforms baseline methods, requiring only around 5,000 specific parameters for each speaker. This demonstrates its efficacy in achieving high-quality and memory-efficient customization of voices. For audio samples, please visit https://speechresearch.github.io/adaspeech/.

**b) Inference** : **Mingjian Chen , Xu Tan et al. (2021)** The system uses two different acoustic encoders to handle varied acoustic conditions: one extracts an utterance-level vector

from the target voice, while the other extracts a series of phoneme-level vectors. Inference uses an auditory predictor to forecast the phoneme-level vectors while deriving the utterance level from a reference voice. Conditional layer normalization is added to the AdaSpeech mel-spectrogram decoder and is modified alongside speaker embedding for adaptation in order to better balance adaptation parameters and voice quality.With little adaptation data available, this fine-tuning procedure often involves 20 sentences or approximately 1 minute of speech.

**2) Efficiently Trainable Text-To-Speech System Based On Deep.**

**a) Abstract**: This paper introduces a novel text-to-speech (TTS) method based on deep convolutional neural networks (CNN), a departure from the commonly used recurrent neural networks (RNN) in recent TTS techniques. RNNs, while effective, often demand substantial computational power and time, sometimes spanning days or weeks for training. In contrast, recent studies have demonstrated that CNN-based sequence synthesis is significantly faster due to its high parallelizability. The main goal of this paper is to showcase that a CNN-based neural TTS system can mitigate the economic costs associated with training. The authors successfully trained their proposed Deep Convolutional TTS in just 15 hours, using a standard gaming PC equipped with two GPUs. Despite the relatively short training time, the quality of the synthesized speech was deemed largely acceptable, highlighting the efficiency and effectiveness of this approach in the realm of TTS.

**b) Inference**: **Hideyuki Tachibana, Katsuya Uenoyama et al 2018** A novel TTS technique based on deep convolutional neural networks, and a technique to train the attention module rapidly. In the experiment, the proposed Deep Convolutional TTS was trained overnight (~15 hours), using an ordinary gaming PC equipped with two GPUs, while the quality of the synthesized speech was almost acceptable. Although the audio quality is far from perfect yet, it may be improved by tuning some hyper-parameters thoroughly, and by applying some techniques developed in the deep learning community.

**3) Espnet-Tts: Unified, Reproducible, And Integratable Open Source End-To-End Text-To-Speech Toolkit**

**a) Abstract**: This paper introduces ESPnet-TTS, an innovative end-to-end text-to-speech (E2E-TTS) toolkit that extends the ESPnet open-source speech processing toolkit. The toolkit includes pre-trained models and recipe samples, making it easy for users to get started and use them as a baseline. One of the notable features is the seamless integration of ASR functions with TTS, enabling ASR-based objective evaluation and semi-supervised learning with both ASR and TTS models. The paper describes the toolkit's design and presents experimental results comparing it with other toolkits. These results demonstrate that ESPnet-TTS achieves state-of-the-art performance, with a mean opinion score (MOS) of 4.25 on the LJSpeech dataset. ESPnet-TTS is available to the public on GitHub at https://github.com/espnet/espnet, making it a

valuable resource for researchers and developers in the field of text-to-speech.

**b) Inference**: **Tomoki Hayashi, Takenori Yoshimura, Tomoki Toda, Kazuya Takeda ,Nagoya University et al (2020)** ESPnet-TTS supports cutting-edge E2E-TTS models, such as Tacotron 2, Transformer TTS, and FastSpeech, while incorporating recipes inspired by the Kaldi automatic speech recognition (ASR) toolkit. These recipes are designed for high reproducibility and are unified with the ESPnet ASR recipe. E2E-TTS systems are more accommodating and further this area of study. The toolbox supports modern E2E-TTS models in addition to numerous TTS recipes whose layout is consistent with ASR recipes, and great repeatability is provided. The findings of the experimental evaluation showed that our models can attain cutting-edge performance equivalent to the other newest toolkits, yielding MOS of on the LJSpeech dataset, 4.25.

**4) Meta-TTS: Meta-Learning for Few-Shot Speaker Adaptive Text-to-Speech**

**a) Abstract**: This paper explores methods for personalizing a speech synthesis system with minimal user voice recordings. Two common approaches are compared: speaker adaptation and speaker encoding. Speaker adaptation fine-tunes a text-to-speech (TTS) model with a few user samples but requires many steps for quality, making it less practical for devices. Speaker encoding encodes user voice into a speaker embedding for TTS. However, it faces challenges with unseen speakers. The paper introduces a novel approach, Meta-TTS, which applies meta-learning (specifically, Model Agnostic Meta-Learning or MAML) to speaker adaptation. Meta-TTS seeks to quickly adapt a multi-speaker TTS model to new speakers by finding an efficient meta-initialization. Experimental results show that Meta-TTS can generate speaker-similar speech with minimal enrollment samples, requiring fewer adaptation steps than traditional methods. It outperforms speaker encoding approaches, even when the latter are pre trained with a larger dataset.

**b) Inference**: **Sung-Feng Huang , Chyi-Jiunn Lin ,Yi-Chen Chen , and Hung-yi Lee et al (2022)** Speaker adaptation involves fine-tuning a multi-speaker text-to-speech (TTS) model with a few enrolled samples. However, this method typically requires a large number of fine-tuning steps for high-quality adaptation, making it less feasible for use on devices. On the other hand, speaker encoding methods encode enrollment utterances into a speaker embedding, allowing the TTS model to synthesize the user's speech based on this embedding. Nevertheless, these speaker encoders face challenges when dealing with unseen speakers. A multi-speaker TTS model's training approach uses Model Agnostic Meta-Learning (MAML), which seeks to quickly identify a great meta-initialization for any few-shot speaker adaptation tasks. A speaker adaptation method baseline and a speaker encoding technique baseline make up the approach (Meta-TTS).

**5) Sentence-Level Emotion Detection from Text Based on Semantic Rules**

**a) Abstract**: This paper delves into the realm of emotion detection from text, a compelling area within natural language processing. Emotion detection involves identifying and categorizing emotions from diverse sources like text, facial expressions, gestures, and speech. The paper introduces an efficient emotion detection technique that relies on a predefined emotional keyword database. This method involves searching for emotional words within the text, analyzing emotion-related words and phrasal verbs, and considering the impact of negation words. The results indicate that this approach outperforms recent methods in the field, suggesting its effectiveness in accurately detecting and categorizing emotions in text.

**b) Inference**: **Dibyendu Seal, Uttam K. Roy and Rohini Basak et al (2020)** Sentiment analysis is a basic form of emotion detection that determines whether the sentiment of the text is positive, negative, or neutral.Techniques range from rule-based systems to machine learning models trained on labeled data.

**6) An effective approach for emotion detection in multimedia text data using sequence based convolutional neural network**

**a) Abstract**: The paper presents a new corpus comprising various emotional expressions extracted from a TV show's transcript. This corpus was manually annotated with the assistance of English experts. The proposed emotion detection framework employs a sequence-based convolutional neural network (CNN) with word embedding, enhanced by an attention mechanism. The attention mechanism enables the CNN to focus on words that have a more significant impact on classification or specific features that require closer attention. The primary goal of this work is to create a framework that can generalize newly collected data, aiding businesses in understanding customer sentiments and facilitating social media monitoring to gauge public opinion on various topics. Experimental results on the dataset demonstrate that the proposed framework effectively detects emotions from text with high precision and accuracy scores, making it a valuable tool in the domain of fine-grained emotion detection.

**b) Inference**: **Kush Shrivastava , Shishir Kumar et al (2019)** Recent trends have ushered in a multimedia era, significantly impacting areas like business, recommendation systems, and information retrieval. While emotion detection from multimedia images and videos has received attention, text data in this context has been relatively understudied. Deep learning has proven superior to traditional methods in sentiment analysis, and this work is inspired by those achievements. It introduces a deep learning framework for fine-grained emotion detection in multimedia text data. Utilize neural networks (e.g., LSTM, GRU) to capture complex patterns and context in the text. Sequence-to-sequence models can predict emotions directly from text inputs.

**7) Emotional Voice Conversion Using Multitask Learning With Text-To-Speech**

**a)** **Abstract**: The paper presents a voice conversion system that leverages multitask learning in conjunction with text-to-speech (TTS) technology. By utilizing multitask learning, this approach is designed to capture and preserve linguistic information while ensuring training stability. Unlike previous methods, it doesn't necessitate explicit alignment to extract abundant text information. The experiments conducted in this study involve voice conversion using a male-Korean-emotional-text-speech dataset, with the objective of converting a neutral voice into an emotional voice. The results indicate that multitask learning plays a significant role in preserving the linguistic content during voice conversion, addressing the limitations of the previous approach.

**b)** **Inference**: **Tae-Ho Kim , Sejik Park , Soo-Young Lee et al (2020)** Voice conversion (VC) is a technology aimed at transforming a person's voice to match different styles while maintaining the underlying linguistic content. Previous state-of-the-art VC methods were based on the sequence-to-sequence (seq2seq) model, but they had limitations in retaining linguistic information. Some attempts were made to address this issue through textual supervision, but this approach required explicit alignment, negating the advantages of the seq2seq model. The model was honed to produce speech based on the input reference for style. The style encoder was also de-advised to remove linguistic components while extracting style information. Without the explicit labeling of an emotion, the feelings were successfully untangled by style encoder.

## 2.2 Inference Drawn

In the realm of addressing web communication challenges for visually impaired individuals, there's a notable absence of comprehensive solutions that integrate Text-to-Speech (TTS), Speech-to-Text (STT), and emotion detection from text. Existing research primarily focuses on isolated aspects of accessibility or individual assistive technologies, without offering a holistic end-to-end solution tailored to the needs of visually challenged users.

While various studies have explored elements such as screen reader compatibility or text-to-speech synthesis, there's a distinct lack of research that combines these functionalities with emotion detection to enhance the user experience further. Moving forward, there's a clear opportunity to bridge this gap by pioneering innovative solutions that seamlessly integrate TTS, STT, and emotion detection within web communication platforms tailored for visually challenged individuals. By incorporating these functionalities will help visually challenged individuals to engage with online communities, fostering inclusivity and empowerment in the digital realm.

## 2.3 Comparison with the Existing System

Table 2 Comparison with the Existing System

| **EXISTING SYSTEMS** | **SOLUTIONS PROVIDED BY PROPOSED SYSTEM** |
|---|---|
| Numerous systems are available that solely offer functionalities encompassing text-to-speech and speech-to-text capabilities. | The proposed system not only integrates text-to-speech (TTS) and speech-to-text (STT) functionalities but also incorporates emotion detection from text. |
| Existing systems face challenges in fully meeting the diverse needs and preferences of visually challenged users, potentially leading to limitations in user customization and interface adaptability. | The project prioritizes the needs and preferences of visually challenged users, offering a user-friendly interface and customizable features to optimize the user experience. |
| One challenge with real-time TTS, STT, and emotion detection capabilities is the potential for increased computational demands and processing delays, which may affect the responsiveness and fluidity of the communication experience for visually impaired users, particularly in situations with limited internet connectivity or older devices. | With real-time TTS, STT, and emotion detection capabilities, the project facilitates dynamic and interactive communication experiences, empowering visually impaired individuals to engage more seamlessly in conversations and social interactions online. |

# Chapter 3: Requirement Gathering for the Proposed System

## 3.1 Introduction to requirement gathering

Requirement gathering is essential for any software or technology project as it helps to ensure that the final product meets the needs and expectations. It helps to define the scope of the project, including the features and functionality that will be included in the final project. Requirement gathering helps to prioritize the requirements based on their importance and impact on the project.

## 3.2 Functional Requirements

1. New user registration: EmoSpeak allows new users to register for the app, providing essential details and preferences required for personalized interaction.
2. Authentication of users: The app authenticates users securely, ensuring that only authorized individuals can access its features.
3. Integration with TTS, STT, and emotion detection: EmoSpeak seamlessly integrates text-to-speech (TTS), speech-to-text (STT), and emotion detection functionalities, enabling visually challenged users to engage in dynamic and emotionally rich conversations.
4. User profile management: Users can manage their profiles, including personal information, preferences, and communication settings.
5. Contact Management: Users are able to add, remove, and manage contacts within the application.
6. Text Messaging: Users are able to send and receive text messages in real-time.
7. Group chat functionality: Group chat functionality is included, allowing users to create and manage group conversations.
8. Emoticons: The application includes a library of emoticons for users to express emotions in their messages.
9. Security and Privacy: The chat application employs encryption mechanisms to ensure the privacy and security of user communications.

## 3.3 Non-Functional Requirements

1. Performance: The application has fast response times and minimal latency, ensuring smooth communication even during peak usage periods.
2. Accessibility: The chat application is accessible to any users.

3. Usability: The chat application has an intuitive and user-friendly interface.

4. Compliance: The chat application complies with relevant legal and regulatory requirements.

## 3.4 Hardware, Software , Technology and tools utilized

### A. Hardware Requirements

1. Processor: Multi-core Processor(2.5 GHz or higher)
2. RAM: Minimum 4GB RAM.
3. Minimum 10 GB Storage Capacity
4. Network Connectivity: Internet Connection

### B. Software Requirements

1. Operating System: Windows 7 or higher, macOS or Linux
2. Flutter Framework.
3. Programming Language-Python 3.8.
4. Java Development Kit
5. Android Studio ver: 2022.1.1
6. Firebase
7. Git
8. Text-To-Speech API
9. Google Cloud Natural Language API
10. Google Colab

### C. Technology and Tools Utilized

1. Flutter Framework: Flutter is an open-source framework developed by Google for building natively compiled applications for mobile, web, and desktop from a single codebase. It's known for its fast development, expressive and flexible UI components, and high performance

2. Python: Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation.

3. Java Development Kit: The Java Development Kit (JDK) is a software package developed by Oracle Corporation (formerly by Sun Microsystems) that provides the necessary tools and resources for developing Java applications.

4. Firebase: Firebase is a comprehensive mobile and web application development platform

developed by Google. It provides a wide range of tools, services, and resources to help developers build high-quality, feature-rich apps.

5. Git: Git is a distributed version control system widely used in software development for tracking changes in source code during the development process.

6. Text-To-Speech API: A Text-to-Speech (TTS) API is an application programming interface that allows developers to integrate TTS capabilities into their applications, websites, or services. TTS technology converts written text into spoken audio, enabling users to listen to content, such as articles, messages, or documents, instead of reading it.

7. Google Cloud Natural Language API: The Google Cloud Natural Language API is a cloud-based service provided by Google that offers advanced natural language processing capabilities. It allows developers to analyze and extract insights from textual content, making it useful for a wide range of applications.

8. Google Colab: Colaboratory, or "Colab" for short, is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser and is especially well suited to machine learning, data analysis and education. More technically, Colab is a hosted Jupyter notebook service that requires no setup to use, while providing access free of charge to computing resources including GPUs.

## 3.5 Constraints

1. Internet access is required.
2. Mapped timing for both the users.

# Chapter 4: Proposed Design

## 4.1 Block diagram of the system

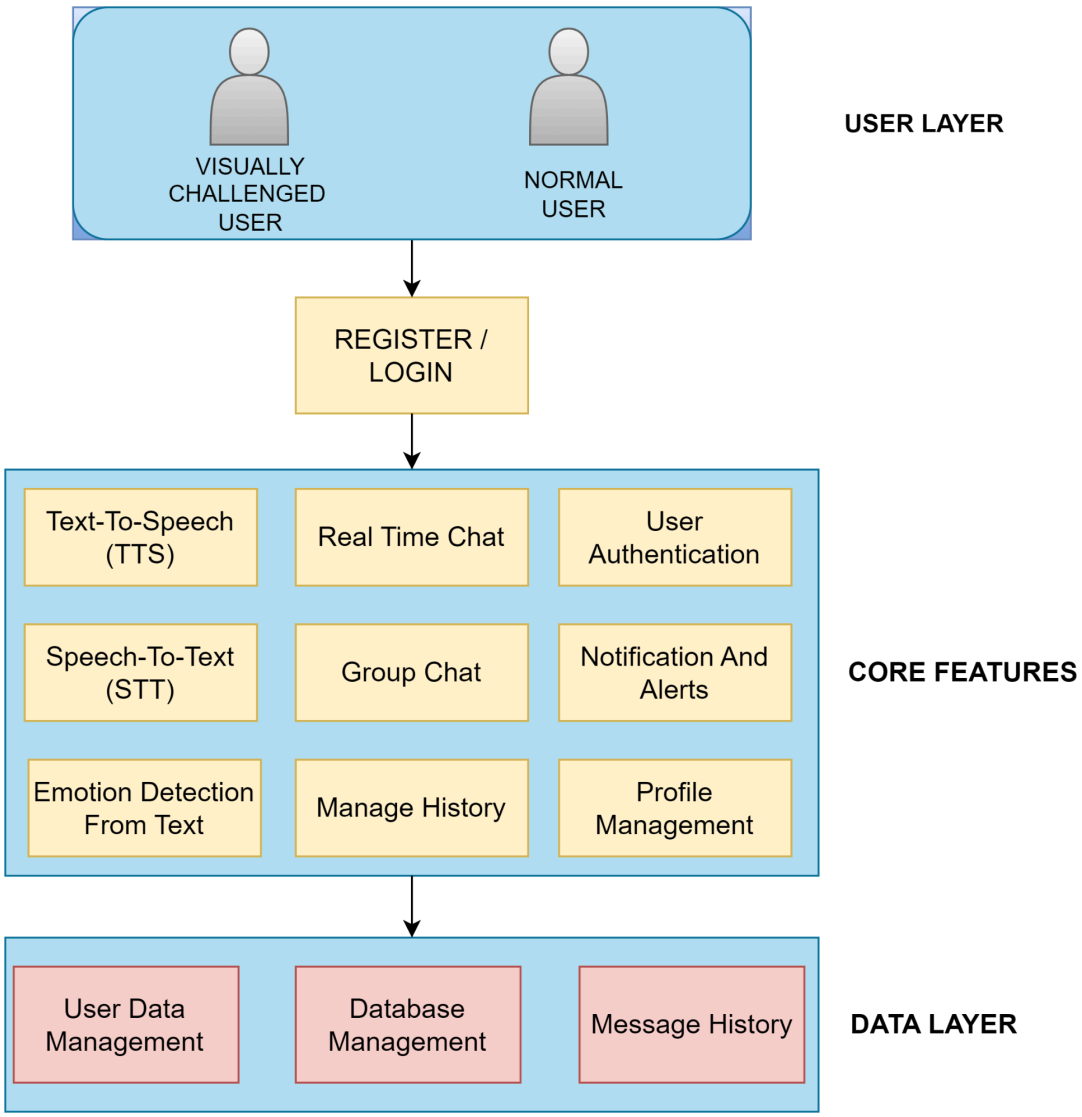



Fig 1 Overview of the System

The block diagram consists of 3 layers:

- User Layer - All the users who can use the system are a part of this layer
- Core Features - This is the core layer which depicts the functionalities of the system
- Data Layer - User's data and history are stored in database

Features:

1. Text-to-Speech (TTS): Converts written text into spoken words.
2. Speech-to-Text (STT): Transcribes spoken words into written text.
3. Emotion Detection from Text: Analyzes text to identify underlying emotional cues or sentiment.
4. Real-Time Chat: Facilitates instant messaging and communication between users in live sessions.
5. Group Chat: Enables multiple users to participate in a single chat conversation simultaneously.
6. Manage History: Stores and organizes past chat conversations for reference and retrieval.
7. Profile Management: Allows users to create, edit, and customize their personal profiles within the application.
8. Notification and Alerts: Notifies users of new messages, updates, or important events through alerts or notifications.
9. User Authentication: Verifies the identity of users accessing the application to ensure secure access and data protection.

## 4.2 Modular Design of the System

1. Text-to-Speech:

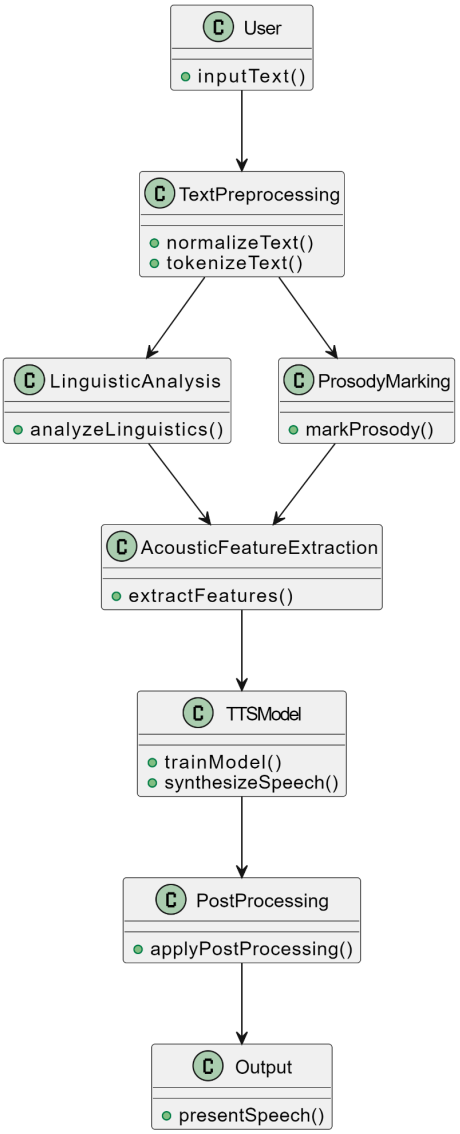



Fig 2 Flowchart of Text-to-Speech Module

1. User Input: In a text-to-speech (TTS) system, "user input" refers to the text or material that users supply for voice synthesis. This input provides the framework for TTS systems to process and produce output that sounds like natural speech. Advanced TTS systems can also take into account user preferences and input context to customize the synthesized speech to specific requirements and emotional nuance.

2. Text Processing: The text input for the synthesizer can be in UTF-8 or transliterated form. Before further processing, input text in UTF-8 format will be translated to the transliterated form.Preprocessing and syllabication modules make up the text processing module. The transliterated text has been preprocessed to eliminate any incorrect characters.

Additionally, depending on full stops and case markers, the preprocessing program adds phrase break signs to the text.

3. Linguistic Analysis: The segmentation of text into sentences and words, the assignment of parts of speech, and syntactic analysis to identify sentence structure and phrasing are all key aspects of text-to-speech (TTS) systems. It also includes the use of suitable pauses and timing for natural speech rhythm, as well as norms for pronunciation, stress and intonation patterns, and prosody. This analysis considers the linguistic context, adjusts for other languages and dialects, and even has the ability to adapt to user preferences and emotional expression. Finally, language analysis makes sure that TTS systems produce coherent, contextually acceptable, and human-like voices from written text.

4. Phenome Generation: In a text-to-speech (TTS) system, phoneme generation entails breaking down written text into a series of phonemes, which are a language's smallest units of sound. To produce genuine and contextually appropriate speech, this procedure involves segmenting the text, employing phoneme dictionaries, applying grapheme-to-phoneme conversion rules, and taking prosody and intonation patterns into account. Modern TTS systems can also take into account emotional nuances in speech. The synthesized speech is then rendered audibly using the created phoneme sequence, resulting in a lifelike, understandable, and emotionally expressive sound. A vital stage in producing high-quality TTS output is phoneme creation.

5. Prosody Modeling: The prosody module predicts prosodic characteristics for the chosen syllables, including pitch, duration, stress, emotions, intensity, etc. The prosody with which voice actors read the prompts during recording varies over the course of the recording. Syllables chosen for concatenation are also chosen from various contexts. These factors cause the synthesized speech to have audible discontinuity brought on by irregular prosodic contours. Classification and Regression Tree (CART) is used to forecast prosody for the chosen units in order to rectify these prosodic contours. Syllable, word, and phrase levels, among others, are three levels at which prosodic qualities can be classified. For instance, vowels are more intense than consonants at the word level. Correct prosody is harder to produce at the phrase level than at the sentence level.

6. Waveform Generation: The act of generating the audible speech signal from intermediate representations, such as phoneme sequences or acoustic parameters, is known as waveform generation in text-to-speech (TTS) systems. To create the speech waveform, either concatenative synthesis—which chooses and combines pre-recorded speech units—or parametric synthesis—manipulates acoustic characteristics. To ensure genuine and emotionally expressive speech, consideration is given to prosody, intonation, voice qualities, and emotional expressiveness. Achieving high-quality and human-like TTS output is mostly dependent on the

waveform generation technology chosen and parameter fine-tuning.

7. Generated Speech Output: The finished product of the synthesis process is the generated voice output in a text-to-speech (TTS) system, which represents the transformed text as realistic speech. This output can be in the form of synthetic speech waveforms or voice recordings that sound like people. To produce speech that is understandable, contextually suitable, and, if desired, emotionally expressive, it combines linguistic analysis, phoneme production, prosody, voice characteristics, and emotional expressiveness. The user experience is impacted by the generated speech output quality in TTS applications including virtual assistants, accessibility aids, and content production.

2. Speech-to-Text:

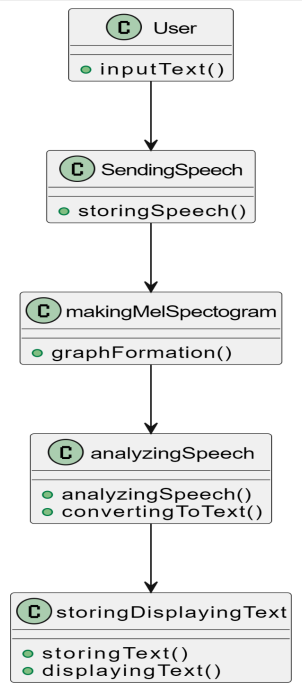



Fig 3 Flowchart of Speech-to-Text Module

1. Speech Input (Audio): The voice input audio in a speech-to-text (STT) system is the initial audio signal or spoken content supplied by the user. To translate the spoken words into written text, this audio input is processed through a number of stages that include signal processing, acoustic analysis, phoneme recognition, and language modeling. The STT system's dependability and effectiveness are significantly impacted by the precision and caliber of the voice input processing. STT systems enable users to translate spoken language into written text for a variety of purposes and find applications in transcription services, voice assistants, accessibility tools, and more.

2. Audio Preprocessing: Speech-to-text (STT) systems use audio processing to convert

spoken language from an audio input signal into written text. There are various steps involved in this process, including audio preprocessing to improve quality, feature extraction to record acoustic properties, acoustic and language modeling to distinguish phonemes and contextual information, and post-processing to polish the result. With applications in transcription services, voice assistants, closed captioning, and more, the final result is a transcribed textual representation of the spoken content that makes spoken language accessible and searchable in written form. Performance of STT systems is greatly influenced by the caliber and precision of audio processing.

3. Feature Extraction: The process of turning audio input into acoustic features that capture the spectral and temporal properties of spoken content is known as feature extraction in a speech-to-text (STT) system. Processes including frame-based analysis, spectral analysis, and the computation of features like Mel-frequency cepstral coefficients (MFCCs) are used to do this. These characteristics capture the core of the audio signal and act as input for speech recognition stages that follow. The selection of features during feature extraction, a crucial first step in STT, has a significant impact on the system's capacity to accurately convert spoken language into text, which affects applications like transcription services, voice assistants, and more.

4. Acoustic Model: In a speech-to-text (STT) system, an element known as an acoustic model learns to associate phonemes or subword units with acoustic properties, which are commonly acquired from audio signals. The ability to identify specific speech sounds and patterns in the audio depends heavily on this model. It's essential for precise speech transcription and recognition in STT applications. The model learns the connections between acoustic variables and phonetic units by being trained on sizable datasets of audio and related transcriptions. The likelihood scores for various phonetic units are provided by the acoustic model during recognition, and these values are then combined with language models and other components to convert spoken language into written text. The acoustic model's performance and training data quality have a considerable impact on the accuracy and performance of the STT system.

5. Language Model: A language model is an essential part of a speech-to-text (STT) system that aids in determining the likelihood of word sequences and contextual information. In order to recognize words and sentences based on the audio signal's phonemes, this model is essential. By taking into account grammatical patterns, word probabilities, and linguistic context, it helps in the transcription of spoken language into written text. The technology increases transcription accuracy and contextually appropriate voice recognition by introducing a language model into STT. The effectiveness of the STT system is significantly influenced by the standard of the language model, the training data, and its comprehension of contextual data.

6. Decoding: Decoding is the process of choosing the most likely word order or transcription in a speech-to-text (STT) system based on the data from the acoustic model and the language model. It's a crucial stage in the STT pipeline when the system compares linguistic data with acoustic data from the audio input to produce the final transcribed text. Decoding algorithms employ a variety of methods, including dynamic programming and neural networks, to identify the ideal word order. Decoding establishes the recognized words or phrases, which has a direct impact on the precision and caliber of the STT system's output. The effectiveness of the decoding process and the method chosen have a big impact on the STT system's overall performance.

7. Text Output (Transcription): The final step in the conversion of the identified words and phrases from the audio input into written text in a speech-to-text (STT) system is text output transcription. The STT process produces a transcription that can be used in a variety of contexts, including closed captioning, voice assistants, and transcription services. The effectiveness of the STT system for making speech accessible and searchable depends on the quality and accuracy of text output transcription, which is essential for ensuring that the spoken language is appropriately and contextually reproduced in written form.

3. Emotion Detection From Text Model:

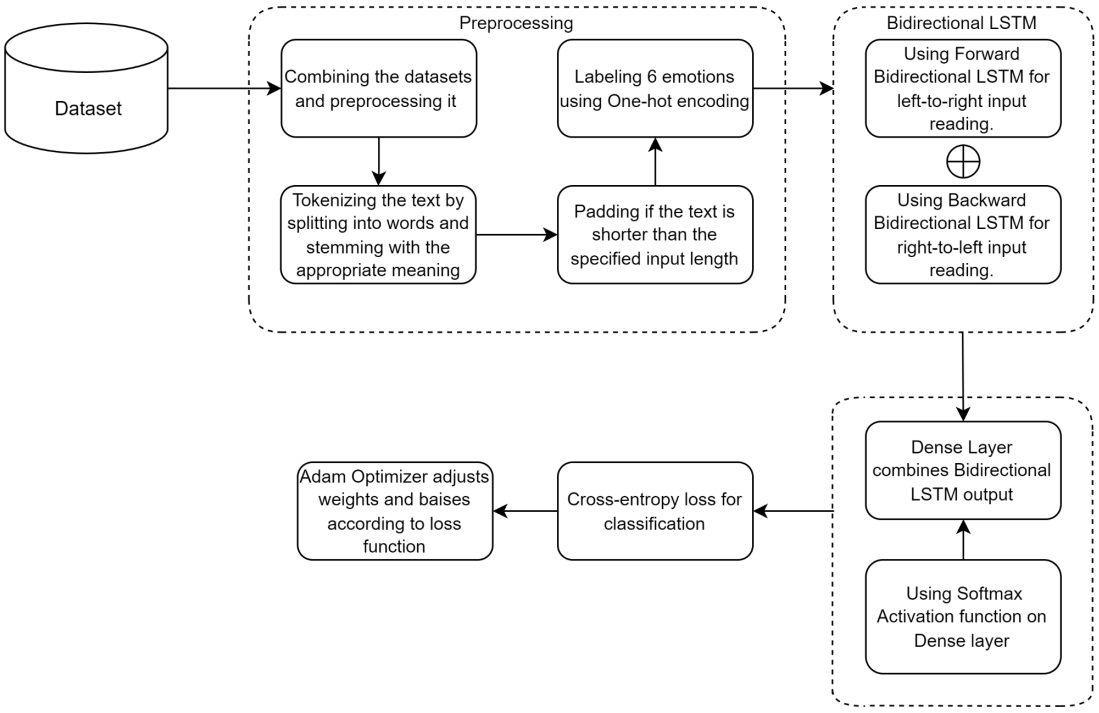



Fig 4 Flowchart of LSTM Emotion Detection from Text Model

1. The inputs taken from the datasets are then preprocessed in the Embedding Layer. Preprocessing in the LSTM approach entails partitioning the dataset into two segments followed

by their amalgamation for tokenization, where text is converted into tokens, with each token representing a single word, and stemming is applied to emphasize word semantics. After preprocessing, each word in the text is likely converted to a numerical index based on a vocabulary.

2. The index represents the word's position in the vocabulary list. The embedding layer takes these word indices as input and looks up their corresponding embedding vectors in the embedding matrix. Subsequently, padding and One-hot-Encoding techniques are applied to these embedding vectors, containing semantic information about the words, and become the actual input to the bidirectional LSTM layers in the model. Embedding matrix: W ∈ R^(16000×100) where W is Word embedding matrix, 16000 is the vocabulary size and 100 is the embedding dimension.

3. The input is transferred to the bidirectional LSTM layer which is the core of the model responsible for learning long-range dependencies in the text sequence. It processes the input sequence in both forward and backward directions using two separate LSTMs. In a Forward LSTM, the sequence is processed from left to right, allowing it to capture the context preceding each word. Conversely, in a Backward LSTM, the sequence is processed from right to left, enabling it to capture the context succeeding each word. Following is the equation of Bidirectional LSTM:

$$h\_lstm = f\_LSTM(h\_embedding, h\_prev)$$

where h_lstm is the combined hidden state from both LSTMs, f_LSTM represents a function encapsulating the LSTM computations, and h_prev is the hidden state from the previous time step (or initial hidden state for the first step).

4. The output of both the LSTMs is concatenated and is transferred to the Dense Layer. Linear transformation of the final hidden states from both LSTMs:

$$z = W\_dense * h\_lstm + b\_dense$$

where W_dense is the weight matrix for the dense layer, combining information from the final hidden states of both forward and backward LSTMs (h_lstm) and b_dense is the bias vector for the dense layer.

5. Subsequently, the output of the Dense layer undergoes the application of the Softmax activation function. This function normalizes the values, ensuring they sum to 1 and represent probabilities for each class. These probabilities indicate the likelihood of the input text

belonging to each of the possible classes. The equation of Normalized probability scores using softmax is

$$a = softmax(z)$$

6. In order to measure the disparity between the model's predictions and the actual values of the target variable, a loss function is employed. In this instance, the chosen loss function is categorical cross-entropy. By minimizing categorical cross-entropy, the model is trained to allocate higher probabilities to the correct classes for each input while reducing probabilities assigned to incorrect ones. This process enhances the model's capacity to precisely predict the class to which an unseen text sample belongs. The equation representing categorical cross-entropy loss, which compares predicted probabilities with true labels, is as follows:

$$L = categorical\_cross\ entropy(y\_true, a)$$

7. However, in machine learning, the goal is to minimize the lost function. Optimizers are algorithms that iteratively update the model's internal parameters (weights and biases) to gradually decrease the loss function. They adjust the parameters in the direction that leads to a steeper decrease in the loss.The Adam Optimizer is implemented with a learning rate set to 0.01 (lr=0.01). This process continues until the model reaches a minimum point in the loss function, hopefully representing a good fit to the data. Hence, the model outputs a vector of probabilities, with each component representing the probability of the input text belonging to a specific class in the multiclass classification problem. The predicted class is typically determined as the one with the highest probability, thus enabling accurate detection of the emotion.

8. Hence, to understand the emotion implied in the message, the 'EmoSpeak' chat application integrates a Text-to-Speech feature powered by a Machine Learning LSTM Model. Upon converting the text message into audio format, it identifies the underlying emotion, facilitating smoother user comprehension and communication.

# 4.3 Project Scheduling & Tracking using Timeline / Gantt Chart

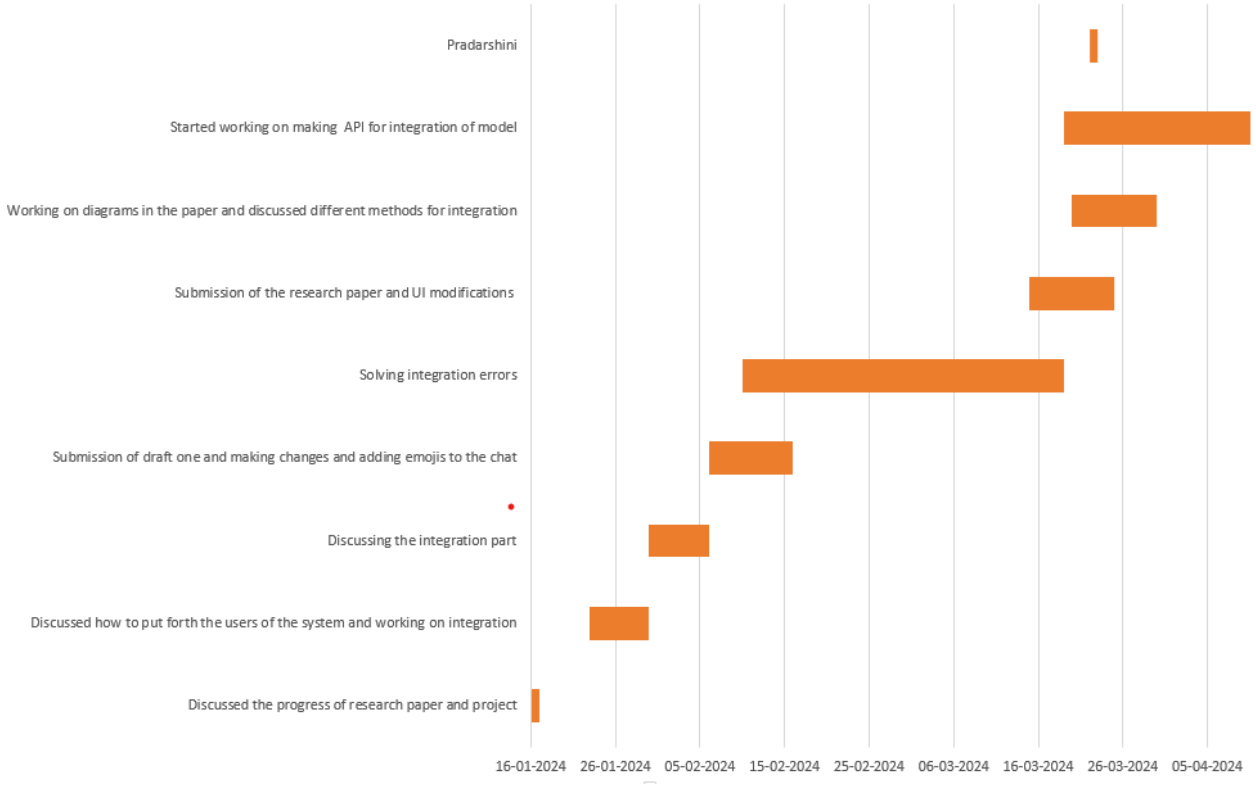



Fig 11 Gantt Chart

# Chapter 5: Implementation of the Proposed System

## 5.1 Methodology employed for development

Many solutions have been put forth in the field of helping the blind, ranging from Text-to-Speech (TTS) to Speech-to-Text (STT) systems. Still, an all-inclusive system that smoothly links the sighted and the individuals having low vision is necessary for efficient communication. This paper presents a solution called "EmoSpeak," which combines a Text-to-Speech (TTS) module with emotion identification from text. This system would combine cutting-edge technology with user-friendly interfaces, encouraging real connections and enabling the blind to easily connect with the outside world. It has great potential to improve the sociability and engagement of people with visual impairments. This function enables visually challenged persons to interact with digital content more successfully and to understand the subtle emotional aspects of text-based communication, leading to more meaningful conversations and facilitating greater inclusivity in social settings.

The comprehensive architecture of “EmoSpeak” comprises multiple levels and components tailored to meet the needs of stakeholders, including visually impaired and sighted users. The initial step involves user registration followed by login to initiate application use. Core features such as Text-to-Speech, Speech-to-Text, Emotion detection from text, real-time and group chat enhance the user experience. Additionally, the system includes user profile management, notifications, alerts, user authentication, and comprehensive profile customization. Access to user data, message histories, and database administration is seamless and reliable through the data layer. Lastly, the architecture prioritizes flawless communication and interoperability to empower visually impaired individuals in navigating the digital realm and fostering genuine connections with sighted users.

## 5.2 Algorithms and flowcharts for the respective modules developed

1. Text-to-Speech:

This algorithm represents a simplified interaction between the Flutter TTS package and the Android Text-to-Speech Engine (Google TTS) within a Flutter app. The Flutter TTS package is used to set up, configure, and trigger text-to-speech conversion, and the Android Text-to-Speech Engine (Google TTS) is responsible for synthesizing the text and generating audible speech output.

2. Speech-to-Text:

This algorithm outlines the steps for implementing speech-to-text conversion in a Flutter app:

- Initialize the Speech-to-Text plugin, which provides speech recognition capabilities.
- Check if speech recognition is available on the device using the isAvailable property of the plugin.
- Start listening for speech input using the listen method, specifying callbacks for handling recognition results and errors. The app listens for spoken words and converts them into text.
- Handle recognition results when speech input is recognized. The recognized text is extracted from the result, and can process it as needed.
- Handle recognition errors that may occur during the speech recognition process. Common errors include permission issues or device limitations.
- Optionally, provide user interface elements to initiate and stop speech recognition. This allows users to control when speech input is recognized.
- Call the perform_speech_to_text function to initiate speech-to-text conversion in the app.

3. Emotion Detection from text:

a. Text Preprocessing:

This section of the code snippet primarily focuses on data preprocessing and exploration. Initially, it maps numeric labels representing emotions to their corresponding emotion categories using a predefined dictionary. Subsequently, it groups the training data based on these emotion categories and generates a count of rows for each emotion, providing insight into the distribution of emotions within the dataset. Visual representation of this distribution is achieved through a bar chart. Additionally, the code checks for missing values across the training, validation, and test datasets, highlighting any potential data quality issues. Finally, the text data from all datasets is combined into a single list, facilitating further preprocessing or analysis. Overall, these steps lay the foundation for subsequent tasks, such as text tokenization and model training, by ensuring the integrity and completeness of the dataset.

b. Tokenization:

This section of the code snippet focuses on text data preprocessing and vocabulary processing. Initially, a Tokenizer is instantiated and trained on the combined text data from all datasets. This Tokenizer is used to tokenize the text data and generate a vocabulary index. Subsequently, a Porter Stemmer is employed to stem the words in the vocabulary, reducing them to their root form. A new Tokenizer is then initialized and trained on the stemmed words, resulting in a vocabulary index for the stemmed words. The preprocessing function defined thereafter splits each text sample into words, applies stemming, tokenizes the stemmed words using the second

Tokenizer, and converts them into sequences of tokens. The preprocessed data is structured as lists containing token sequences and their corresponding labels. These steps collectively ensure that the text data is prepared and tokenized effectively for subsequent processing and model training tasks.

c. Prepare Data for Model Training:

In this section, the preprocessed data is further manipulated to prepare it for model training. The `preprocess_data()` function is applied to both the training and validation datasets, transforming the text samples into token sequences and their corresponding labels. Examples of text samples before and after preprocessing are displayed for illustration. Subsequently, the data is split into features (`train_X`, `val_X`) containing token sequences and labels (`train_y`, `val_y`) representing emotion categories. The longest sequence length among the training data is determined, and both `train_X` and `val_X` are padded with zeros to ensure uniform input length for the model. Finally, the data is converted into NumPy arrays, and the labels are one-hot encoded to facilitate model training. These steps collectively prepare the data for effective utilization in training the LSTM model for emotion classification.

d. Model Training:

This section of the code snippet defines a sequential deep learning model using Keras for text classification tasks, particularly for emotion detection from text data. The model architecture comprises an Embedding layer, which converts integer-encoded words into dense vectors to capture semantic relationships between words. This layer is followed by a Bidirectional LSTM layer that processes input sequences bidirectionally to capture temporal dependencies effectively. The model outputs probabilities for multiclass classification using a Dense layer with a softmax activation function. After defining the model architecture, it is compiled with appropriate loss and optimization functions. Subsequently, the model is trained on the preprocessed training data for 25 epochs while monitoring performance on the validation data. Finally, the model summary is printed to provide an overview of its architecture and parameters. This sequential model configuration aims to effectively capture and classify emotions expressed in textual data.

## 5.3 Datasets source and utilization

In the Emotion Detection Model, the dataset comprises English Twitter messages distributed across three CSV files, namely training, testing, and validation sets, containing 16,000, 2,000, and 2,000 instances, respectively. Each message is annotated with one of six emotion categories: sadness, joy, love, anger, fear, and surprise. These datasets serve as input for emotion detection from text employing a Long Short-Term Memory (LSTM) approach. Initially, emotions are categorized as specified in TABLE 3.

Table 3 Distribution Of Input Text According To Various Emotional Categories

| Label | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Emotion | Sadness | Joy | Love | Anger | Fear | Surprised |
| No. of Sentences | 4666 | 5362 | 1304 | 2159 | 1937 | 572 |

# Chapter 6: Results and Discussion

## 6.1 Screenshots of User Interface (UI) for the respective module

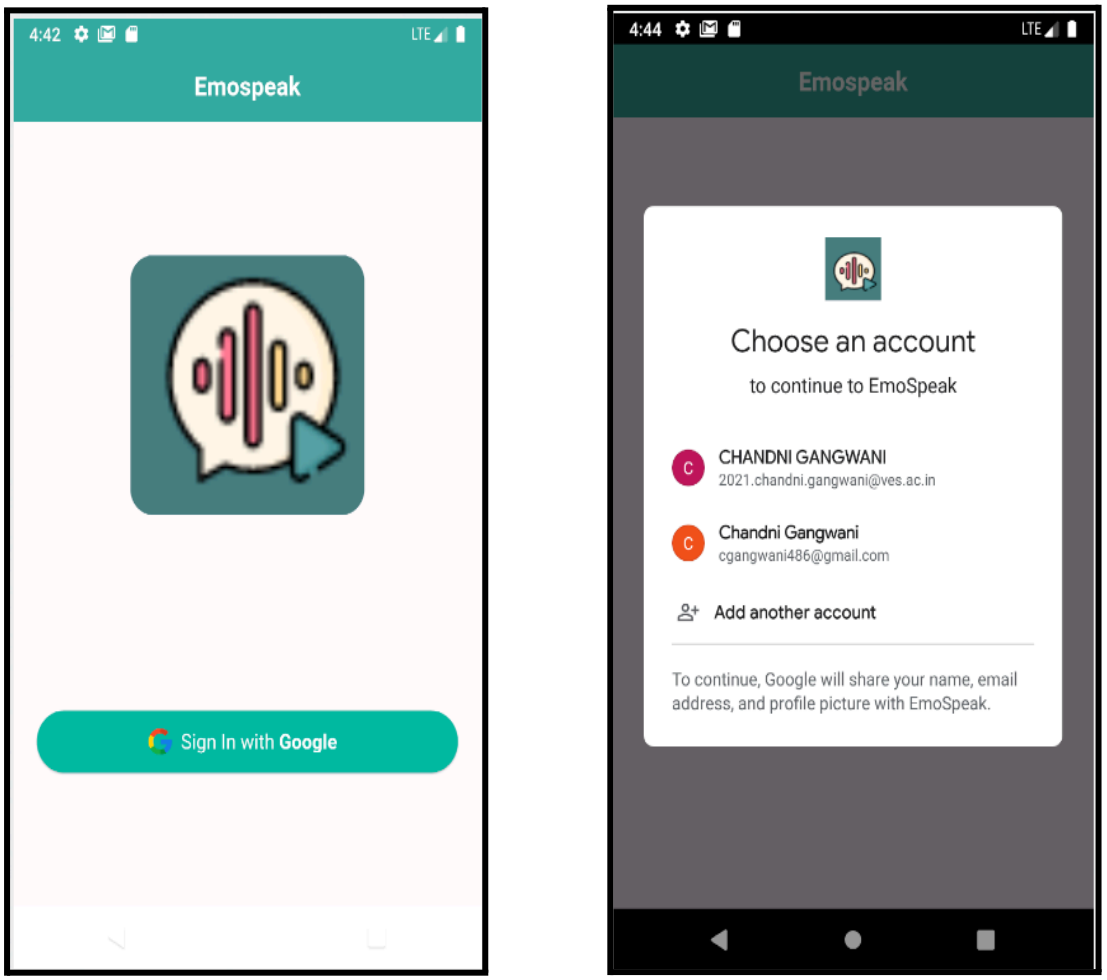



Fig 5 Sign-in page and Login Interface of EmoSpeak application

Fig 5 presents two screenshots showcasing the EmoSpeak mobile application, starting with the sign-in page and the login interface. It provides users with the choice to pick an account for entry. Beneath these options lies a button labeled "Add another account" for incorporating additional accounts. At the bottom of the screen, informative text assures users that Google will share their name, email address, and profile picture with EmoSpeak to proceed.

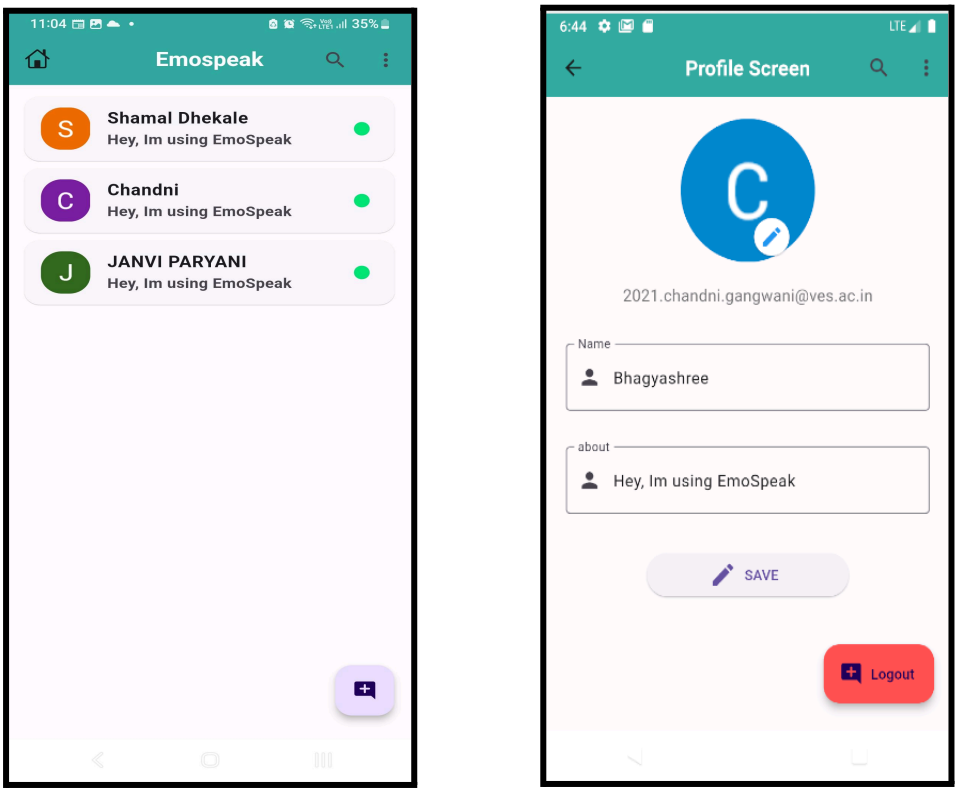



Fig 6 User selection chat interface and Profile screen for users

Fig 6 represents the interface that allows users to select a specific user with whom they wish to engage in a chat. Once a user is selected, the system opens an individual chat window dedicated to that particular user. The second screenshot depicts the user profile screen where users can manage their personal information, preferences, and settings. It provides a comprehensive overview of the user's account, including their profile picture, name, about, and a logout button.

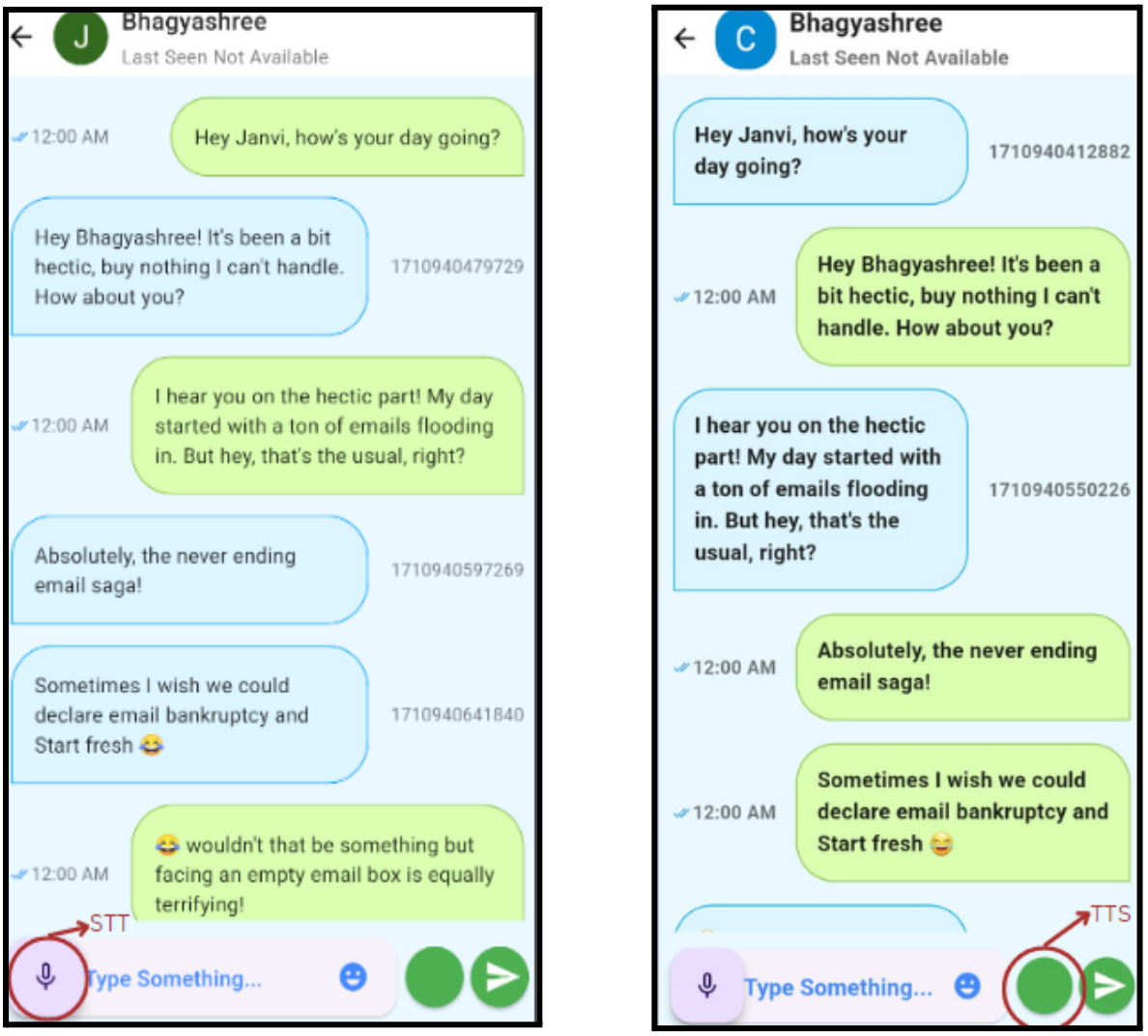



Fig 7 Speech-to-Text (STT)and  Text-to-Speech(TTS) enabled chat functionality.

Fig 7 displays two screenshots of converting the message in Text-to-Speech and Speech-to-text. The first screenshot depicts the action of the user tapping on the microphone button to initiate speech input. The spoken words are then automatically transcribed into text, ready to be sent to the recipient user. In the second screenshot, upon receiving the message, the recipient clicks on the green button, prompting the generation of audio of the latest message received, simultaneously detecting the emotion expressed.

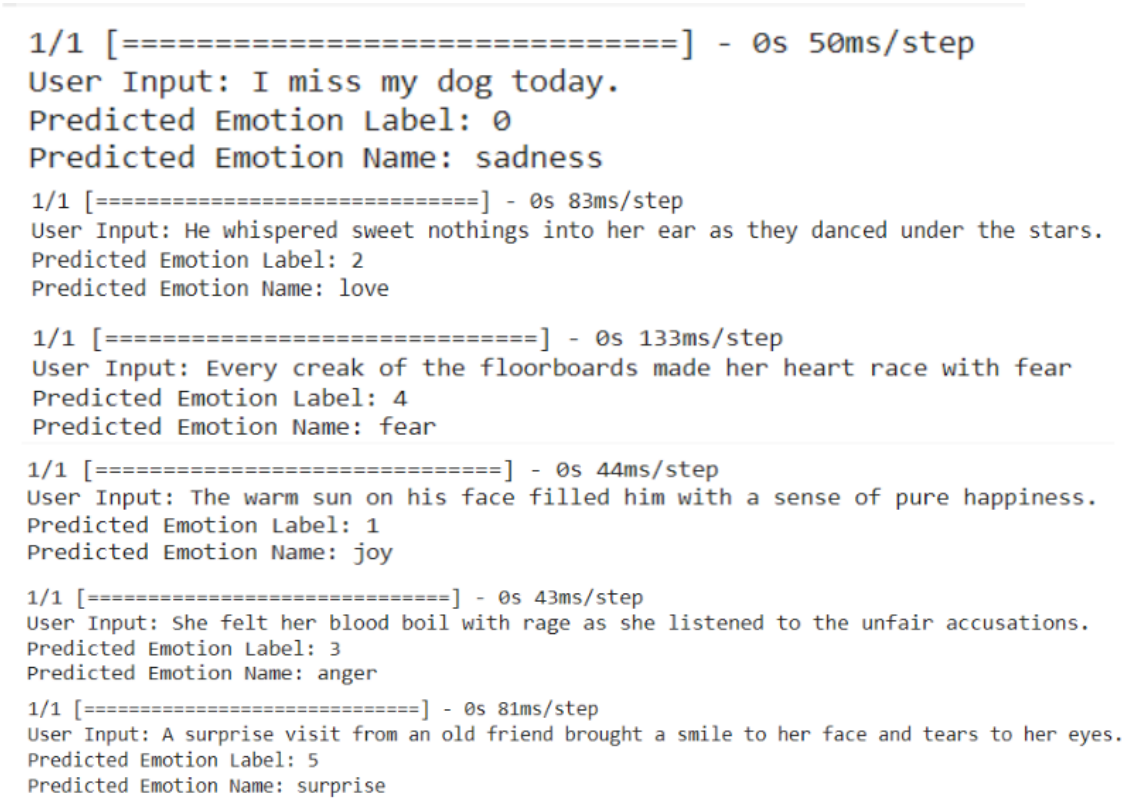



Fig 8 Examples on Emotion Detection from Text

Figure 8 showcases six sentences, each capturing a distinct emotion, accurately detected by the model.

## 6.2 Performance Evaluation measures

A. Determination of efficiency:

Precision is defined as the ratio of correctly classified positive samples (True Positive) to the total number of classified positive samples (either correctly or incorrectly).

$$\text{Precision} = TP / (TP+FP)$$

B. Determination of accuracy:

Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right. Formally, accuracy has the following definition:

$$\text{Accuracy} = (TP+TN) / ( TP+FP+TN+FN)$$

C. Report on sensitivity(recall) analysis:

Recall of positive class is also termed sensitivity and is defined as the ratio of the True Positive to the number of actual positive cases. It can intuitively be expressed as the ability of the classifier to capture all the positive cases. It is also called the True Positive Rate (TPR).

$$\text{Recall} = TP / ( TP+FN)$$

D. Determination of F1-score:

The F1 score is the harmonic mean of precision and recall. It provides a single metric that balances between precision and recall. It is calculated as the weighted average of precision and recall, where the F1 score reaches its best value at 1 and worst at 0.

$$\text{F1 Score} = \text{Precision+Recall} / 2\times\text{Precision}\times\text{Recall}$$

## 6.3. Input Parameters / Features considered

Input parameters consist of the text message sent by the user. This text message is then classified into six categories of emotions joy, sadness, happy , anger , fear ,love and surprise

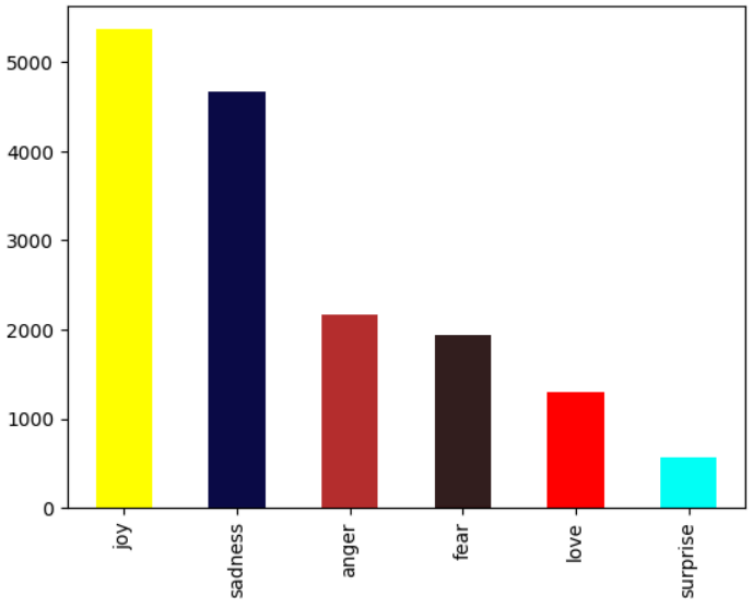



Fig 9 Features considered and their count in the dataset

## 6.4. Graphical and statistical output

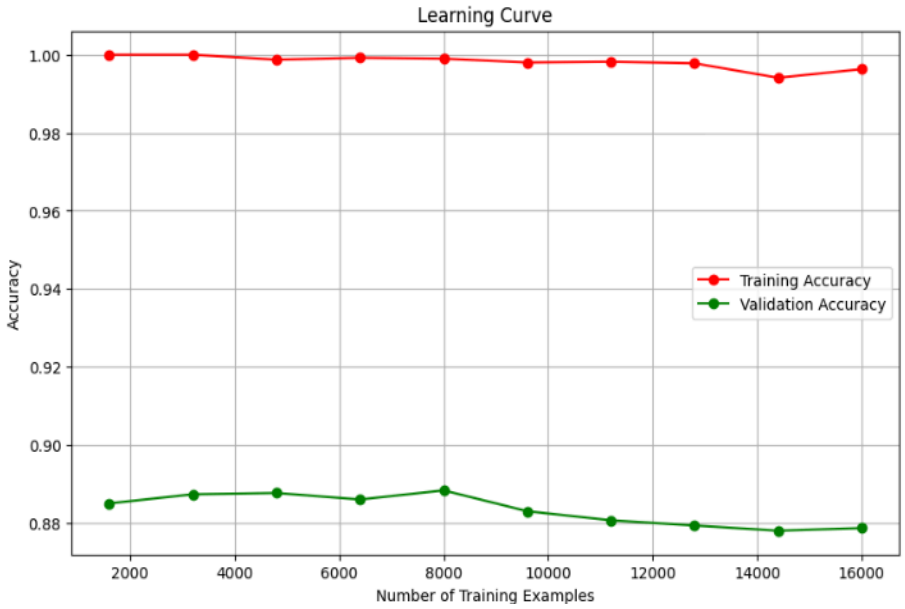



Fig 10 Learning Curve of Number of training examples vs Accuracy

Fig 9 displays the graph of a learning curve. It shows how well the model performs on a task as the number of training examples increases. On the graph, the x-axis corresponds to the number of training examples, while the y-axis denotes accuracy.There are two lines on the graph - the red line and the green line. The red line represents the accuracy of the model on the training data and the green line represents validation accuracy of the model on a separate set of data that the model has not been trained on. Here, the validation accuracy is around 0.88-0.90, that is, around 88%-90% accuracy which indicates how well the model is performing on the unseen data.

Table 4 Performance Metrics For Emotion Detection Model

| Class | Emotion | Parameters | | |
|---|---|---|---|---|
| | | Precision | Recall | F1-score |
| 0 | Sadness | 1.00 | 1.00 | 1.00 |
| 1 | Joy | 1.00 | 1.00 | 1.00 |
| 2 | Love | 0.99 | 1.00 | 0.99 |
| 3 | Anger | 1.00 | 1.00 | 1.00 |
| 4 | Fear | 0.99 | 0.99 | 0.99 |
| 5 | Surprised | 0.98 | 0.99 | 0.99 |

The high precision, recall, and F1-score values showcased in Table 4 across all classes indicate exceptional performance, underscoring the model's effectiveness in diverse categories. With an overall accuracy of 0.9963125, the model correctly classifies approximately 99.63% of the instances, further emphasizing its robustness and reliability.

## 6.5. Inference drawn

- From all the previous published papers, we have observed that there is no such project which provides a complete end-to-end solution. There is just a comparison between various Machine Learning algorithms to improve the accuracy of predicting emotions from the text.
- We took into consideration all the related work and proposed one such solution which will help visually challenged users to communicate easily with the outside world.
- The second part of the project consists of an implementation of detecting emotion from text which will help visually challenged users to understand the context behind the texts.

# Chapter 7: Conclusion

## 7.1 Limitations:

1. Speech Recognition Accuracy: The accuracy of STT systems may vary depending on factors like background noise, speech clarity, and accents. Visually challenged users may face difficulties in articulating words clearly, leading to lower accuracy in speech recognition. This could result in inaccurate transcription of spoken messages, affecting the overall performance of the system.
2. Emotion Detection Reliability: Emotion detection from text using machine learning models relies heavily on the quality and relevance of the input data. The effectiveness of the LSTM model in accurately detecting emotions depends on the diversity and representativeness of the training dataset. Limited training data or biases in the dataset may lead to misinterpretation or misclassification of emotions, impacting the reliability of the system.
3. Natural Language Understanding: Understanding and interpreting natural language, especially in conversational contexts, pose significant challenges. The LSTM model may struggle with understanding nuanced or ambiguous expressions, slang, or colloquial language commonly used in informal conversations. This could result in misclassification of emotions or miscommunication between users.
4. Real-Time Processing Constraints: Real-time processing of speech input, transcription, emotion detection, and TTS output requires efficient handling of computational resources. Complex LSTM models for emotion detection may demand significant computational power and memory, leading to latency issues or performance degradation, particularly on resource-constrained devices or networks.
5. Privacy and Security Concerns: Speech data collected for STT purposes raises privacy and security concerns, especially for visually challenged users who heavily rely on voice interactions. Ensuring the confidentiality and integrity of user data, including spoken messages and emotion profiles, is crucial to maintain trust and compliance with privacy regulations.
6. User Interface Accessibility: Designing an intuitive and accessible user interface for visually challenged users poses additional challenges. The chat application must be compatible with screen readers, support voice commands, and provide alternative navigation methods to ensure a seamless user experience for individuals with visual impairments.

## 7.2 Conclusion:

To sum up, the creation of EmoSpeak, a chat program designed for people with visual impairments, tackles the major obstacles these people encounter when using technology, especially when text-based communication is involved. EmoSpeak provides a comprehensive solution to improve communication accessibility and emotional expression for this demographic by integrating Speech-to-Text and Text-to-Speech functions with an LSTM architecture-based Emotion Detection Model. The LSTM model classifies sentiments into six unique emotions by accurately interpreting the emotional context of text by capturing sequential dependencies in the data. Visually impaired people may now interact more meaningfully and inclusively thanks to EmoSpeak, which bridges the gap between text-based messaging platforms and their specific needs. This creative method encourages emotional intelligence and connectedness in everyday life in addition to giving people the ability to speak on their own.This feature framework can be integrated into existing chat applications and can make them more powerful.

## 7.3 Future Scope:

EmoSpeak's future evolution entails two main objectives: broadening language support to serve diverse global users and refining its user interface for enhanced accessibility. Through language expansion, EmoSpeak aims to foster inclusivity, while ongoing user interface optimization, driven by user feedback and usability studies, ensures continued user-friendliness for visually impaired individuals. These endeavors reinforce EmoSpeak's position as a pivotal tool for accessible and emotionally expressive communication.

# References

[1] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized Affect Representations for Emotion Recognition. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

[2] H. Tachibana, K. Uenoyama and S. Aihara, "Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 4784-4788, doi: 10.1109/ICASSP.2018.8461829.

[3] T. Hayashi et al., "Espnet-TTS: Unified, Reproducible, and Integratable Open Source End-to-End Text-to-Speech Toolkit," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 7654-7658, doi: 10.1109/ICASSP40776.2020.9053512.

[4] S. -F. Huang, C. -J. Lin, D. -R. Liu, Y. -C. Chen and H. -y. Lee, "Meta-TTS: Meta-Learning for Few-Shot Speaker Adaptive Text-to-Speech," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 1558-1571, 2022, doi: 10.1109/TASLP.2022.3167258.

[5] D. Seal, U. K. Roy, and R. Basak, “Sentence-Level Emotion Detection from Text Based on Semantic Rules,” Advances in Intelligent Systems and Computing, pp. 423–430, Jun. 2019, doi: 10.1007/978-981-13-7166-0_42.

[6] Shrivastava, K., Kumar, S. & Jain, D.K. An effective approach for emotion detection in multimedia text data using sequence based convolutional neural networks. Multimed Tools Appl 78, 29607–29639 (2019).

[7] K. Patil, A. Kharat, P. Chaudhary, S. Bidgar and R. Gavhane, "Guidance System for Visually Impaired People," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 2021, pp. 988-993, doi: 10.1109/ICAIS50930.2021.9395973.

[8] A. Karkar and S. Al-Maadeed, "Mobile Assistive Technologies for Visual Impaired Users: A Survey," 2018 International Conference on Computer and Applications (ICCA), Beirut, Lebanon, 2018, pp. 427-433, doi: 10.1109/COMAPP.2018.8460406.

[9]Khan, Akif, Shah Khusro, and Iftikhar Alam. "BlindSense: An Accessibility-inclusive Universal User Interface for Blind People." Engineering, Technology & Applied Science Research 8.2 (2018).

[10] V. Raval and A. Shah, “icuean iot application for indian currency recognition in vernacular languages for visually challenged people,” in 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence. IEEE, 2017, pp. 577–581.

[11] A. Vashistha, P. Sethi, and R. Anderson, “Bspeak: An accessible voice-based crowdsourcing marketplace for low-income blind people,” in Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM, 2018, p. 57.

[12] S. Malgaonkar, P. Shah, D. Panchal, and S. Pradhan, “Awaaz: A bridge between android phones and the visually impaired,” International Journal of Computer Applications, vol. 134, no. 1, pp. 42–47, 2016.

[13] G. Lee, L. C. Quero, J. Yang, H. Jung, J. Son, and J. Cho, “Slate master: a tangible braille slate tutor for mobile devices,” in Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services. ACM, 2017, p. 108.

[14] N. K. Dim, K. Kim, and X. Ren, “Designing motion marking menus for people with visual impairments,” International Journal of HumanComputer Studies, vol. 109, pp. 79–88, 2018.

[15] A. S. Martinez-Sala, F. Losilla, J. C. Sanchez-Aarnoutse, and J. Garc ´ ´ıaHaro, “Design, implementation and evaluation of an indoor navigation system for visually impaired people,” Sensors, vol. 15, no. 12, pp. 32 168–32 187, 2015.

# Research Paper Details

## a. Paper published

# EmoSpeak: An Emotionally Intelligent TTS System for Visually Impaired


Yugchhaya Galphat
*Computer Department*
*Vivekanand Education Society's Institute of Technology*
Mumbai, India.
yugchhaya.dhote@ves.ac.in

Bhagyashree Vaswani
*Computer Department*
*Vivekanand Education Society's Institute of Technology*
Mumbai, India.
2021.bhagyashree.vaswani@ves.ac.in

Chandni Gangwani
*Computer Department*
*Vivekanand Education Society's Institute of Technology*
Mumbai, India.
2021.chandni.gangwani@ves.ac.in

Shamal Dhekale
*Computer Department*
*Vivekanand Education Society's Institute of Technology*
Mumbai, India.
2021.shamal.dhekale@ves.ac.in



***Abstract*—The paper delves into the obstacles confronting individuals with visual impairments, especially those experiencing blurred vision due to medical conditions or aging, when navigating technology, notably chat services. People with weak eyesight face issues such as inaccessible user interfaces, challenges in reading and inputting text, interpreting visual cues, and engaging in real-time interactions. The central aim is to propose a chat application specifically tailored to enhance communication for this demographic. The proposed solution integrates cutting-edge speech-to-text and text-to-speech conversion, alongside emotion detection using Long Short-Term Memory (LSTM technology). This holistic approach seeks to create a more accessible and inclusive digital communication platform, thereby empowering users to better understand and express their emotions in online interactions.**



***Keywords*—Text-To-Speech, Speech-To-Text, Emotion Detection, Long Short-Term Memory.**


### I. Introduction

Blindness is a severe visual impairment that significantly reduces one's ability to see. It is characterized by total blindness or the inability to see forms, colors, or light. In order to go around, blind persons usually rely on support from other senses such as touch, hearing, smell, and taste and adaptive techniques. These senses are essential for people to comprehend emotions in their daily lives.

Worldwide, a minimum of 2.2 billion individuals experience either near or distance vision challenges. While vision loss can impact individuals across all age groups, the majority of those affected by vision impairment and blindness are typically aged 50 years and older. Types of low vision are Central vision impairment (difficulty seeing objects in the center of the field of vision), Peripheral vision impairment (difficulty seeing objects in the peripheral vision), Nocturnal vision impairment (difficulty seeing in low-light conditions), Blurred or hazy vision.

The paper primarily focuses on individuals with low vision, elderly individuals facing difficulty in reading or seeing clearly, and also offers insights applicable to dyslexic individuals struggling with fluent word reading and spelling. They often experience feelings of exclusion from the outside world. They desire to be treated equally and wish to communicate with those considered "normal". However, they encounter difficulties when using technology, especially when dealing with text-based content. For instance, navigating social media or chat applications can be challenging for them, ultimately resulting in their isolation from society. These challenges include difficulties in accessing and navigating digital platforms due to inadequate accessibility features, compatibility issues with screen reader software, and struggles with reading small text sizes or complex fonts. Moreover, they encounter challenges in interpreting emotional nuances conveyed through written language. These challenges encompass difficulties in discerning subtle emotional cues such as nuances or humor, which are commonly conveyed visually through font styles, emojis, or punctuation marks. The lack of alternative text descriptions for visual elements like emoticons or images further impedes their ability to grasp the emotional context of written content. Consequently, these obstacles contribute to potential misunderstandings and communication barriers for individuals with visual impairments. This highlights the importance of accessible and inclusive technology to bridge the gap and foster connections within the community. This paper presents "EmoSpeak," a chat application tailored for individuals with visual impairments. "EmoSpeak" is crafted to facilitate robust interaction, empowering visually impaired users to engage confidently with their environment. The system primarily facilitates text-to-speech and speech-to-text conversions, empowering users to communicate seamlessly.

Furthermore, the EmoSpeak application is integrated with an Emotion Detection from text Model for enhancing communication and emotional understanding among visually impaired individuals. The technology guarantees smooth interaction, which enables visually impaired people and sighted users to communicate effectively. A speech message sent by a user is transformed to text for the recipient. On the other hand, when a user receives a text message from the sender, it is converted into voice and identified emotions are also communicated. The model categorizes the sentiment into 6 emotions- joy , sadness , fear , anger , love and surprise. This method improves the overall communication experience of visually impaired users by guaranteeing that they keep a true link with the outside world, especially in text-based communication.

The following section offers a comprehensive review of existing literature on Text-to-Speech systems and textual emotion detection. Section 3 outlines the specifics of the proposed algorithm. Section 4 presents a comprehensive array of experiments aimed at evaluating the efficacy of the approach. Finally, Section 5 offers concluding remarks and

delves into prospective directions for future research initiatives.

## II. Related work

Over the years, there has been significant progress in assistive technologies, particularly in meeting the needs of individuals with visual impairments. These technologies encompass a wide range of tools and systems aimed at enhancing independence, mobility, and overall quality of life. K. Patil *et al*[1] have introduced a wearable device comprising five integrated components, including voice-over assistants for tasks such as understanding surroundings, searching for objects, recognizing faces with emotions, and reading. Notably, the integration of assistive technologies with mobile applications has led to substantial improvements in social communication and accessibility for visually impaired individuals.

A comprehensive survey by A. Karkar and S. Al-Maadeed [2] has explored various mobile-based systems tailored for individuals with visual impairments. This investigation spans a broad spectrum of applications, encompassing general assistive mobile applications like Awaaz [6], emphasizing the utilization of the Text-to-Speech (TTS) feature for vocalizing desired text. Additionally, the survey includes innovations such as BlindSense [3], an Android app utilizing semantic ontology to dynamically adjust user interfaces of common applications. Furthermore, notable mentions include iCure [4], a mobile application designed for detecting counterfeit Indian currency, and BSpeak [5], aiding economically disadvantaged blind individuals in earning income by transcribing audio files through speech.

Moreover, the survey covers advanced sensory-based applications like H-Slate[7], simulating the usage of physical braille devices, and MMM [8], utilizing Accelerometer as a sensor for motion marking menu. Additionally, mobile-based navigation systems such as SUGAR [9], an indoor navigation system, RSNAVI [10], facilitating navigation to specific locations with obstacle detection, and Poster [11], employing context awareness for indoor navigation, are highlighted. Furthermore, outdoor navigation solutions like iMove [12] and Off-Road [13], offering obstacle awareness, weather updates, social news, emergency calls, with tactile and auditory feedback, are discussed. Also included are applications like NavCog [14], an iOS app guiding visually impaired individuals to designated destinations, and Transport Assistant [15], aiding transportation and indoor/outdoor navigation, featuring image recognition, navigation, and vocal commands.

In highlighting the crucial role of assistive technologies for those with visual impairments, it's essential to underscore the importance of Text-to-Speech (TTS) and Speech-to-Text (STT) systems. Addressing this need, R. Ani, E. Maria, *et al* developed "Smart Specs" [16], a solution tailored for visually impaired users. This project integrates a Raspberry Pi with a built-in camera for capturing printed text images, utilizes Tesseract OCR for text recognition, and employs eSpeak for speech synthesis, ultimately delivering synthesized speech via headphones. The primary aim is to provide a compact and open-source method of converting printed text into audible speech. Conversely, S. Ghatak *et al* introduced "SocialWeb" [17], adopting a "keyboard-less" approach to facilitate user-friendly website exploration for individuals with limited computer proficiency. By leveraging screen readers and speech-to-text (STT) conversion technologies, the system integrates Web Speech API and acoustic fingerprinting for password validation, thereby enhancing accessibility and security, ensuring seamless engagement with social networking sites for visually impaired users.

The interpretation of emotions from text plays a crucial role in digital integration and fostering a sense of community inclusion for visually impaired individuals. Research conducted by S. Al-Saqqa *et al* [18] highlights the predominance of machine learning techniques in text-based emotion detection, driven by their implicit inference capabilities. Consequently, future investigations may explore deep learning methodologies, particularly when augmented by well-annotated datasets. Among these methodologies, LSTM emerges as a promising candidate for sentiment analysis, proficient in handling sequential textual data. Notably, M.-H. Su, *et al* [19] introduced an LSTM-based methodology for text emotion recognition, integrating semantic and emotional word vectors derived from word2vec and affective lexicons. By incorporating autoencoder bottleneck features for dimensionality reduction and LSTM for contextual emotion modeling, the proposed approach achieved a recognition accuracy of 70.66% on the NLPCC-MHMC-TE dataset, surpassing CNN-based methods by 5.33%. This study underscores the effectiveness of feature integration in bolstering performance metrics.

## III. Proposed Solution

Many solutions have been put forth in the field of helping the blind, ranging from Text-to-Speech (TTS) to Speech-to-Text (STT) systems. Still, an all-inclusive system that smoothly links the sighted and the individuals having low vision is necessary for efficient communication. This paper presents a solution called "EmoSpeak," which combines a Text-to-Speech (TTS) module with emotion identification from text. This system would combine cutting-edge technology with user-friendly interfaces, encouraging real connections and enabling the blind to easily connect with the outside world. It has great potential to improve the sociability and engagement of people with visual impairments. This function enables visually challenged persons to interact with digital content more successfully and to understand the subtle emotional aspects of text-based communication, leading to more meaningful conversations and facilitating greater inclusivity in social settings.

Fig 1. represents a comprehensive architecture of "EmoSpeak" comprising multiple levels and components tailored to meet the needs of stakeholders, including visually impaired and sighted users. The initial step involves user registration followed by login to initiate application use. Core features such as Text-to-Speech, Speech-to-Text, Emotion detection from text, real-time and group chat enhance the user experience. Additionally, the system includes user profile management, notifications, alerts, user authentication, and comprehensive profile customization. Access to user data, message histories, and database administration is seamless and reliable through the data layer. Lastly, the architecture prioritizes flawless communication and interoperability to empower visually impaired individuals in navigating the digital realm and fostering genuine connections with sighted users.

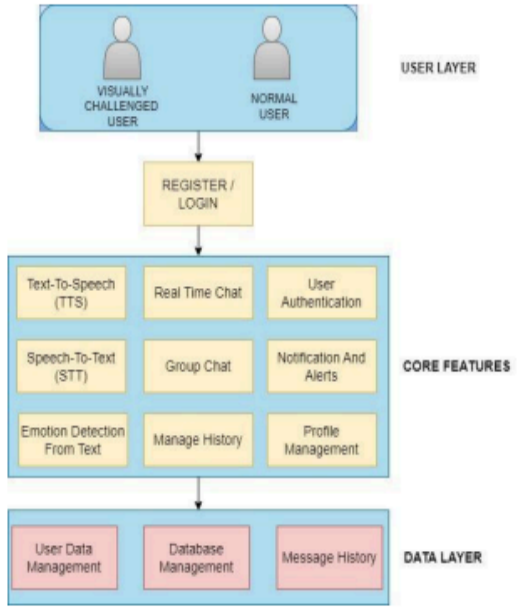



Fig 1. Overview of the system

### A. *Text-to-Speech*

The `flutter_tts` flutter package is utilized for integrating Text-to-Speech (TTS) functionality into Emospeak, the chat application. This integration empowers users to listen to text messages as spoken audio directly within the chat interface, enhancing accessibility and user experience. By leveraging the capabilities of the `flutter_tts` package, Emospeak seamlessly converts text messages into spoken words, providing users with an alternative means of communication and making the application more versatile and user-friendly.

### B. *Speech-to-Text*

Emospeak integrates the `speech_to_text` flutter package to enable effortless Speech-to-Text (STT) functionality. This integration enhances accessibility and user experience by allowing users to convert spoken words into text directly within the chat interface. By leveraging the `speech_to_text` package, Emospeak facilitates seamless input of messages through speech recognition, significantly enhancing the application's accessibility and user experience.

### C. *Emotion Detection*

While current Text-to-Speech (TTS) systems are proficient at converting text to speech, they frequently lack the ability to convey emotions effectively. This shortfall presents challenges for users in comprehending the emotional undertones of the content. Integrating emotional cues into TTS systems is pivotal for improving user experience. By incorporating emotion detection capabilities, this application can help users to understand the emotional content of the text. This enhancement results in more expressive synthesis, enabling users to better perceive and engage with the emotional context conveyed in the synthesized speech. Therefore, EmoSpeak will assist individuals in achieving more effective communication.

In the Emotion Detection Model, the dataset comprises English Twitter messages distributed across three CSV files, namely training, testing, and validation sets, containing 16,000, 2,000, and 2,000 instances, respectively. Each message is annotated with one of six emotion categories: sadness, joy, love, anger, fear, and surprise. These datasets serve as input for emotion detection from text employing a Long Short-Term Memory (LSTM) approach. Initially, emotions are categorized as specified in TABLE I.

TABLE I. DEPICTS THE DISTRIBUTION OF INPUT TEXT ACCORDING TO VARIOUS EMOTIONAL CATEGORIES.

| **Label** | **0** | **1** | **2** | **3** | **4** | **5** |
|---|---|---|---|---|---|---|
| **Emotion** | **Sad** | **Joy** | **Lov** | **Ang** | **Fea** | **Sur** |
| **No of sentences** | **4666** | **5362** | **1304** | **2159** | **1937** | **572** |

- **Sad: Sadness; Joy:Joy; Lov:Love; Ang:Anger; Fea:Fear; Sur: Surprise**

The inputs taken from the datasets are then preprocessed in the Embedding Layer. Preprocessing in the LSTM approach entails partitioning the dataset into two segments followed by their amalgamation for tokenization, where text is converted into tokens, with each token representing a single word, and stemming is applied to emphasize word semantics. After preprocessing, each word in the text is likely converted to a numerical index based on a vocabulary.

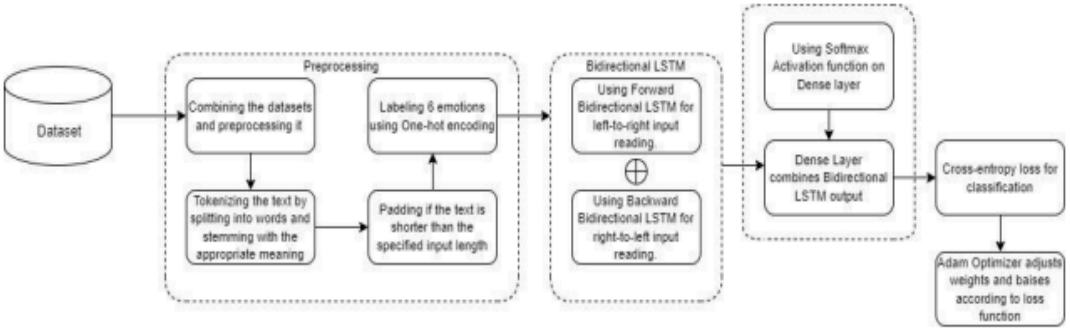



Fig 2. Long Short Term Memory Emotion Detection Model

The index represents the word's position in the vocabulary list. The embedding layer takes these word indices as input and looks up their corresponding embedding vectors in the embedding matrix. Subsequently, padding and One-hot-Encoding techniques are applied to these embedding vectors, containing semantic information about the words, and become the actual input to the bidirectional LSTM layers in the model. Embedding matrix: W ∈ R^(16000×100) where W is Word embedding matrix, 16000 is the vocabulary size and 100 is the embedding dimension.

The input is transferred to the bidirectional LSTM layer which is the core of the model responsible for learning long-range dependencies in the text sequence. It processes the input sequence in both forward and backward directions using two separate LSTMs. In a Forward LSTM, the sequence is processed from left to right, allowing it to capture the context preceding each word. Conversely, in a Backward LSTM, the sequence is processed from right to left, enabling it to capture the context succeeding each word. Following is the equation of Bidirectional LSTM:

$$h\_lstm = f\_LSTM(h\_embedding, h\_prev) \quad (1)$$

where h_lstm is the combined hidden state from both LSTMs, f_LSTM represents a function encapsulating the LSTM computations, and h_prev is the hidden state from the previous time step (or initial hidden state for the first step).

The output of both the LSTMs is concatenated and is transferred to the Dense Layer. Linear transformation of the final hidden states from both LSTMs:

$$z = W\_dense * h\_lstm + b\_dense \quad (2)$$

where W_dense is the weight matrix for the dense layer, combining information from the final hidden states of both forward and backward LSTMs (h_lstm) and b_dense is the bias vector for the dense layer.

Subsequently, the output of the Dense layer undergoes the application of the Softmax activation function. This function normalizes the values, ensuring they sum to 1 and represent probabilities for each class. These probabilities indicate the likelihood of the input text belonging to each of the possible classes. The equation of Normalized probability scores using softmax is

$$a = softmax(z) \quad (3)$$

In order to measure the disparity between the model's predictions and the actual values of the target variable, a loss function is employed. In this instance, the chosen loss function is categorical cross-entropy. By minimizing categorical cross-entropy, the model is trained to allocate higher probabilities to the correct classes for each input while reducing probabilities assigned to incorrect ones. This process enhances the model's capacity to precisely predict the class to which an unseen text sample belongs. The equation representing categorical cross-entropy loss, which compares predicted probabilities with true labels, is as follows:

$$L = categorical\_crossentropy(y\_true, a) \quad (4)$$

However, in machine learning, the goal is to minimize the lost function. Optimizers are algorithms that iteratively update the model's internal parameters (weights and biases) to gradually decrease the loss function. They adjust the parameters in the direction that leads to a steeper decrease in the loss.The Adam Optimizer is implemented with a learning rate set to 0.01 (lr=0.01). This process continues until the model reaches a minimum point in the loss function, hopefully representing a good fit to the data. Hence, the model outputs a vector of probabilities, with each component representing the probability of the input text belonging to a specific class in the multiclass classification problem. The predicted class is typically determined as the one with the highest probability, thus enabling accurate detection of the emotion.

Hence, to understand the emotion implied in the message, the 'EmoSpeak' chat application integrates a Text-to-Speech feature powered by a Machine Learning LSTM Model. Upon converting the text message into audio format, it identifies the underlying emotion, facilitating smoother user comprehension and communication.

## IV. Implementation and Results

The EmoSpeak application is implemented using Flutter, integrating Speech-to-Text (SST) and Text-to-Speech (TTS) functionalities along with the Emotion Detection Model. The LSTM Emotion Detection Model was trained on text datasets. After thorough testing and training, the following results were obtained.

TABLE II. PERFORMANCE METRICS FOR EMOTION DETECTION MODEL

| Class | Emotion | Parameters | | |
|---|---|---|---|---|
| | | Precision | Recall | F1-score |
| 0 | Sadness | 1.00 | 1.00 | 1.00 |
| 1 | Joy | 1.00 | 1.00 | 1.00 |
| 2 | Love | 0.99 | 1.00 | 0.99 |
| 3 | Anger | 1.00 | 1.00 | 1.00 |
| 4 | Fear | 0.99 | 0.99 | 0.99 |
| 5 | Surprise | 0.98 | 0.99 | 0.99 |

The high precision, recall, and F1-score values showcased in TABLE II across all classes indicate exceptional performance, underscoring the model's effectiveness in diverse categories. With an overall accuracy of 0.9963125, the model correctly classifies approximately 99.63% of the instances, further emphasizing its robustness and reliability.

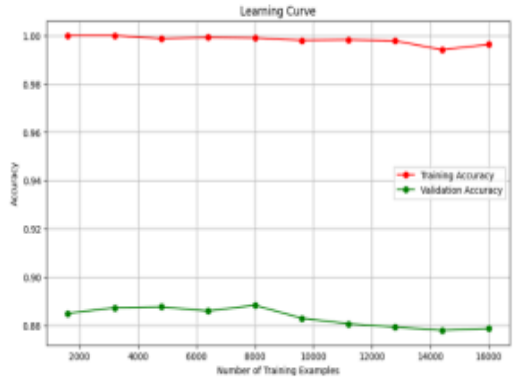



Fig 3 Learning Curve of Number of training examples vs Accuracy

Fig 3 displays the graph of a learning curve. It shows how well the model performs on a task as the number of training examples increases. On the graph, the x-axis corresponds to the number of training examples, while the y-axis denotes accuracy.There are two lines on the graph - the red line and the green line. The red line represents the accuracy of the model on the training data and the green line represents validation accuracy of the model on a separate set of data that the model has not been trained on. Here, the validation accuracy is around 0.88-0.90, that is, around 88%-90% accuracy which indicates how well the model is performing on the unseen data.

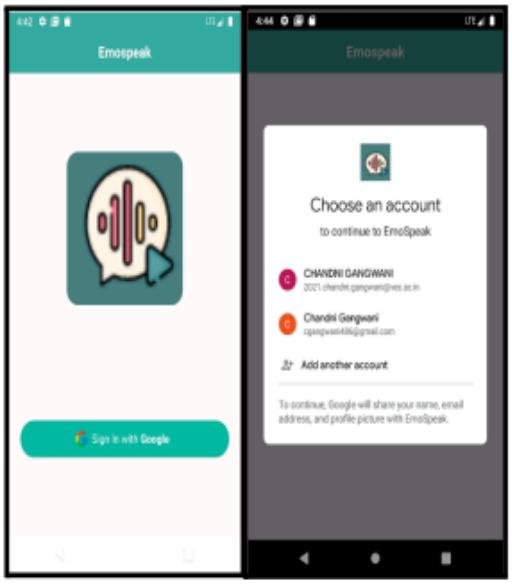



Fig 4 Sign-in page and Login Interface of EmoSpeak application

Fig 4 presents two screenshots showcasing the EmoSpeak mobile application, starting with the sign-in page and the login interface. It provides users with the choice to pick an account for entry. Beneath these options lies a button labeled "Add another account" for incorporating additional accounts. At the bottom of the screen, informative text assures users that Google will share their name, email address, and profile picture with EmoSpeak to proceed.

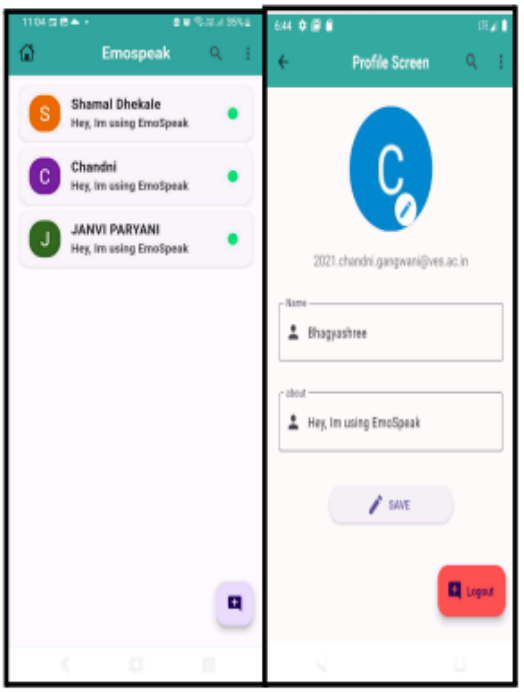



Fig 5 User selection chat interface and Profile screen for users

Fig 5 represents the interface that allows users to select a specific user with whom they wish to engage in a chat. Once a user is selected, the system opens an individual chat window dedicated to that particular user. The second screenshot depicts the user profile screen where users can manage their personal information, preferences, and settings. It provides a comprehensive overview of the user's account, including their profile picture, name, about, and a logout button.

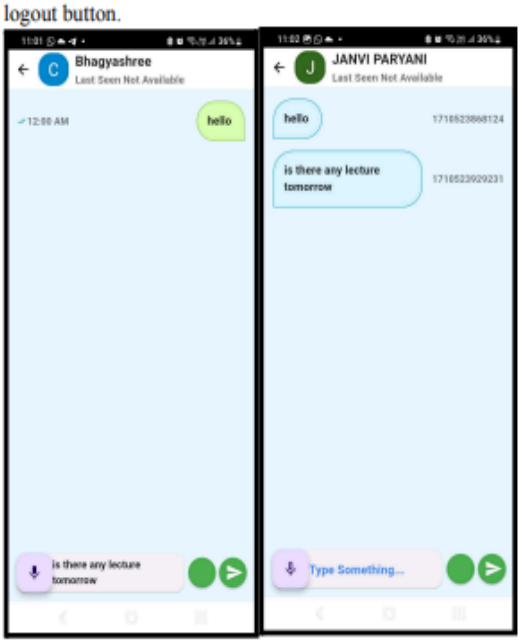



Fig 6 Speech-to-Text (STT)and Text-to-Speech(TTS) enabled chat functionality.

Fig 6 displays two screenshots of converting the message in Text-to-Speech and Speech-to-text. The first screenshot depicts the action of the user tapping on the microphone button to initiate speech input. The spoken words are then automatically transcribed into text, ready to be sent to the recipient user. In the second screenshot, upon receiving the message, the recipient clicks on the green button, prompting the generation of audio of the latest message received, simultaneously detecting the emotion expressed. The second screenshot

## V. Conclusion and Future Work

To sum up, the creation of EmoSpeak, a chat program designed for people with visual impairments, tackles the major obstacles these people encounter when using technology, especially when text-based communication is involved. EmoSpeak provides a comprehensive solution to improve communication accessibility and emotional expression for this demographic by integrating Speech-to-Text and Text-to-Speech functions with an LSTM architecture-based Emotion Detection Model. The LSTM model classifies sentiments into six unique emotions by accurately interpreting the emotional context of text by capturing sequential dependencies in the data. Visually impaired people may now interact more meaningfully and inclusively thanks to EmoSpeak, which bridges the gap between text-based messaging platforms and their specific needs. This creative method encourages emotional intelligence and connectedness in everyday life in addition to giving people the ability to speak on their own.This feature framework can be integrated into existing chat applications and can make them more powerful.

EmoSpeak's future evolution entails two main objectives: broadening language support to serve diverse global users and refining its user interface for enhanced accessibility. Through language expansion, EmoSpeak aims to foster inclusivity, while ongoing user interface optimization, driven by user feedback and usability studies, ensures continued user-friendliness for visually impaired individuals. These endeavors reinforce EmoSpeak's position as a pivotal tool for accessible and emotionally expressive communication.

## References


[1] K. Patil, A. Kharat, P. Chaudhary, S. Bidgar and R. Gavhane, "Guidance System for Visually Impaired People," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 2021, pp. 988-993, doi: 10.1109/ICAIS50930.2021.9395973.

[2] A. Karkar and S. Al-Maadeed, "Mobile Assistive Technologies for Visual Impaired Users: A Survey," 2018 International Conference on Computer and Applications (ICCA), Beirut, Lebanon, 2018, pp. 427-433, doi: 10.1109/COMAPP.2018.8460406.

[3] Khan, Akif, Shah Khusro, and Iftikhar Alam. "BlindSense: An Accessibility-inclusive Universal User Interface for Blind People." Engineering, Technology & Applied Science Research 8.2 (2018).

[4] V. Raval and A. Shah, "icuean iot application for indian currency recognition in vernacular languages for visually challenged people," in 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence. IEEE, 2017, pp. 577–581..

[5] A. Vashistha, P. Sethi, and R. Anderson, "Bspeak: An accessible voice-based crowdsourcing marketplace for low-income blind people," in Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM, 2018, p. 57.

[6] S. Malgaonkar, P. Shah, D. Panchal, and S. Pradhan, "Awaaz: A bridge between android phones and the visually impaired," International Journal of Computer Applications, vol. 134, no. 1, pp. 42–47, 2016.

[7] G. Lee, L. C. Quero, J. Yang, H. Jung, J. Son, and J. Cho, "Slate master: a tangible braille slate tutor for mobile devices," in Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services. ACM, 2017, p. 108.

[8] N. K. Dim, K. Kim, and X. Ren, "Designing motion marking menus for people with visual impairments," International Journal of HumanComputer Studies, vol. 109, pp. 79–88, 2018.

[9] A. S. Martinez-Sala, F. Losilla, J. C. Sanchez-Aarnoutse, and J. Garc ´ ´ıaHaro, "Design, implementation and evaluation of an indoor navigation system for visually impaired people," Sensors, vol. 15, no. 12, pp. 32 168–32 187, 2015.

[10] R. Ivanov, "Rsnavi: an rfid-based context-aware indoor navigation system for the blind," in Proceedings of the 13th international conference on computer systems and technologies. ACM, 2012, pp. 313–320.

[11] S. K. Long, N. D. Karpinsky, H. Doner, and J. D. Still, "Using a mobile ¨ application to help visually impaired individuals explore the outdoors," in Advances in design for inclusion. Springer, 2016, pp. 213–223.

[12] H. Kacorri, S. Mascetti, A. Gerino, D. Ahmetovic, H. Takagi, and C. Asakawa, "Supporting orientation of people with visual impairment: Analysis of large scale usage data," in Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility. ACM, 2016, pp. 151–159.

[13] S. K. Long, N. D. Karpinsky, H. Doner, and J. D. Still, "Using a mobile ¨ application to help visually impaired individuals explore the outdoors," in Advances in design for inclusion. Springer, 2016, pp. 213–223.

[14] D. Ahmetovic, C. Gleason, C. Ruan, K. Kitani, H. Takagi, and C. Asakawa, "Navcog: a navigational cognitive assistant for the blind," in Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services. ACM, 2016, pp. 90–99.

[15] G. G. Shenoy, M. A. Wagle, and K. Connelly, "Leveling the playing field for visually impaired using transport assistant," arXiv preprint arXiv:1703.02103, 2017.

[16] R. Ani, E. Maria, J. J. Joyce, V. Sakkaravarthy and M. A. Raja, "Smart Specs: Voice assisted text reading system for visually impaired persons using TTS method," 2017 International Conference on Innovations in Green Energy and Healthcare Technologies (IGEHT), Coimbatore, India, 2017, pp. 1-6, doi: 10.1109/IGEHT.2017.8094103.

[17] S. Ghatak, A. Lodh, E. Saha, A. Goyal, A. Das and S. Dutta, "Development of a keyboardless social networking website for visually impaired: SocialWeb," 2014 IEEE Global Humanitarian Technology Conference - South Asia Satellite (GHTC-SAS), Trivandrum, India, 2014, pp. 232-236, doi: 10.1109/GHTC-SAS.2014.6967589.

[18] S. Al-Saqqa, H. Abdel-Nabi and A. Awajan, "A Survey of Textual Emotion Detection," 2018 8th International Conference on Computer Science and Information Technology (CSIT), Amman, Jordan, 2018, pp. 136-142, doi: 10.1109/CSIT.2018.8486405.

[19] M. -H. Su, C. -H. Wu, K. -Y. Huang and Q. -B. Hong, "LSTM-based Text Emotion Recognition Using Semantic and Emotional Word Vectors," 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), Beijing, China, 2018, pp. 1-6, doi: 10.1109/ACIIAsia.2018.8470378.

## b. Certificate of publication / Project Competition

We have submitted the research paper on International Conference On Advancements In Power, Communication And Intelligent Systems - APCI 2024 and we are eagerly anticipating the results of our submission.

## c. Plagiarism report

# EmoSpeak: An Emotionally Intelligent TTS System for Visually Impaired

*by* Shamal Dhekale

**Submission date:** 19-Mar-2024 02:08PM (UTC+0530)
**Submission ID:** 2324619188
**File name:** copy_of_research_paper_42.pdf (1.41M)
**Word count:** 3492
**Character count:** 21145

# EmoSpeak: An Emotionally Intelligent TTS System for Visually Impaired

Yugchhaya Galphat
*Computer Department*
*Vivekanand Education Society's Institute of Technology*
Mumbai, India.
yugchhaya.dhote@ves.ac.in

Bhagyashree Vaswani
*Computer Department*
*Vivekanand Education Society's Institute of Technology*
Mumbai, India.
2021.bhagyashree.vaswani@ves.ac.in

Chandni Gangwani
*Computer Department*
*Vivekanand Education Society's Institute of Technology*
Mumbai, India.
2021.chandni.gangwani@ves.ac.in

Shamal Dhekale
*Computer Department*
*Vivekanand Education Society's Institute of Technology*
Mumbai, India.
2021.shamal.dhekale@ves.ac.in

***Abstract*—The paper delves into the obstacles confronting individuals with visual impairments, especially those experiencing blurred vision due to medical conditions or aging, when navigating technology, notably chat services. People with weak eyesight face issues such as inaccessible user interfaces, challenges in reading and inputting text, interpreting visual cues, and engaging in real-time interactions. The central aim is to propose a chat application specifically tailored to enhance communication for this demographic. The proposed solution integrates cutting-edge speech-to-text and text-to-speech conversion, alongside emotion detection using Long Short-Term Memory (LSTM technology). This holistic approach seeks to create a more accessible and inclusive digital communication platform, thereby empowering users to better understand and express their emotions in online interactions.**

***Keywords*—Text-To-Speech, Speech-To-Text, Emotion Detection, Long Short-Term Memory.**

## I. Introduction

Blindness is a severe visual impairment that significantly reduces one's ability to see. It is characterized by total blindness or the inability to see forms, colors, or light. In order to go around, blind persons usually rely on support from other senses such as touch, hearing, smell, and taste and adaptive techniques. These senses are essential for people to comprehend emotions in their daily lives.

Worldwide, a minimum of 2.2 billion individuals experience either near or distance vision challenges. While vision loss can impact individuals across all age groups, the majority of those affected by vision impairment and blindness are typically aged 50 years and older. Types of low vision are Central vision impairment (difficulty seeing objects in the center of the field of vision), Peripheral vision impairment (difficulty seeing objects in the peripheral vision), Nocturnal vision impairment (difficulty seeing in low-light conditions), Blurred or hazy vision.

The paper primarily focuses on individuals with low vision, elderly individuals facing difficulty in reading or seeing clearly, and also offers insights applicable to dyslexic individuals struggling with fluent word reading and spelling. They often experience feelings of exclusion from the outside world. They desire to be treated equally and wish to communicate with those considered "normal". However, they encounter difficulties when using technology, especially when dealing with text-based content. For instance, navigating social media or chat applications can be challenging for them, ultimately resulting in their isolation from society. These challenges include difficulties in accessing and navigating digital platforms due to inadequate accessibility features, compatibility issues with screen reader software, and struggles with reading small text sizes or complex fonts. Moreover, they encounter challenges in interpreting emotional nuances conveyed through written language. These challenges encompass difficulties in discerning subtle emotional cues such as nuances or humor, which are commonly conveyed visually through font styles, emojis, or punctuation marks. The lack of alternative text descriptions for visual elements like emoticons or images further impedes their ability to grasp the emotional context of written content. Consequently, these obstacles contribute to potential misunderstandings and communication barriers for individuals with visual impairments. This highlights the importance of accessible and inclusive technology to bridge the gap and foster connections within the community. This paper presents "EmoSpeak," a chat application tailored for individuals with visual impairments. "EmoSpeak" is crafted to facilitate robust interaction, empowering visually impaired users to engage confidently with their environment. The system primarily facilitates text-to-speech and speech-to-text conversions, empowering users to communicate seamlessly.

Furthermore, the EmoSpeak application is integrated with an Emotion Detection from text Model for enhancing communication and emotional understanding among visually impaired individuals. The technology guarantees smooth interaction, which enables visually impaired people and sighted users to communicate effectively. A speech message sent by a user is transformed to text for the recipient. On the other hand, when a user receives a text message from the sender, it is converted into voice and identified emotions are also communicated. The model categorizes the sentiment into 6 emotions- joy , sadness , fear , anger , love and surprise. This method improves the overall communication experience of visually impaired users by guaranteeing that they keep a true link with the outside world, especially in text-based communication.

The following section offers a comprehensive review of existing literature on Text-to-Speech systems and textual emotion detection. Section 3 outlines the specifics of the proposed algorithm. Section 4 presents a comprehensive array of experiments aimed at evaluating the efficacy of the approach. Finally, Section 5 offers concluding remarks and

delves into prospective directions for future research initiatives.

## II. Related work

Over the years, there has been significant progress in assistive technologies, particularly in meeting the needs of individuals with visual impairments. These technologies encompass a wide range of tools and systems aimed at enhancing independence, mobility, and overall quality of life. K. Patil *et al*[1] have introduced a wearable device comprising five integrated components, including voice-over assistants for tasks such as understanding surroundings, searching for objects, recognizing faces with emotions, and reading. Notably, the integration of assistive technologies with mobile applications has led to substantial improvements in social communication and accessibility for visually impaired individuals.

A comprehensive survey by A. Karkar and S. Al-Maadeed [2] has explored various mobile-based systems tailored for individuals with visual impairments. This investigation spans a broad spectrum of applications, encompassing general assistive mobile applications like Awaaz [6], emphasizing the utilization of the Text-to-Speech (TTS) feature for vocalizing desired text. Additionally, the survey includes innovations such as BlindSense [3], an Android app utilizing semantic ontology to dynamically adjust user interfaces of common applications. Furthermore, notable mentions include iCure [4], a mobile application designed for detecting counterfeit Indian currency, and BSpeak [5], aiding economically disadvantaged blind individuals in earning income by transcribing audio files through speech.

Moreover, the survey covers advanced sensory-based applications like H-Slate[7], simulating the usage of physical braille devices, and MMM [8], utilizing Accelerometer as a sensor for motion marking menu. Additionally, mobile-based navigation systems such as SUGAR [9], an indoor navigation system, RSNAVI [10], facilitating navigation to specific locations with obstacle detection, and Poster [11], employing context awareness for indoor navigation, are highlighted. Furthermore, outdoor navigation solutions like iMove [12] and Off-Road [13], offering obstacle awareness, weather updates, social news, emergency calls, with tactile and auditory feedback, are discussed. Also included are applications like NavCog [14], an iOS app guiding visually impaired individuals to designated destinations, and Transport Assistant [15], aiding transportation and indoor/outdoor navigation, featuring image recognition, navigation, and vocal commands.

In highlighting the crucial role of assistive technologies for those with visual impairments, it's essential to underscore the importance of Text-to-Speech (TTS) and Speech-to-Text (STT) systems. Addressing this need, R. Ani, E. Maria, *et al* developed "Smart Specs" [16], a solution tailored for visually impaired users. This project integrates a Raspberry Pi with a built-in camera for capturing printed text images, utilizes Tesseract OCR for text recognition, and employs eSpeak for speech synthesis, ultimately delivering synthesized speech via headphones. The primary aim is to provide a compact and open-source method of converting printed text into audible speech. Conversely, S. Ghatak *et al* introduced "SocialWeb" [17], adopting a "keyboard-less" approach to facilitate user-friendly website exploration for individuals with limited computer proficiency. By leveraging screen readers and speech-to-text (STT) conversion technologies, the system integrates Web Speech API and acoustic fingerprinting for password validation, thereby enhancing accessibility and security, ensuring seamless engagement with social networking sites for visually impaired users.

The interpretation of emotions from text plays a crucial role in digital integration and fostering a sense of community inclusion for visually impaired individuals. Research conducted by S. Al-Saqqa *et al* [18] highlights the predominance of machine learning techniques in text-based emotion detection, driven by their implicit inference capabilities. Consequently, future investigations may explore deep learning methodologies, particularly when augmented by well-annotated datasets. Among these methodologies, LSTM emerges as a promising candidate for sentiment analysis, proficient in handling sequential textual data. Notably, M.-H. Su, *et al* [19] introduced an LSTM-based methodology for text emotion recognition, integrating semantic and emotional word vectors derived from word2vec and affective lexicons. By incorporating autoencoder bottleneck features for dimensionality reduction and LSTM for contextual emotion modeling, the proposed approach achieved a recognition accuracy of 70.66% on the NLPCC-MHMC-TE dataset, surpassing CNN-based methods by 5.33%. This study underscores the effectiveness of feature integration in bolstering performance metrics.

## III. Proposed Solution

Many solutions have been put forth in the field of helping the blind, ranging from Text-to-Speech (TTS) to Speech-to-Text (STT) systems. Still, an all-inclusive system that smoothly links the sighted and the individuals having low vision is necessary for efficient communication. This paper presents a solution called "EmoSpeak," which combines a Text-to-Speech (TTS) module with emotion identification from text. This system would combine cutting-edge technology with user-friendly interfaces, encouraging real connections and enabling the blind to easily connect with the outside world. It has great potential to improve the sociability and engagement of people with visual impairments. This function enables visually challenged persons to interact with digital content more successfully and to understand the subtle emotional aspects of text-based communication, leading to more meaningful conversations and facilitating greater inclusivity in social settings.

Fig 1. represents a comprehensive architecture of “EmoSpeak” comprising multiple levels and components tailored to meet the needs of stakeholders, including visually impaired and sighted users. The initial step involves user registration followed by login to initiate application use. Core features such as Text-to-Speech, Speech-to-Text, Emotion detection from text, real-time and group chat enhance the user experience. Additionally, the system includes user profile management, notifications, alerts, user authentication, and comprehensive profile customization. Access to user data, message histories, and database administration is seamless and reliable through the data layer. Lastly, the architecture prioritizes flawless communication and interoperability to empower visually impaired individuals in navigating the digital realm and fostering genuine connections with sighted users.

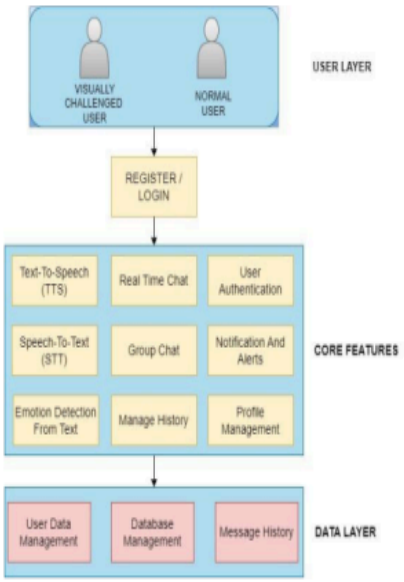



Fig 1. Overview of the system

### A. *Text-to-Speech*

The `flutter_tts` flutter package is utilized for integrating Text-to-Speech (TTS) functionality into Emospeak, the chat application. This integration empowers users to listen to text messages as spoken audio directly within the chat interface, enhancing accessibility and user experience. By leveraging the capabilities of the `flutter_tts` package, Emospeak seamlessly converts text messages into spoken words, providing users with an alternative means of communication and making the application more versatile and user-friendly.

### B. *Speech-to-Text*

Emospeak integrates the `speech_to_text` flutter package to enable effortless Speech-to-Text (STT) functionality. This integration enhances accessibility and user experience by allowing users to convert spoken words into text directly within the chat interface. By leveraging the `speech_to_text` package, Emospeak facilitates seamless input of messages through speech recognition, significantly enhancing the application's accessibility and user experience.

### C. *Emotion Detection*

While current Text-to-Speech (TTS) systems are proficient at converting text to speech, they frequently lack the ability to convey emotions effectively. This shortfall presents challenges for users in comprehending the emotional undertones of the content. Integrating emotional cues into TTS systems is pivotal for improving user experience. By incorporating emotion detection capabilities, this application can help users to understand the emotional content of the text. This enhancement results in more expressive synthesis, enabling users to better perceive and engage with the emotional context conveyed in the synthesized speech. Therefore, EmoSpeak will assist individuals in achieving more effective communication.

In the Emotion Detection Model, the dataset comprises English Twitter messages distributed across three CSV files, namely training, testing, and validation sets, containing 16,000, 2,000, and 2,000 instances, respectively. Each message is annotated with one of six emotion categories: sadness, joy, love, anger, fear, and surprise. These datasets serve as input for emotion detection from text employing a Long Short-Term Memory (LSTM) approach. Initially, emotions are categorized as specified in TABLE I.

TABLE I. DEPICTS THE DISTRIBUTION OF INPUT TEXT ACCORDING TO VARIOUS EMOTIONAL CATEGORIES.

| Label | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Emotion | Sad | Joy | Lov | Ang | Fea | Sur |
| No of sentences | 4666 | 5362 | 1304 | 2159 | 1937 | 572 |

- Sad: Sadness; Joy:Joy; Lov:Love; Ang:Anger; Fea:Fear; Sur: Surprise

The inputs taken from the datasets are then preprocessed in the Embedding Layer. Preprocessing in the LSTM approach entails partitioning the dataset into two segments followed by their amalgamation for tokenization, where text is converted into tokens, with each token representing a single word, and stemming is applied to emphasize word semantics. After preprocessing, each word in the text is likely converted to a numerical index based on a vocabulary.

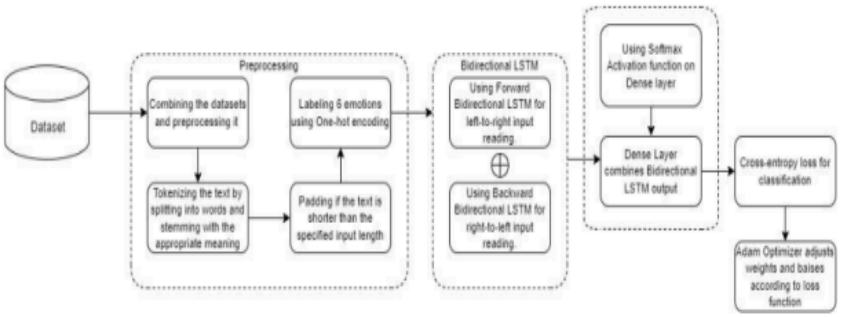



Fig 2. Long Short Term Memory Emotion Detection Model

The index represents the word's position in the vocabulary list. The embedding layer takes these word indices as input and looks up their corresponding embedding vectors in the embedding matrix. Subsequently, padding and One-hot-Encoding techniques are applied to these embedding vectors, containing semantic information about the words, and become the actual input to the bidirectional LSTM layers in the model. Embedding matrix: W ∈ R^(16000×100) where W is Word embedding matrix, 16000 is the vocabulary size and 100 is the embedding dimension.

The input is transferred to the bidirectional LSTM layer which is the core of the model responsible for learning long-range dependencies in the text sequence. It processes the input sequence in both forward and backward directions using two separate LSTMs. In a Forward LSTM, the sequence is processed from left to right, allowing it to capture the context preceding each word. Conversely, in a Backward LSTM, the sequence is processed from right to left, enabling it to capture the context succeeding each word. Following is the equation of Bidirectional LSTM:

$$h\_lstm = f\_LSTM(h\_embedding, h\_prev)$$

(1)

where h_lstm is the combined hidden state from both LSTMs, f_LSTM represents a function encapsulating the LSTM computations, and h_prev is the hidden state from the previous time step (or initial hidden state for the first step).

The output of both the LSTMs is concatenated and is transferred to the Dense Layer. Linear transformation of the final hidden states from both LSTMs:

$$z = W\_dense * h\_lstm + b\_dense$$

(2)

where W_dense is the weight matrix for the dense layer, combining information from the final hidden states of both forward and backward LSTMs (h_lstm) and b_dense is the bias vector for the dense layer.

Subsequently, the output of the Dense layer undergoes the application of the Softmax activation function. This function normalizes the values, ensuring they sum to 1 and represent probabilities for each class. These probabilities indicate the likelihood of the input text belonging to each of the possible classes. The equation of Normalized probability scores using softmax is

$$a = softmax(z)$$

(3)

In order to measure the disparity between the model's predictions and the actual values of the target variable, a loss function is employed. In this instance, the chosen loss function is categorical cross-entropy. By minimizing categorical cross-entropy, the model is trained to allocate higher probabilities to the correct classes for each input while reducing probabilities assigned to incorrect ones. This process enhances the model's capacity to precisely predict the class to which an unseen text sample belongs. The equation representing categorical cross-entropy loss, which compares predicted probabilities with true labels, is as follows:

$$L = categorical\_crossentropy(y\_true, a)$$

(4)

However, in machine learning, the goal is to minimize the lost function. Optimizers are algorithms that iteratively update the model's internal parameters (weights and biases) to gradually decrease the loss function. They adjust the parameters in the direction that leads to a steeper decrease in the loss.The Adam Optimizer is implemented with a learning rate set to 0.01 (lr=0.01). This process continues until the model reaches a minimum point in the loss function, hopefully representing a good fit to the data. Hence, the model outputs a vector of probabilities, with each component representing the probability of the input text belonging to a specific class in the multiclass classification problem. The predicted class is typically determined as the one with the highest probability, thus enabling accurate detection of the emotion.

Hence, to understand the emotion implied in the message, the 'EmoSpeak' chat application integrates a Text-to-Speech feature powered by a Machine Learning LSTM Model. Upon converting the text message into audio format, it identifies the underlying emotion, facilitating smoother user comprehension and communication.

## IV. Implementation and Results

The EmoSpeak application is implemented using Flutter, integrating Speech-to-Text (SST) and Text-to-Speech (TTS) functionalities along with the Emotion Detection Model. The LSTM Emotion Detection Model was trained on text datasets. After thorough testing and training, the following results were obtained.

TABLE II. PERFORMANCE METRICS FOR EMOTION DETECTION MODEL

| Class | Emotion | Parameters | | |
|---|---|---|---|---|
| | | Precision | Recall | F1-score |
| 0 | Sadness | 1.00 | 1.00 | 1.00 |
| 1 | Joy | 1.00 | 1.00 | 1.00 |
| 2 | Love | 0.99 | 1.00 | 0.99 |
| 3 | Anger | 1.00 | 1.00 | 1.00 |
| 4 | Fear | 0.99 | 0.99 | 0.99 |
| 5 | Surprise | 0.98 | 0.99 | 0.99 |

The high precision, recall, and F1-score values showcased in TABLE II across all classes indicate exceptional performance, underscoring the model's effectiveness in diverse categories. With an overall accuracy of 0.9963125, the model correctly classifies approximately 99.63% of the instances, further emphasizing its robustness and reliability.

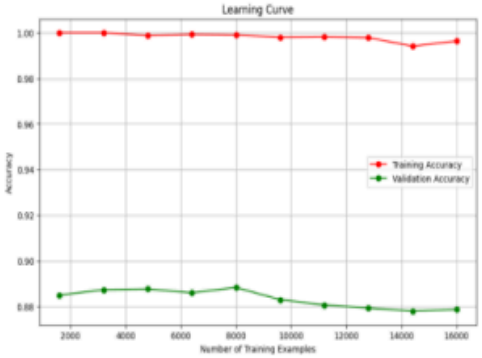



Fig 3 Learning Curve of Number of training examples vs Accuracy

Fig 3 displays the graph of a learning curve. It shows how well the model performs on a task as the number of training examples increases. On the graph, the x-axis corresponds to the number of training examples, while the y-axis denotes accuracy.There are two lines on the graph - the red line and the green line. The red line represents the accuracy of the model on the training data and the green line represents validation accuracy of the model on a separate set of data that the model has not been trained on. Here, the validation accuracy is around 0.88-0.90, that is, around 88%-90% accuracy which indicates how well the model is performing on the unseen data.

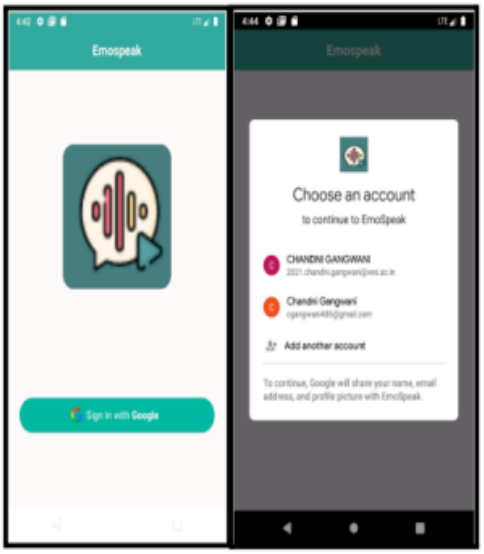



Fig 4 Sign-in page and Login Interface of EmoSpeak application

Fig 4 presents two screenshots showcasing the EmoSpeak mobile application, starting with the sign-in page and the login interface. It provides users with the choice to pick an account for entry. Beneath these options lies a button labeled "Add another account" for incorporating additional accounts. At the bottom of the screen, informative text assures users that Google will share their name, email address, and profile picture with EmoSpeak to proceed.

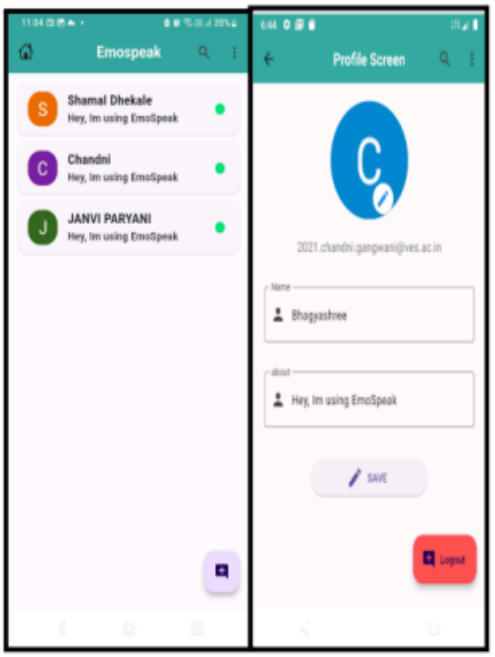



Fig 5 User selection chat interface and Profile screen for users

Fig 5 represents the interface that allows users to select a specific user with whom they wish to engage in a chat. Once a user is selected, the system opens an individual chat window dedicated to that particular user. The second screenshot depicts the user profile screen where users can manage their personal information, preferences, and settings. It provides a comprehensive overview of the user's account, including their profile picture, name, about, and a logout button.

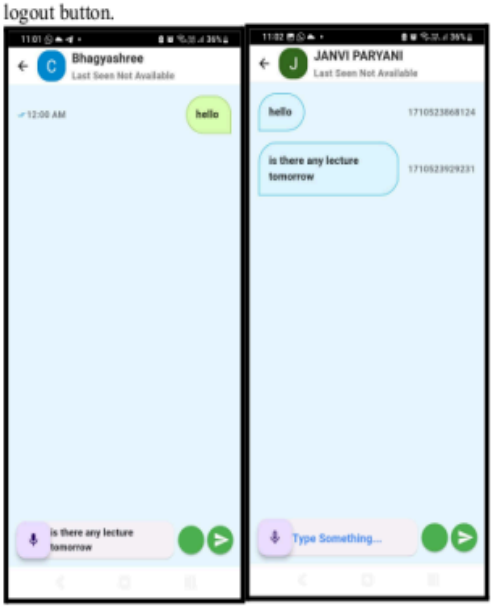



Fig 6 Speech-to-Text (STT)and Text-to-Speech(TTS) enabled chat functionality

Fig 6 displays two screenshots of converting the message in Text-to-Speech and Speech-to-text. The first screenshot depicts the action of the user tapping on the microphone button to initiate speech input. The spoken words are then automatically transcribed into text, ready to be sent to the recipient user. In the second screenshot, upon receiving the message, the recipient clicks on the green button, prompting the generation of audio of the latest message received, simultaneously detecting the emotion expressed. The second screenshot

## V. Conclusion and Future Work

To sum up, the creation of EmoSpeak, a chat program designed for people with visual impairments, tackles the major obstacles these people encounter when using technology, especially when text-based communication is involved. EmoSpeak provides a comprehensive solution to improve communication accessibility and emotional expression for this demographic by integrating Speech-to-Text and Text-to-Speech functions with an LSTM architecture-based Emotion Detection Model. The LSTM model classifies sentiments into six unique emotions by accurately interpreting the emotional context of text by capturing sequential dependencies in the data. Visually impaired people may now interact more meaningfully and inclusively thanks to EmoSpeak, which bridges the gap between text-based messaging platforms and their specific needs. This creative method encourages emotional intelligence and connectedness in everyday life in addition to giving people the ability to speak on their own.This feature framework can be integrated into existing chat applications and can make them more powerful.

EmoSpeak's future evolution entails two main objectives: broadening language support to serve diverse global users and refining its user interface for enhanced accessibility. Through language expansion, EmoSpeak aims to foster inclusivity, while ongoing user interface optimization, driven by user feedback and usability studies, ensures continued user-friendliness for visually impaired individuals. These endeavors reinforce EmoSpeak's position as a pivotal tool for accessible and emotionally expressive communication.

## References

# EmoSpeak: An Emotionally Intelligent TTS System for Visually Impaired

ORIGINALITY REPORT

| 3% | 2% | 3% | 1% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | "Sustainable Development through Machine Learning, AI and IoT", Springer Science and Business Media LLC, 2023<br>Publication | 1% |
| 2 | Nupur Giri, Tamanna Saini, Kalpesh Bhole, Anuraj Bhosale, Tanishqa Shetty, Alka Subramanyam, Swati Shelke. "Detection of Dyscalculia Using Machine Learning", 2020 5th International Conference on Communication and Electronics Systems (ICCES), 2020<br>Publication | 1% |
| 3 | Yifan Zhao, Zhanhui Hu, Rongjun Liu. "TBGD: Deep Learning Methods on Network Intrusion Detection Using CICIDS2017 Dataset", Journal of Physics: Conference Series, 2023<br>Publication | 1% |
| 4 | sanskrit.jnu.ac.in<br>Internet Source | 1% |

| | | | |
|---|---|---|---|
| Exclude quotes | Off | Exclude matches | < 1% |
| Exclude bibliography | Off | | |

EmoSpeak: An Emotionally Intelligent TTS System for Visually Impaired

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

PAGE 1

PAGE 2

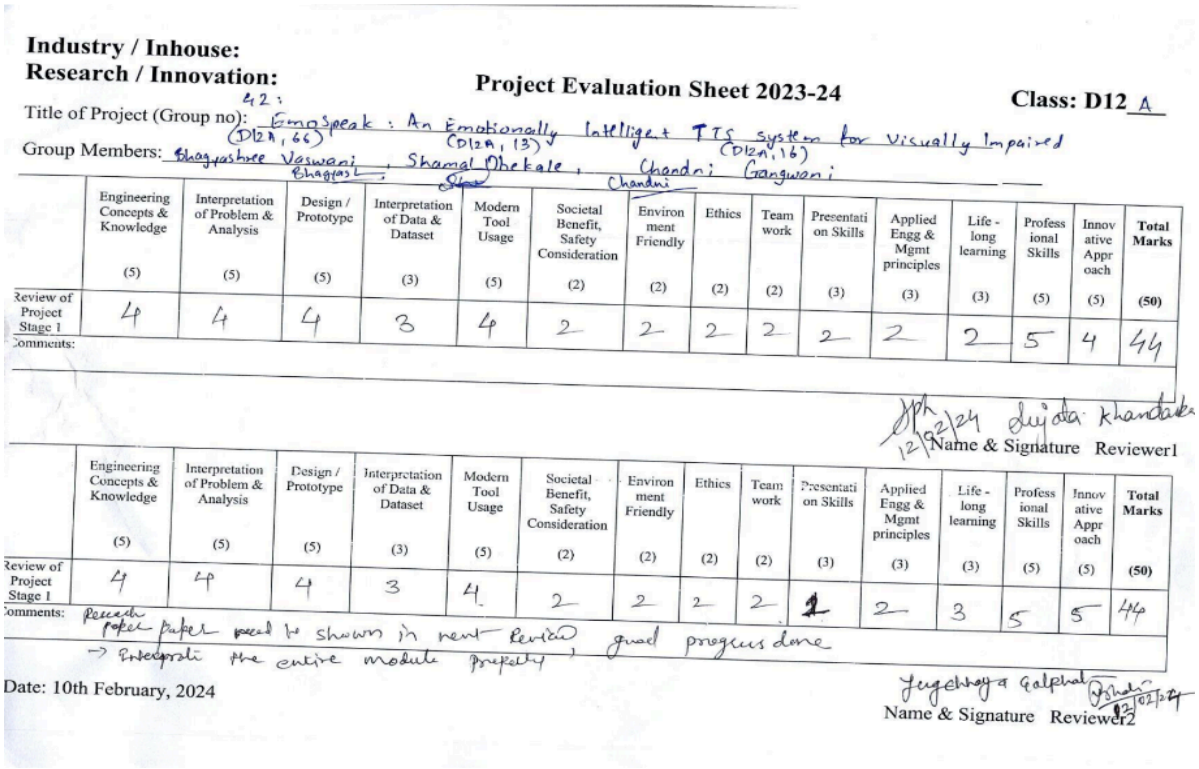
PAGE 3

PAGE 4

PAGE 5

PAGE 6

## d. Project review sheet



**Industry / Inhouse:**
**Research / Innovation:**

**Project Evaluation Sheet 2023-24**

**Class: D12_A**

Title of Project (Group no): 42: EmoSpeak : An Emotionally Intelligent TTS system for Visually Impaired

Group Members: Bhagyashree Vaswani (D12A, 66), Shamal Dhekale (D12A, 13), Chandni Gangwani (D12A, 16)

| | Engineering Concepts & Knowledge (5) | Interpretation of Problem & Analysis (5) | Design / Prototype (5) | Interpretation of Data & Dataset (3) | Modern Tool Usage (5) | Societal Benefit, Safety Consideration (2) | Environment Friendly (2) | Ethics (2) | Team work (2) | Presentation Skills (3) | Applied Engg & Mgmt principles (3) | Life - long learning (3) | Professional Skills (5) | Innovative Approach (5) | Total Marks (50) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Review of Project Stage 1 | 4 | 4 | 4 | 3 | 4 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 5 | 4 | 44 |

Comments:

12/2/24 Sujata Khandekar
Name & Signature Reviewer1

| | Engineering Concepts & Knowledge (5) | Interpretation of Problem & Analysis (5) | Design / Prototype (5) | Interpretation of Data & Dataset (3) | Modern Tool Usage (5) | Societal Benefit, Safety Consideration (2) | Environment Friendly (2) | Ethics (2) | Team work (2) | Presentation Skills (3) | Applied Engg & Mgmt principles (3) | Life - long learning (3) | Professional Skills (5) | Innovative Approach (5) | Total Marks (50) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Review of Project Stage 1 | 4 | 4 | 4 | 3 | 4 | 2 | 2 | 2 | 2 | 2 | 2 | 3 | 5 | 5 | 44 |

Comments: Research paper paper need to shown in next Review good progress done
→ Incorporate the entire module properly

Date: 10th February, 2024

Jugeshwara Gaikwad 12/02/24
Name & Signature Reviewer2