

Multi-Model Approach to Recommend Personalized Music Playlist

Project ID: TMP – 2023 – 24 - 065

Project Proposal Report

Gunasekara C.M – IT20665852

Supervised by – Mr. Thusithanjana Thilakarathna

B.Sc. (Hons) Degree in Information Technology
Department of Information Technology

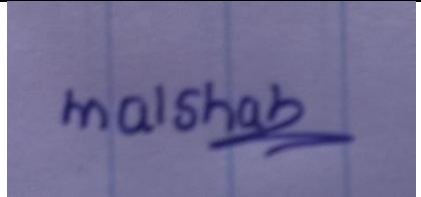
Sri Lanka Institute of Information Technology
Sri Lanka



August 2023

1 Declaration

We declare that this is our own work, and this proposal does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of our knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Name	Student ID	Signature
Gunasekara C.M	IT20665852	

The above candidates are carrying out research for the undergraduate Dissertation under my supervision.

Name of supervisor: Mr. Thusithanjana Thilakarathna

Name of co-supervisor: Dr. Darshana Kasthurirathna



.....

.....08/25/2023.....

Signature of the supervisor:

Date



.....

.....08/25/2023.....

Signature of the Co - supervisor:

Date

2 Abstract

This research focuses on an innovative system that can identify users' emotional states by listening to their vocal frequencies. In addition to comprehending the emotional context of users' speech interactions, this dynamic system also offers recommendations and customized playlist ideas that are sympathetic to their feelings. The Objective is to create a unique and intelligent system that combines voice-based emotion detection with customized playlist generation. The goal of this system is to improve user experiences by establishing a dynamic and emotionally resonant relationship between the user and the program. Our technology enables users to express themselves through voice, creating a distinctive and emotionally resonant interaction experience. It does this by merging signal processing, machine learning, and real-time analysis. The technology recognizes and categorizes emotions when users issue voice commands, giving them the option to pick between their original requests and recommendations that are emotionally consistent. This abstraction symbolizes a paradigm shift in how people and technology interact, where technology adjusts to users' emotional needs and uses voice and music to create a personalized and emotionally rich journey. In a word, my contribution represents the convergence of emotion-aware technology with user-driven interaction, leading to a system that communicates with and caters to users' emotions through voice-based commands and playlist creation.

Key words: -emotion based music recommendations, personalize music recommendations,

Table of Contents

1	Declaration	2
2	Abstract	3
3	Introduction	6
3.1	<i>Background & literature survey</i>	7
3.2	<i>Research Gap</i>	9
3.3	<i>Research Problem</i>	11
4	Objective	12
4.1	<i>Main Objective</i>	12
4.2	<i>Sub Objectives</i>	12
5	Requirements	13
5.1	<i>Functional Requirements</i>	13
5.2	<i>Non-functional requirements</i>	14
6	Methodology	15
6.1	<i>Overall System Diagram</i>	16
6.2	<i>Component overview Diagram</i>	17
6.3	<i>Gantt Chart</i>	18
6.4	<i>Work Breakdown Chart</i>	19
7	Limitation & challenges	20
8	Testing plan	22
9	Budget and Commercialization	23
10	REFERENCES	25
11	Appendix	26
11.1	<i>Plagiarism test</i>	26

List of figures

Figure 1 - System Overview Diagram	16
Figure 2 - component Overview Diagram	17
Figure 3 – grant chart.....	18
Figure 4 – work breakdown chart.....	19

List of Tables

Table 1 - Subscription Types	23
Table 2 - Budget Plan3	24

List of Abbreviations

Abbreviations	Description
AI	Artificial Intelligence
ML	Machine Learning
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network

3 Introduction

The seamless integration of technology and emotions has emerged as a critical frontier in today's context of human-computer interaction, providing richer, more immersive experiences. As technology advances, so does our ability to create systems that not only recognize but also respond to human emotions in real time. This paper makes an important contribution at the intersection of signal processing, machine learning, and personalized content curation by introducing an innovative approach that uses voice-based emotion detection to create personalized playlists that resonate with users' emotional states.

The ability to recognize and respond to human emotions accurately has far-reaching implications for a variety of applications, ranging from entertainment and mental well-being to immersive narratives and uses in therapy. Consider a situation in which an individual's emotional state, as captured by the distinct frequencies and intricacies of their voice, serves as the foundation for a tech-driven encounter that customizes its responses and offerings to the user's emotional state. This study gives light on the path to fulfilling this vision, emphasizing the dynamic interplay of technology and emotion.

In essence, our suggested system uses a complex array of advanced signal processing techniques to analyse the tiny nuances of users' vocal frequencies in real-time. We intend to have meticulously crafted an emotion detection model that demonstrates a remarkable proficiency in categorizing an extensive frequency of emotional states, all extracted from the combination of user voice inputs, through the intricate combining of revolutionary machine learning algorithms. Importantly, this model functions in the background, effectively understanding and decoding emotions while users interact with the application's interface intuitively. However, the main strength of our technology is its capacity to go beyond simple emotion recognition. It incorporates an additional level of sympathetic intelligence, delivering contextually aware recommendations as well as intimately selected playlist choices.

In conclusion, our work highlights the revolutionary potential of the cohesive marriage of technology and human emotion. Our system represents an ambitious step toward a future in which technology is not merely a passive observer of emotions, but an active, empathetic participant forging deep, resonant connections through its intuitive understanding of human sentiment, thanks to the fusion of advanced signal processing, intricate machine learning paradigms, and customized curation of content.

3.1 Background & literature survey

In the area of human-computer interaction, the combination of emotion recognition with personalized suggestion is gaining attention. The capacity to reliably determine users' emotional states from voice inputs and then deliver contextually appropriate playlist suggestions has the potential to transform human interactions with technology. This section begins with an introduction of the background and context of voice-based emotion recognition and playlist construction, and then moves on to a detailed literature review that sheds light on key research contributions, problems, and present gaps.

1. Voice-Based Emotion Detection

The human voice is a strong tool for communicating emotions, with nuances in pitch, tone, rhythm, and other acoustic characteristics. Voice-based emotion recognition research has advanced greatly, with studies utilizing approaches such as Mel-Frequency Cepstral Coefficients (MFCCs), prosodic characteristics, and deep learning models ([1]; [2]). The results of these studies highlight the possibility for properly recognizing emotional states in real time from speech signals.

2. Personalized Playlist Creation

Personalized content suggestion systems have evolved to accommodate the preferences of consumers, increasing engagement and pleasure. Mounika and Charitha's [3] work is an example of the integration of sound signals and emotions for mood-enhancing music suggestions. However, present research frequently focuses on stand-alone suggestions that do not include real-time emotion detection.

3. Contextual Recommendations and Seamless Integration

Despite advances in voice-based emotion recognition as well as customized recommendation systems, there is still a research gap in the seamless integration of these technologies for real-time engagement. The integration challenge has several dimensions, including creating an intuitive user interface for voice input recognition [2], algorithms that generate emotionally coherent playlists [3], and responding contextually to users' emotional cues [1].

4. User Engagement and Satisfaction

A research gap exists in the user-centric element of emotion-aware technology. Although the research evaluated touch on user experiences, there is a lack of a full empirical analysis of how the seamless integration of emotion recognition and playlist building affects user engagement, contentment, and emotional connection.

5. cohesive playlist generation

The problem of creating emotionally cohesive playlists that represent customers' emotional experiences is yet unsolved. Existing research focuses on mood-boosting recommendations [3] however, there is a gap in research on keeping emotional consistency while respecting user preferences.

6. Proposed Research Contribution:

By establishing an integrated system that smoothly blends voice-based emotion recognition with customized playlist production, this research attempts to bridge the observed gaps. Deep learning models, context-aware interaction interfaces, and ethical concerns will be used in the proposed system to provide a transformational interaction experience that matches with users' emotional states, improves user engagement, and builds emotional ties with technology.

In conclusion, combining voice-based emotion recognition with playlist creating has the possibility of generating sympathetic and contextually relevant user experiences. The research begins by addressing the highlighted deficiencies by creating an innovative system that modifies how people engage with technology, so generating a new paradigm in human-computer interaction.

3.2 Research Gap

According to my Literature Review, music selection and personalization are now hot topics among researchers. Our ultimate objective is to create a multi-model method to offer tailored music playlists and assess users' current sentiment based on a given voice clip to improve accuracy and personalization. It has the potential to dominate the market among all other existing music players. According to my Literature Review, music selection and personalization are now hot topics among researchers. Our ultimate objective is to create a multi-model method to offer tailored music playlists and assess users' current sentiment based on a given voice clip to improve accuracy and personalization. It has the potential to dominate the market among all other existing music players.

While we reviewed research articles, they give useful insights into voice-based emotion recognition and music selection but there is a huge research gap in the seamless integration of these technologies to produce a holistic and emotionally intelligent engagement experience. Existing research focuses primarily on individual aspects—for example, emotion detection or music recommendation—rather than addressing the complexities of real-time interaction, coherent playlist generation, contextual recommendations, user engagement, and generalization across diverse contexts. This void necessitates the creation of a complete system that connects these elements to build an integrated platform that recognizes and responds to users' emotional states while delivering customized and contextually relevant content.

The research gap can be defined specifically as follows

1. Holistic Integration of Components

While the articles provided touch on emotion recognition and music selection, there remains a research gap in developing an integrated system that easily integrates both of these aspects. The aim is to create a platform that properly identifies emotions in real time and converts this knowledge into emotionally appropriate music recommendations.

2. Real-Time Interaction and Contextual recommendations

None of the papers reviewed thoroughly address the real-time interaction component, in which the system dynamically responds to recognized emotions in users and provides contextually

sensitive suggestions. This research gap demands the development of an interface that engages users in meaningful interactions and modifies its replies based on the emotional indicators of the users.

3. User-Centric Engagement and Satisfaction

Although user experience is inferred in the studies examined, there is a research gap in experimentally studying how the integration of emotion-aware technology improves user engagement, contentment, and emotional connection with the system. This gap necessitates user research to measure the effects on user experiences.

4. User-Centric Engagement and Satisfaction

Although user experience appears in the reviewed papers, there is a research gap in experimentally studying how the integration of emotion-aware technology improves user engagement, contentment, and emotional connection with the system. This gap necessitates user research to measure the effects on user experiences.

5. Coherent and Diverse Playlist Generation

While the studies explore emotion-linked recommendations, there is a research gap for creating playlists that are both emotionally coherent and diverse, allowing for an emotional journey across the suggested content. This research gap emphasizes the difficulty of developing an algorithm that balances emotion alignment with user preferences.

Finally, the research gap is in the creation of an integrated voice-based emotion recognition and playlist-generating system that seamlessly blends real-time interaction, coherent playlist curation, contextual suggestions, and user engagement. The proposed research intends to bridge this gap by developing a holistic system that integrates all of these aspects, boosting human-computer connection through empathetic interactions.

3.3 Research Problem

As mentioned before, music and suitable music suggestions play an important part in all aspects of life nowadays. There are several successful mobile applications and IoT devices in use throughout the world, including Spotify, iTunes, Alexa, Amazon Music, and others. These programs have been enormously popular and are still actively used throughout the world. Unlike previous study, I am worried with an unexplored region in music streaming app development. Music listening habits are strongly tied to users' current moods. As an example, if I use a music streaming service if my mood is happy at the moment, I will undoubtedly hear pleasant mood songs.

The primary goal of the system we propose is to overcome this issue by introducing an innovative method. My aim is to create an audio categorization model using convolutional neural networks (CNNs). This model will analyze voice patterns and frequencies to properly predict users' moods using deep learning technologies. We will be able to construct a unified mobile app by merging our emotion classification model with our music recommendation system.

4 Objective

4.1 Main Objective

The primary objective of the proposed system is to give the user a customized music playlist. The major purpose is to improve the usability and accuracy of music recommendations. There are difficult actions to take in order to reach this aim. They are taking a high-quality selfie, assessing the facial features, studying the surroundings, taking a voice command to detect the user's sentiment, and generating a customized playlist utilizing these inputs. I came up with a sub-goal for the previously mentioned system: categorize emotional states based on voice inputs.

4.2 Sub Objectives

I divided my component into a few particular objectives, which I have listed below.

- Collect vocal data inputs into emotion classification model to identify various emotions.
- converts raw voice signals into a set of acoustic features using feature extraction.
- Train the emotion classification model.
- Evaluate model performance
- Optimize for real-time processing
- Iterative model improvement
- Train the music recommender model
- Develop the mobile application accordingly.
- Integrate above features into the developed mobile application.

5 Requirements

5.1 Functional Requirements

The most important aspect for the developer to build up the solution are functional requirements. The primary pillars of a system are functional requirements, which specify the system's ultimate objective and user expectations. During my research, I discovered the following critical functional needs.

- Input the real-time voice clip

Firstly, interface should provide a mic to get a user's voice clip as a input and it should be high quality for further processing. Since we consider the voice frequency the voice clip should be uploaded real-time.



- System should be able to extract emotion based on voice clip.
- System should generate a personalized playlist accordingly.
- User should be able to listen to playlist from the software application.
- System should be able to track and train the models for future recommendations.

5.2 Non-functional requirements

- **Performance**

Real-Time Processing: To deliver a smooth real-time experience, emotion assessment and playlist production should occur within milliseconds.

Scalability: The system should be able to handle an increase in user load without sacrificing performance.

- **Accuracy**

Accuracy of Emotion Detection: The emotion classification model should achieve a high rate of correct identification of diverse emotional states from speech inputs.

Playlist Relevance: To improve user happiness, the created playlists should precisely correspond with the identified emotions of the users.

- **Security and privacy**

Data Security: The system must follow data security rules and ensure that user voice data is private and securely stored.

User Data Protection: Put in place safeguards to protect user data and prevent unwanted access.

- **Usability**

User-Friendly Interface: The user interface should be simple and easy to use in order to provide a pleasant user experience.

clear Feedback: To increase user engagement, provide explicit feedback on identified emotions and music recommendations.

6 Methodology

The mechanism for the suggested Multi model music recommender system that improves the user's customized experience is described below. There are tools and technologies that will be employed in the system's implementation.

Algorithms

- CNN
- RCNN
- Hybrid models

(Algorithm selection may vary throughout implementation dependent on the optimal method)

Integrated development environment (IDE)

- PyCharm or Anaconda

Databases

- MongoDB
- Firebase

Backend

- Feature extraction
- Voice frequency Analysis
- Python - handle algorithms OpenCV framework

Overall, the process described above comprises data collecting, data preprocessing, creating a machine learning model, testing, and training the model, deploying the system, and maintaining and updating the system to enhance its performance. The success of the system is determined by the quality and amount of data used to train the model, the accuracy and dependability of the machine learning algorithms, and the system's efficacy in categorizing user information.

6.1 Overall System Diagram

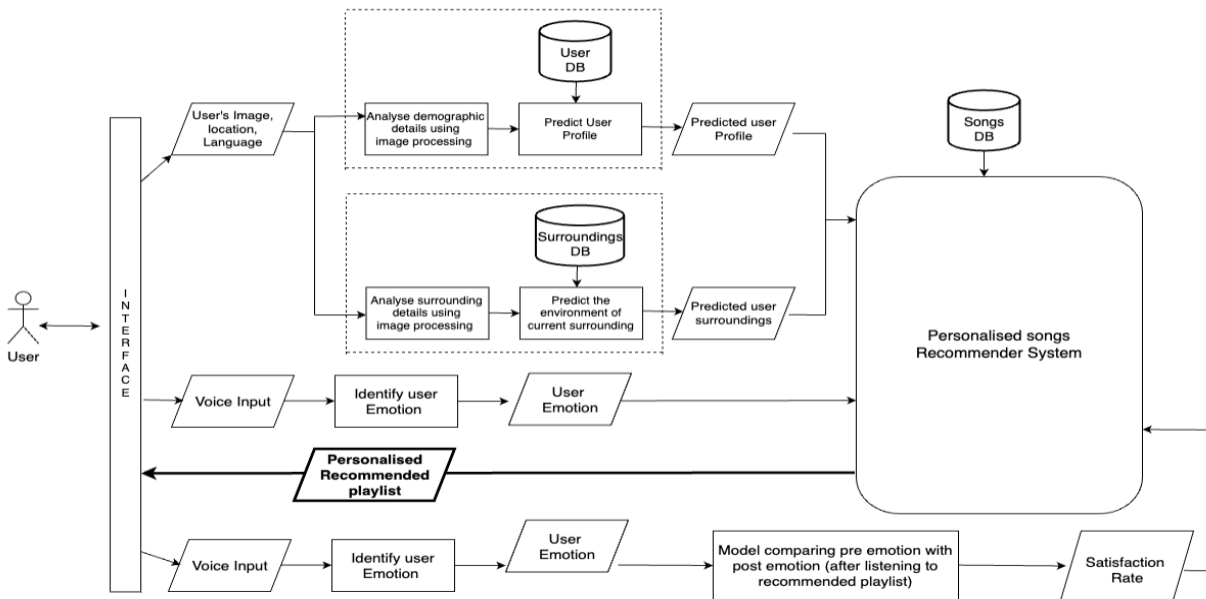


Figure 1 - System Overview Diagram

To alleviate the cold start problem, the user must first capture a photograph of their face in order to collect facial specifications to forecast their age, gender, geographical location (county / island), region, and nationality. The application then often captures user input using a variety of methods, such as speech recognition, picture processing to extract the mood / emotion, and the surroundings. Then there is a new component to assess the post feelings after listening to the recommended music to improve the recommendation's accuracy in the future. After combining all of these models, algorithms, and subcomponents, the user will have individualized play lists for various moods and events (Figure 5).

6.2 Component overview Diagram

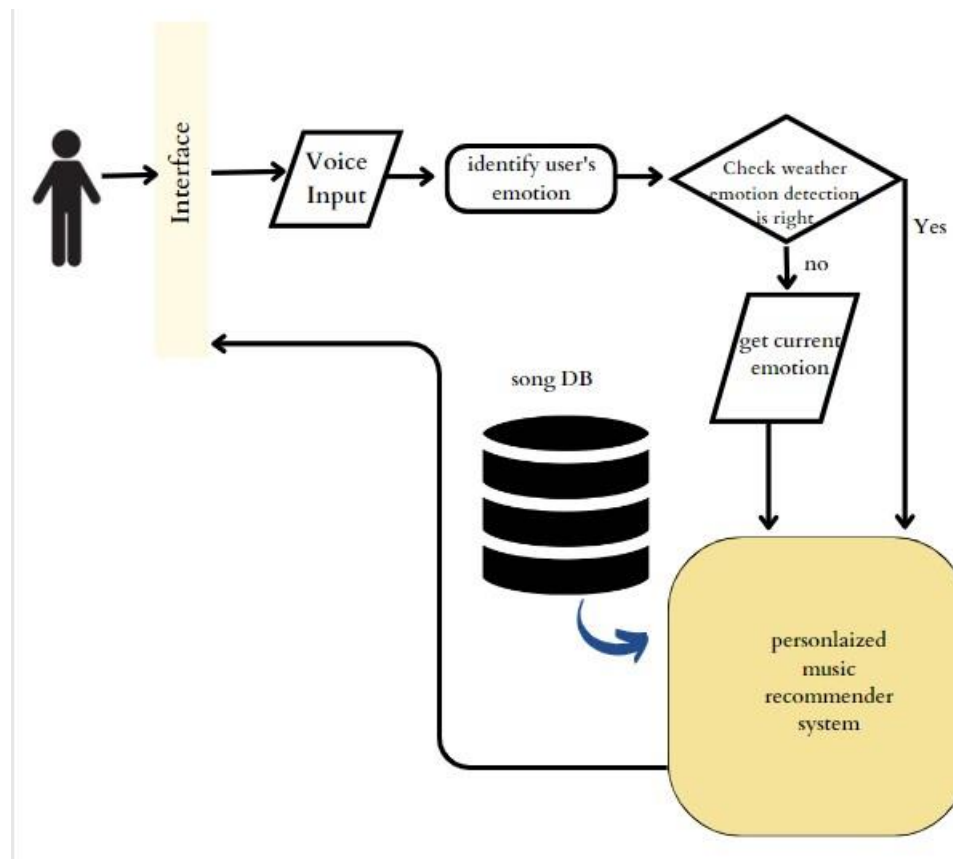


Figure 2 - component Overview Diagram

6.3 Gantt Chart

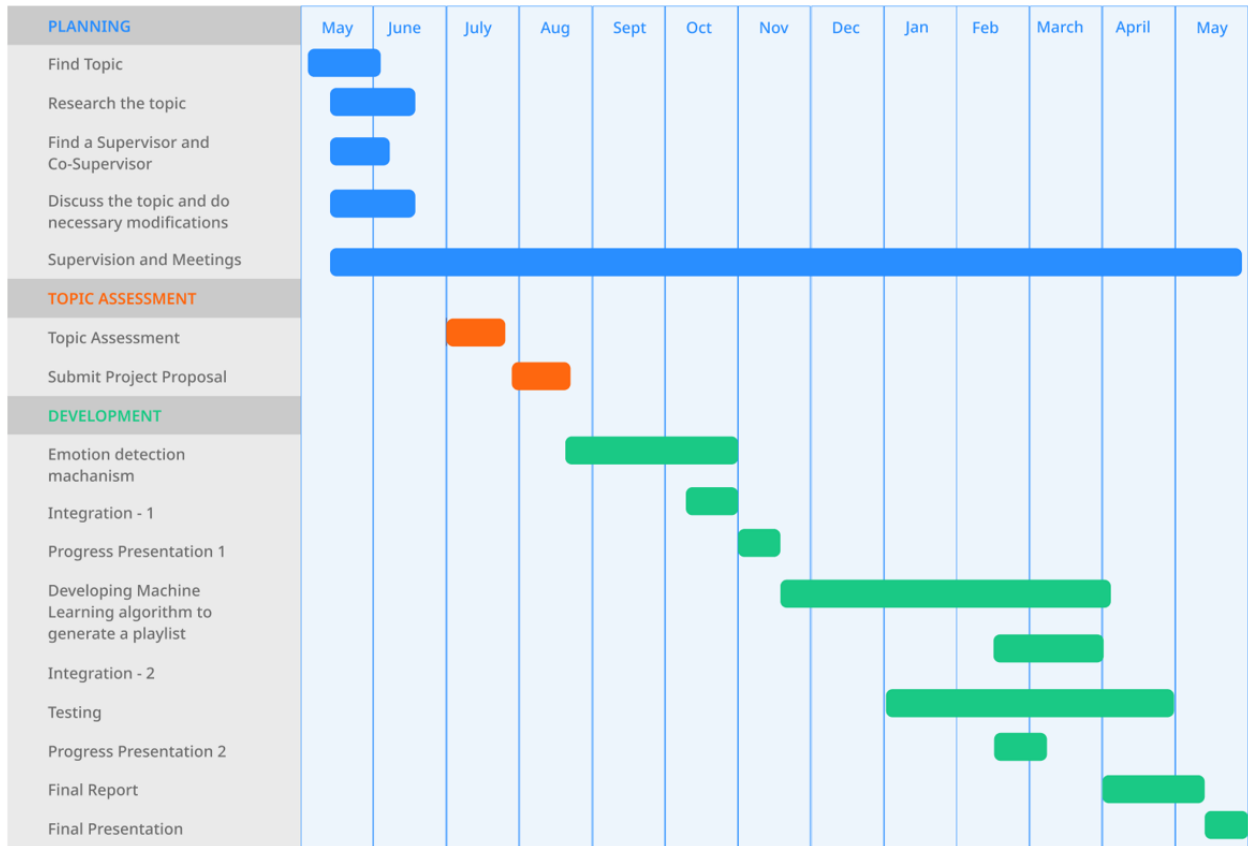


Figure 3 – grant chart

6.4 Work Breakdown Chart



Figure 4 – work breakdown chart

7 Limitation & challenges

By obtaining a real-time voice clip from the user, my research scenario gives a fascinating and new method to addressing the issues of customizing music suggestions. However, there are various restrictions and obstacles that I confront when applying the aforesaid approach, as is always the case.

- **Data Bias**

The model's success is heavily dependent on the variety and representativeness of the training dataset. If the dataset is skewed toward specific demographics or emotional expressions, the model's performance may differ among user groups.

- **Emotional Subjectivity**

Emotions are complicated and subjective phenomena that might be difficult to adequately represent using only speech signals. The model may struggle with emotional expression subtleties and cultural variances.

- **Ambiguity**

Certain emotional states may present identically in vocal features, resulting in categorization uncertainty. Differentiating between mild emotions such as "surprise" and "fear" may be challenging.

- **Real-Time Processing**

Achieving genuine real-time processing can be difficult, including optimization and hardware considerations in order to ensure minimal latency while retaining accurate emotion recognition.

Furthermore, there are some challenges I identified through my research component. They are;

- **Ethical Considerations**

Concerns concerning user privacy, data security, and potential abuse arise when dealing with voice data. Implementing strong privacy safeguards and guaranteeing transparency become critical.

- **Bias and Fairness**

It is difficult to create an emotion recognition algorithm that is devoid of biases and handles all demographic groups equitably. It is difficult to address inherent biases in data collecting and model training.

- **Interpretability**

Deep learning algorithms used for emotion detection might be challenging to understand. Understanding how the model makes predictions, particularly for complex inputs like speech, remains difficult.

- **User Variability**

Because of factors such as mood swings, health issues, or external influences, users' voices may vary, affecting the model's ability to recognize emotions effectively.

- **Real-World Noise**

Voice signals gathered in real-world contexts may contain background noise, variable recording circumstances, and artifacts, all of which can impair model performance.

- **Engagement and Feedback Loop**

It is critical to provide an engaging user experience that encourages people to submit correct emotional inputs. Furthermore, creating a feedback loop for continual model improvement necessitates active user engagement.

8 Testing plan

Optimum music mobile application testing will be place at various phases of the project. It aids in the discovery of issues in each component and makes it simple to solve them discretely rather than the entire project. As a result, the entire testing process will be comprised of multiple phases and methods.

1. Unit Testing

Unit testing will be performed individually for each component, such as the emotion classification model and the music recommendation model, in which case the issue in each component will be detected and repaired. The researchers will concentrate on two primary factors here:

- a) component performance testing.
- b) Component accuracy testing.

Both testing methods mentioned above will be significant for the general usability of the final product.

2.integration testing

Component integration will be a primary responsibility of this research project; components will be merged one by one and tested concurrently since integration might create major defects in the system.

3. Final Testing

Final testing will be performed to ensure that the system is running smoothly. The completed product will be tested using various test scenarios and sample data. The mobile application will be distributed to certain selected beta users as the second round of final testing, and their input will be collected.

The beta users are going to evaluate the mobile application's user experience, and the researchers will fine-tune the app's user interface to create a better end-user experience.

9 Budget and Commercialization

As music players are used daily by people this project can have a huge commercial value. People are willing to pay a fair amount for a better music player. As there are some market leaders like Spotify, Apple Music and Deezer available, the price of the music player must be competitive and fair. Almost every (Spotify, Apple Music, and Deezer) player costs approximately \$10 per month. Some people find it expensive and not worthy enough to buy that subscription model, Therefore the below subscription model is proposed for the commercialization of this mobile app.

Mainly this app will be available in the major app stores such as Google Play Store, Apple AppStore, and Huawei App Gallery. There will be two models for this mobile,

Table 1 - Subscription Types

	Free version	Paid version (<\$10/month)
Advertisements	Yes Advertisement networks such as Google AdSense /Admob will run on this version of the mobile app	No No advertisements will be displayed in the mobile app
Monthly charges for the users.	No Revenue will be generated from the advertisements showed to the user while the user is using the mobile application.	Yes Revenue will be generated from the monthly charges paid by the user.
Features	All features	All features

The final mobile application will be focused on different user groups; therefore, it will be marketed to each user group using different methods,

1. The younger generation – Social media advertisements (Facebook/Instagram)
2. Mature generation – Newspapers etc.
3. Musical experts – Face-to-face demonstrations with musical associations/groups.

Below is the budget that has been planned for the project, Charges will be changed according to time to time, and final charges will be based on the consumption of the resource used in the cloud environment.

Table 2 - Budget Plan³

Description	Amount (USD)
1. AWS Cloud database (S3) for facial images <ul style="list-style-type: none"> • To store user images collected through mobile app. 	0.023 per GB / Month
2. AWS Cloud database (EFS) for user demographic data. <ul style="list-style-type: none"> • To store demographic data of the users. 	0.30 per GB / Month
3. AWS glacier to store User logging from the mobile application.	Storing = \$0.004 per GB / Month Retrieving = \$0.01 per GB
4. Paper Publications and documentation.	50-150

10 REFERENCES

- [1] K. R. Nambiar and S. Palaniswamy, "Speech Emotion Based Music Recommendation," 2022 3rd International Conference for Emerging Technology (INCET), Belgaum, India, 2022, pp. 1-6, doi: 10.1109/INCET54531.2022.9824457.
- [2] K. S. Krupa, G. Ambara, K. Rai and S. Choudhury, "Emotion aware Smart Music Recommender System using Two Level CNN," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2020, pp. 1322-1327, doi: 10.1109/ICSSIT48917.2020.9214164.
- [3] V. Mounika and Y. Charitha, "Mood -Enhancing Music Recommendation System based on Audio Signals and Emotions," 2023 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 2023, pp. 1766-1772, doi:10.1109/ICICT57646.2023.10134211.

11 Appendix

11.1 Plagiarism test

The screenshot shows a web browser window displaying a Turnitin Feedback Studio report. The browser's address bar shows the URL: https://ev.turnitin.com/app/carta/en_us/?student_user=1&lang=en_us&u=1144538920&o=2041698873&ts=. The page header includes the "feedback studio" logo, the user name "Sameeri Subasinghe", and the document title "Malshan - Report". The main content area displays the following text:

Multi-Model Approach to Recommend Personalized Music Playlist

Project ID: TMP – 2023 – 24 - 065

Project Proposal Report

Gunasekara C.M – IT20665852

The bottom of the report shows a status bar with "Page: 1 of 26" and "Word Count: 3571". On the right side, there is a vertical toolbar with icons for various functions, including a red icon with the number "19". The Windows taskbar at the bottom of the screen shows the time as 4:56 PM on 8/25/2023.