A PERSONAL TRAINER APP TO SELF-TRAIN AND IMPROVE PRESENTATION SKILLS

21_22J_002

Project Proposal Report

Wanigasinghe N.T

B.Sc. (Hons) Degree in Software Engineering.

Department of Computer Science and Software Engineering

Sri Lanka Institute of Information Technology

Sri Lanka

June 2021

A PERSONAL TRAINER APP TO SELF-TRAIN AND IMPROVE PRESENTATION SKILLS

21_22J_002

Project Proposal Report

B.Sc. (Hons) Degree in Software Engineering.

Department of Computer Science and Software Engineering

Sri Lanka Institute of Information Technology

Sri Lanka

June 2021

Declaration of The Candidate & Supervisor

I declare that this is my work, and this proposal does not incorporate without acknowledgment 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 acknowledgment is made in the text.

Name	Student ID	Signature
Wanigasinghe N. T	IT18229912	

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

Signature of the supervisor   Date

Acknowledgment

I would like to thank my research supervisors, Dr. Shyam Reyal and Co-supervisor Ms. Samanthi Eranga Siriwardene, and Ms. Veerandi Kulasekara for supporting me in this research. Dr. Shyam, Ms. Samanthi, and Ms. Veerandi have been supportive and have given me the freedom to pursue various projects without objection. They have also provided insightful discussions about the research. I thank all the present members of the group: Himanshi, Chathurdi, and Kethmini, and Ms. Manuri for her good advice during the past few days.

I am also deeply thankful to my informants. I want to acknowledge and appreciate their help and transparency during my research. Their information has helped me complete this proposal. Finally, I would like to express my gratitude to all the people who help me by providing their valuable assistance and time during this research.

Abstract

Presentation is a fundamental means of communicating not only words but also a vast range of human emotions. Consequently, speech processing applications, like human-machine interfacing and speech recognition, could benefit from the introduction of a reliable method for automatic recognition of human emotions through the presentation. Speaker recognition is a method that recognizes a speaker from the characteristics of a voice. Speaker recognition technologies are widely utilized in many domains. Prosody in speech plays a range of functions, such as sentence mode, narrow focus word-stress by pitch-accent or dynamic-accent, prosodic styles (e.g., storytelling, news-reading), and emotional speech. Prosody modeling is an important component in modern text-to-speech (TTS) frameworks. By explicitly providing prosody features to the TTS model, the design of synthesized utterances can be controlled by the system. By using natural language processing techniques, the application will predict the user's emotions and prosody. While they are practicing, the system will analyze the speaking pattern using automatic speech recognition and provide feedbacks and points for their presentation. So, users will be able to correct their emotions and prosodic styles while they are presenting. Thus, user can improve their presentation skills from their home with this personal trainer app. This component performs in terms of emotions and prosody naturalness on the presenter's speech by the audiovisual.

Keywords: emotions, prosody, tone, pitch, text-to-speech, natural language processing, automatic speech recognition

Table of Contents

Declaration of The Candidate & Supervisor 3

Acknowledgment 4

Abstract 5

Table of Contents 6

List of Figures 7

List of Tables 7

List of Abbreviations 7

1 Introduction 8

1.1 Background & Literature Survey 8

1.2 Literature Survey 11

1.3 Research Gap 15

1.4 Research Problem 17

2 Objectives 19

2.1 Main Objective 19

2.2 Specific Objectives 19

3 Methodology 20

3.1 Project Overview 20

3.2 System Overview Diagram 21

3.3 System Overview 22

3.4 Suggest user, the match or mismatch between topic tone and emotions used to present the story using audio analysis - Individual Component 24

Architectural Design 24

Architectural Diagram 26

3.4.1 The flow of the system – Individual Component 27

3.4.2 The flow of the prosody modeling – Individual Component 30

3.5 Work Break Down Chart and Gantt Chart 33

3.6 Requirement Analysis 35

Functional Requirements 35

Non-Functional Requirements 35

4 Budget and Budget Justification 37

5 References 38

List of Figures

Figure 1 Summary of responses who like to self-evaluate presentation 9
Figure 2 Summary of responses who consider proper tonality while presenting 10
Figure 3 Summary of responses who consider emotional control impact while presenting 10
Figure 4 Summary of responses who followed tool for predict presentation skills 17
Figure 5 Summary of responses average number for practice a presentation 17
Figure 6 Project overview diagram 20
Figure 7 Architectural Diagram – individual component 24
Figure 8 System Architectural Diagram 25
Figure 9 Software development life cycle 26
Figure 10 Overview of model architecture and prosody modeling components [7] 29
Figure 11 Prosody modeling architecture [7] 30
Figure 12 General pipeline for TTS systems with prosody modeling [7] 31
Figure 13 Sign-up for the System - Wireframe 32
Figure 14 Sign-in for the system - Wireframe 32
Figure 15 - Main Home page of the system - Wireframe 33
Figure 16 - Output of the system | Video Analysis – Wireframe 34
Figure 17 - Output of the system | Slide Analysis - Wireframe 34
Figure 18 Work breakdown chart 36
Figure 19 Project Gantt chart 37

List of Tables

Table 1 Research Comparison Table 16

Table 2 Budget and budget justification 31

List of Abbreviations

TTS - Text to Speech

STT - Speech to Text

API - Application Programming Interface

NLP - Natural Language Processing

DNN - Deep Neural Network

ASR - Automatic Speech Recognition

SDK - Software Development Kit

1 Introduction

1.1 Background & Literature Survey

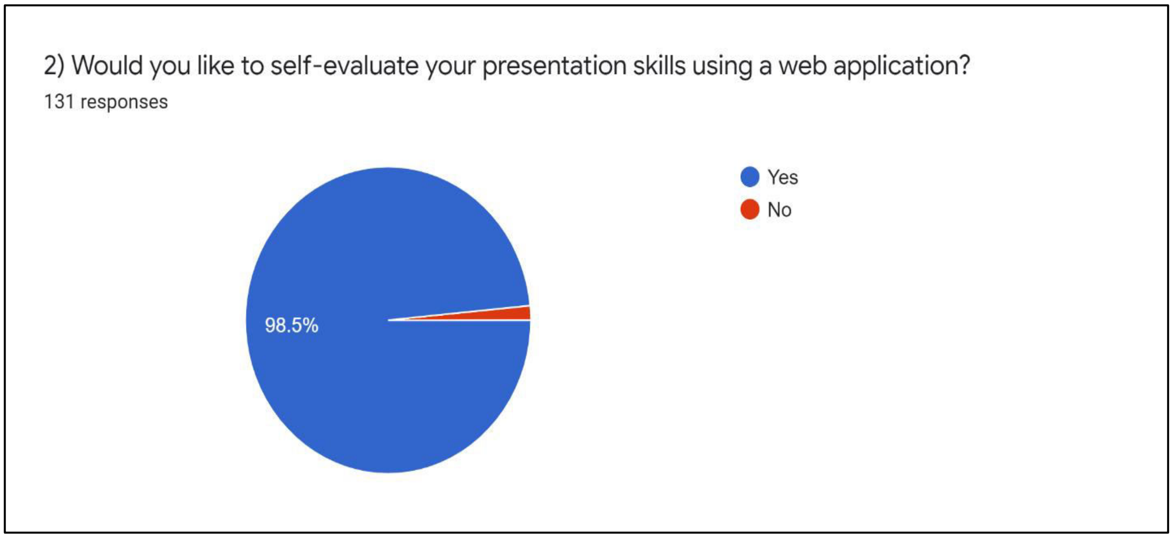
Delivering a presentation can cause even the most confidence among us to break a sweat. Whether you're giving a formal presentation, giving a public speech, or leading a group discussion, expressing your message with poise, confidence, and conviction is critical for engaging with your audience and motivating them by establishing trust, interest, and credibility. It might be at work, university, social occasion, or in connection with the class or voluntary activity.

Presenters must put proven tactics and tools into practice, test out creative ideas, and learn to communicate clearly and confidently through oral presentations and small group exercises. Discover the powerful impact of storytelling and practical persuasion skills to authentically illustrate the message. Presenter/Speaker must effectively organize materials to blend analytical and emotional content into a compelling story and incorporate dynamic introductions and memorable endings into their presentations. And a presentation must be delivered with the skills needed to engage, inform, and inspire others.

Interactions between humans and machines are increasing every day. Therefore, the system which provides self-assessment for the presenter to monitor gaps between speech delivering and the effects and self-reflection is helpful for presenters to attain higher presentation performance.

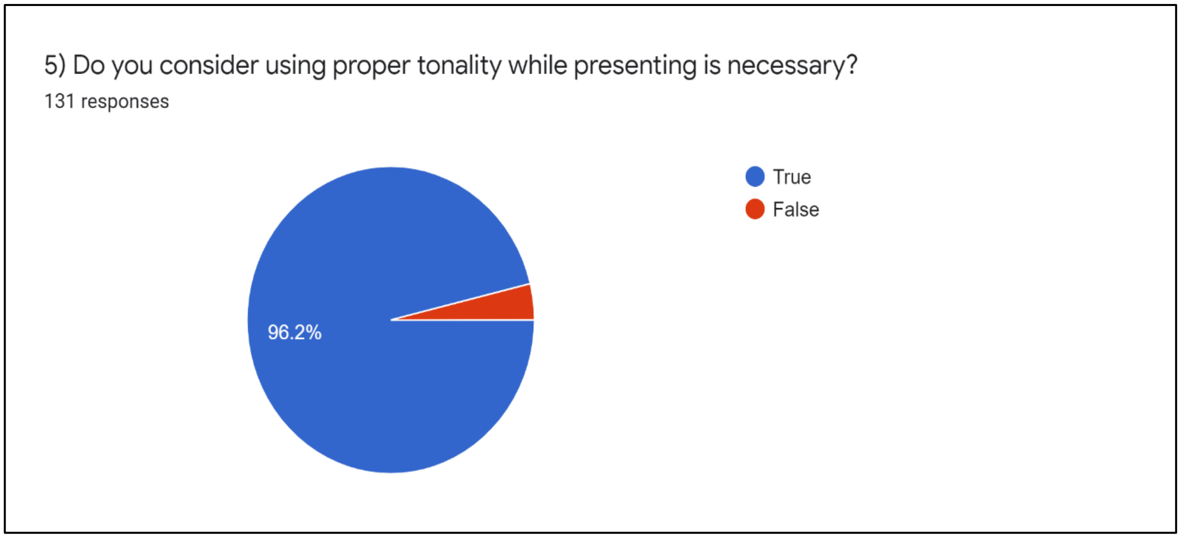
Improving the quality of human-machine voice interaction is the primary goal of automatic speech emotion recognition, where speech communication is more efficient, common, and fast in order to pass information between human beings and machines. There are many applications and systems of voice-based automatic emotional speech recognition tools in real life. Examples include voice portals, assessing the quality of the service provided by call centers, surveillance to detect negative emotions, conversation analysis, gaming, and many more.

The accurate recognition of emotions from speech faces many challenges: it is exceedingly difficult to reliably process speech in real-time. Noise and reverberations are examples of these challenges. Differences in speakers, sentences, and speech rates directly affect the pitch and energy contours of speech, which are the most extracted speech features. Determining which speech features are most effective in the classification of emotions remains a challenge. Also, nowadays, people like to self-evaluate their presentation skills using digital software or tool. People who like to self-evaluate their presentation is shown in Figure 1. And when presenting a presentation, people are thinking about their tone controlling, emotions expressing, and how their voice impacts their presentation. As shown in Figure 2 and Figure 3.


2) Would you like to self-evaluate your presentation skills using a web application?

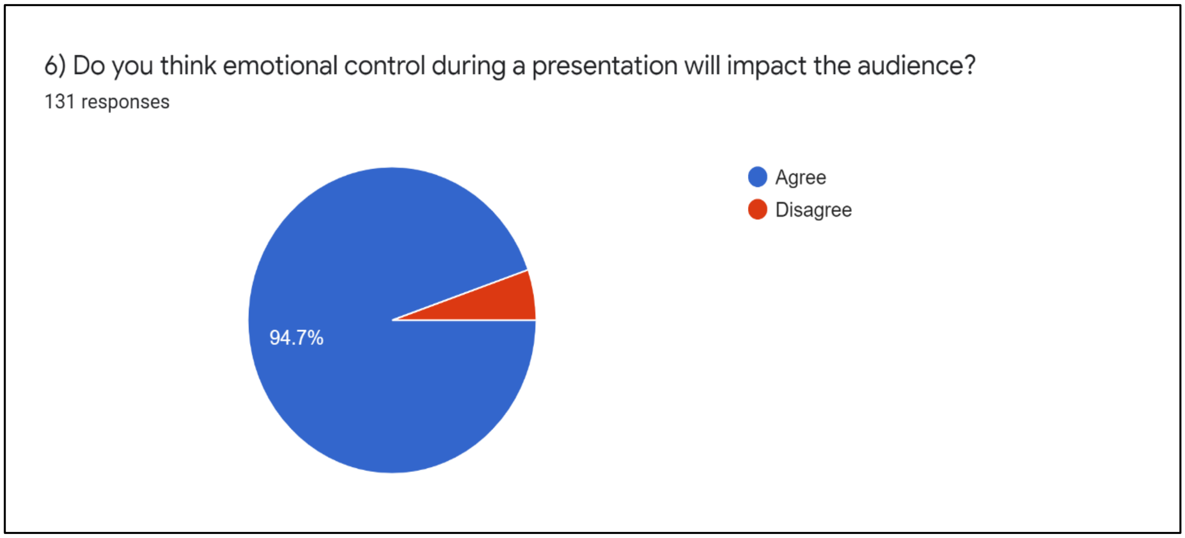
131 responses

Figure 1 Summary of responses who like to self-evaluate presentation


5) Do you consider using proper tonality while presenting is necessary?

131 responses

Figure 2 Summary of responses who consider proper tonality while presenting


6) Do you think emotional control during a presentation will impact the audience?

131 responses

Figure 3 Summary of responses who consider emotional control impact while presenting

For these requirements, we propose a solution based on Natural Language Processing and Computer Vision techniques to cater to the requirement for the user to do a presentation beforehand using a mobile responsive web application. The proposed system will assist anyone who needed to practice a presentation beforehand by Identifying the emotions, body language using video and audio, by checking the pronunciations and vocabulary, and by checking presentation quality. Overall, the system will give a rating and feedback to the presenter about the performance before they go into the actual presentation.

1.2 Literature Survey

Checking the applicability of emotions, tone, and prosody for the speech is the most essential part of the implementation of the system. During this literature review, I even have studied various research papers to spot existing research aspects of applicability of emotions and prosody for Automatic speech recognition (ASR).

During the search, several research papers have been found above topic. In 2017, V.V.V. Raju et al. [1] has mentioned, emotions in speech majorly inflict the changes within the prosody parameters of pitch, duration, and energy. Tuning the modification factors values for the relative differences between the neutral and emotional data sets outcomes in prosody alteration of pitch, duration, and energy [1]. For speech recognition, the neutral version of the emotive speech is lead to achieving uniform and non-uniform prosody modification algorithms. During the study, the IITKGP-SESC corpus is used for building the ASR system. The speech recognition system for emotions (anger, happiness, and compassion) is evaluated. An improvement within the performance of ASR is observed when the prosody modified emotive utterance is used for speech recognition in place of original emotive utterance. In that paper, the uniform and non-uniform prosody modification method is explored for generating a neutral version of an emotive utterance to enhance the performance of the ASR system [1]. The database consisting of spontaneous natural emotions is required for the analysis of emotional speech. During this paper, they have been used standard simulated databases such as the Danish emotion speech database (DES) and Berlin emotion speech database (EMO-DB) [1]. The purpose of that paper is to obtain a neutral version of emotive speech and to assess the effectiveness of speech recognition systems on prosody adjusted speech.

In 2017, B. Sisman et al. [2] has mentioned, Speaker characteristics are embedded in spectral features as well as prosody features. So far, the spectral mapping mechanism is central to the study of voice conversion, and conversion of prosody is still a challenge to be improved. In addition, the research on how to assess the accuracy of prosody transformation in voice conversion is lacking. With this motivation, the researchers focus on research on analysis and evaluation of different prosody transformation techniques and propose a new analysis approach to achieve a more reliable prosody conversion evaluation [2]. In this research paper, I found the evaluation of prosody transformation the evaluation of Fundamental Frequency conversion. Frequency is an essential prosody feature that should be taken care of in a compressive voice conversion framework. So far, the evaluation of the converted prosody features is performed by looking at Pearson Correlation Coefficient and Root Mean Square Error (RMSE). Unfortunately, these techniques do not explicitly measure the Frequency alignment between the source and target signals. In here Dynamic Time Warping (DTW) is used assess the performance of a prosody transformation framework. DTW, is a well-known technique for measuring similarity between two temporal sequences which may vary in speed. Therefore, in this paper, I found a new technique to assess the accuracy of prosody transformation [2]. In human-computer interaction, the naturalness of communication is always dependent on the system's ability to recognize the emotions present in human speech. In an emotion recognition system, the features extracted from the speech signal play a vital role in the task. In 2018, V. V. Raju et al. [3] has mentioned, the best features for this task of emotion recognition are always dependent on the strong basis of relating the speech characteristics with the emotional state of the speaker. In this paper, the emotionally stressed regions are detected using the strength of the excitation (SOE) and the fundamental frequency of the speech signal. A linear kernel support vector machine (SVM) with a two-stage binary decision logic is used to further classify these emotionally stressed regions for the various emotions. This classifier is modeled using the differenced prosody features extracted by considering the relative difference of the prosody components between normal and emotionally stressed regions. This experimentation is carried out using the Berlin emotion speech (EMO-DB) database [3]. The evidence of SOE and frequency is used to identify the emotionally stressed and normal regions. The emotion recognition system is built utilizing an SVM classifier and the different prosody features derived from these regions. And the proposed approach produced a better performance over the existing emotion recognition systems [3].

In 2018, N. Ang, D. Bein et al. [4] has mentioned, when we analyze human voices, we can process audio data of a human speech and detects irregularities in pitch, volume, timbre, or word choice of the user, as well as perform sentiment analysis on the text extract from the audio file in order to detect the human's emotional state [4]. In 2016, V. V. V. Raju et al. [5] has mentioned, most ASR systems are trained using neutral speech and the performance of such systems degrade when tested with emotional speech [5]. The prosody features of the source emotional utterances are modified according to the target neutral utterances using the Flexible Analysis Synthesis Tool (FAST). In the FAST, Dynamic Time Warping (DTW) is used to align the source emotional and target neutral utterances. Components of the prosody like intonation, duration, and excitation source are manipulated to include the specified features into the source utterance. The modified (source emotional) utterances are then used for testing the ASR system which is trained using neutral speech. In this study, three emotions (compassion, happiness, and anger) are considered for the analysis [5].

In 2021, C. -M. Chien et al [7], has mentioned, prosody modeling is an essential component in modern text-to-speech (TTS) frameworks. In order to synthesize human-like speech utterances by TTS, it is important to model the variation in speech signals, including rhythm, intonation, and stress, etc. [7]. These characteristics, collectively referred to as prosody, are not included in the text transcripts, but they are critical for conveying information not explicitly stated in the texts. Providing additional prosody information to the TTS model is referred to as prosody modeling, which enables expressive and controllable speech synthesis. In this work, I found fine-grained prosody modeling, where the prosody of an utterance is represented as a sequence of prosody features instead of a single sentence-level prosody feature. Each fine-grained prosody feature encodes the prosody associated with a speech segment, such as a phoneme or a word. Aside from enabling local prosody control in speech synthesis, fine-grained prosody modeling also further reduces the complexity of the TTS task itself, since the information contained in each fine-grained prosody feature is explicitly assigned to a speech segment. This makes prosody modeling especially a key component in the non- autoregressive TTS framework [7].

Therefore, according to my literature survey, I hope to use tone, pitch, stress, intonation, and prosody for analysis of the emotions of the speech according to the [2] [3] [4] [5] [6]

[7] [15] articles. And, according to the [10] [11] [12] [14] articles, a deep neural network can be used to build an ASR system.

1.3 Research Gap

According to the research papers which I have read during these past few days it has caught my attention that few types of research have been conducted for targeting emotions and prosody detecting for a presentation. Many research papers have been conducted on emotions that are expressing during a speech [11] [3] [5] [12]. And also, many systems have been created for checking prosody while speaking but such an expert system has not yet been created to catch emotions and prosody when delivering a presentation. Therefore, when creating an application to predict presentation skills there should be a way to detect emotions and also prosody levels. But the exciting systems predict only limited types of emotions while speaking. Most of the time the exciting products measure the emotions of anger, disgust, fear, happiness, neutral, sadness, and surprise [10]. And most of prosody detection exciting application detects normal and stressed regions [3], compassion, happiness, and anger [5] of a speech. But we have considered both when we will create a system for predict presentation skills and get feedback for the presenter's emotions and prosodic features during presenting. Therefore, the best solution for catch both we can use the intonation, stress, rhythm, tone, and pitch of a voice while presenting the story. Therefore, the proposed system focuses on emotions and prosody detection for presenting and given feedback on how to tone controlling impact for their presentation. When I have read research papers, I found most of the time desktop applications are only working on computers. So, the proposed system is a mobile responsive web application. Then anyone can use it from anywhere. Because of that, there is a clear research gap between the point of use prosodic and emotions detecting and usage for exciting products.

So, the system needs to develop a complete package with all those components to predict emotions and prosodic styles and give feedback for their presentation. As shown below table presenter should be able to identify these emotional and prosodic components while their presenting.

	Solution	*Research* A [3]	*Research* B [9]	*Research* C [13]	*Research* D [14]	*Research* E [11]	*Research* F [5]	PRESENTLY
Detect	*Angry*	√	√		√	√	√	√
	Anxiety							√
	Desperation				√			√
	Sadness	√	√		√			√
	Fear				√			√
	Interest							√
	Amusement					√		√
	Neutral	√			√		√	√
	Pleasure	√				√	√	√
	Paper Reading		√	√				√
	Fluency, murmuring sound detection							√
	Track nervousness							√
	Personalized system							√
	Application giving feedback for presentation							√

Table 1 Research Comparison Table

1.4 Research Problem

Usually, in our day-to-day life, we face mostly business presentations and educational presentations in our working place or university. But what is the real value of having good presentation skills? If you are planning to apply for a job, the global job market is requiring good presentation skills from you. But for non-native English-speaking countries like Sri Lanka, India, etc. It is hard to speak the English language as a second language because of some reasons. The difficulties one may encounter when learning any language are, pronunciation, spelling, morphology, syntax, vocabulary, phraseology, and fear to speak.

To overcome this issue, there should be a way to improve your presentation skills by yourself. There are not many applications for predict your presentation skills and give feedback. So, we can see the result of *Figure 4,* most people do not use any automated application or system to predict their presentation and get feedback for their present.

People are accustomed to practicing presentations beforehand, preferably with a friend, roommate, or teammate who will listen, take notes, and offer suggestions and feedback afterward, and time is taken to do the presentations. As a result of *Figure 5,* most of the people are practicing their presentations 1-5 times before the actual presentation. If there is an application to use for these scenarios it will be especially useful and efficient.

When you are presenting you must do it professionally. For those expressing emotions and feelings for an audience is important. Book reading while presenting does not make a good presenter. So, that this system will help you to identify emotions and prosody style mistakes while you are presenting. The proposed solution will be a better opportunity for people who need to improve their presentation skills. Then they can face their presentation without afraid and nervousness. Our target is to maximize the presenting level of users and provide a personal trainer to self-improve their skills.

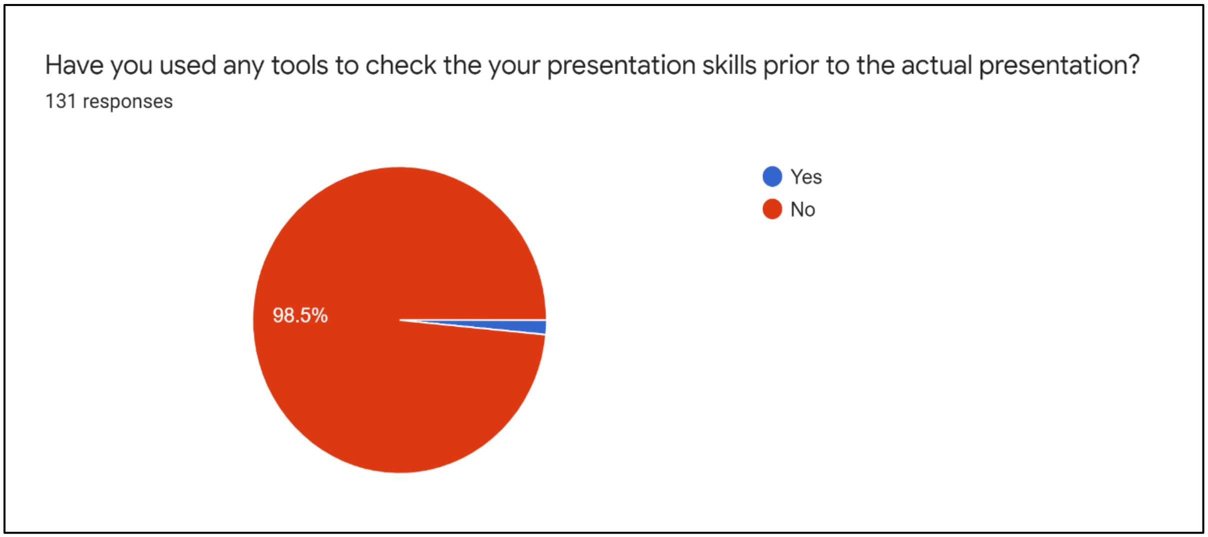


Figure 4 Summary of responses who followed tool for predict presentation skills

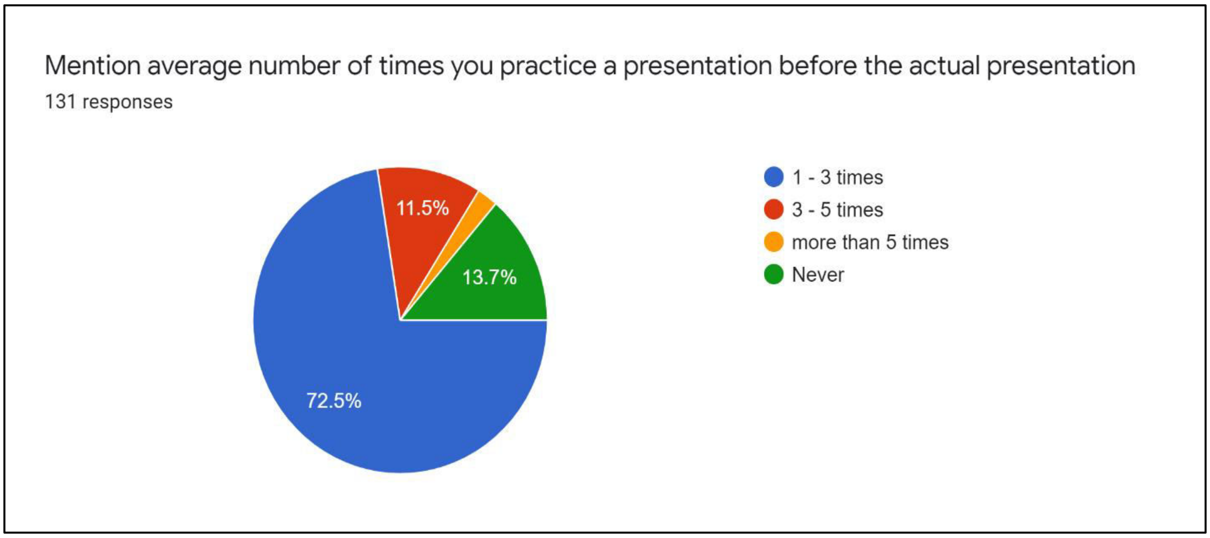


Figure 5 Summary of responses average number for practice a presentation

2 Objectives

2.1 Main Objective

The main objective of the implementation of "Presently" is to work as a personal trainer app to self-train and improve presentation skills. This Mobile Responsive Web Application would help presenters to prepare beforehand for the presentations to deliver a successful speech to the audience. This application will help users to find the incorrect pronunciation and vocabulary mistakes when delivering a presentation. And to suggest user, the match or mismatch between topic tone and emotions used to present the story using audio analysis. Also, to suggest to the user what emotions and enhancements are used to present the story using video analysis. Finally, the system suggests that how to attract an audience effectively by analyzing slides for accuracy of content and aesthetics using computer vision and rules of design-best-practices.

2.2 Specific Objectives

The main objective can be further divided into several specific objectives. The following are some specific objectives for the personal trainer app to self-train and improve presentation skills.

- Implement more accurate and intelligent applications to identify presenters' emotions and prosody levels.
- Suggest user, the match or mismatch between topic tone and emotions used to present the story using audio analysis.

3 Methodology

3.1 Project Overview

The proposed system comprises a mobile responsive web application that helps to improve the presentation skills of a presenter. First, the presenter must upload the video or audio and the presentation slides to the proposed system. According to the presenters' preference, he/she can upload either the video or the audio clip of the presentation to the system. The system facilitates the presenter to check the presentation skills with or without uploading the presentation slides.

Using the uploaded audio clip, firstly the system will check for the pronunciation mistakes done by the presenter. Then the system will check the vocabulary errors and finally will provide feedback, and a rating for the performance. The proposed system will provide the facility to detect the emotions and the body language of the presenter when they upload the video of the presentation. The system will also provide feedback and a rating for the presenter's performance. Other than that, the proposed system can detect emotions, tonality, prosody, and voice qualities that match with the type of presentation or the speech only by using the uploaded audio clip and it will also provide feedback and the rating for the performance.

Moreover, the system will provide the functionality to upload the presentation slides to the system. The system will check the presentations' slides quality by using text and image classification techniques. By analyzing the slides, the system will detect the grammar, attractiveness using color themes, etc. This will also provide feedback and a rating for the quality of the slides. The proposed system is capable to work even without the uploaded presentation slides.

3.2 System Overview Diagram

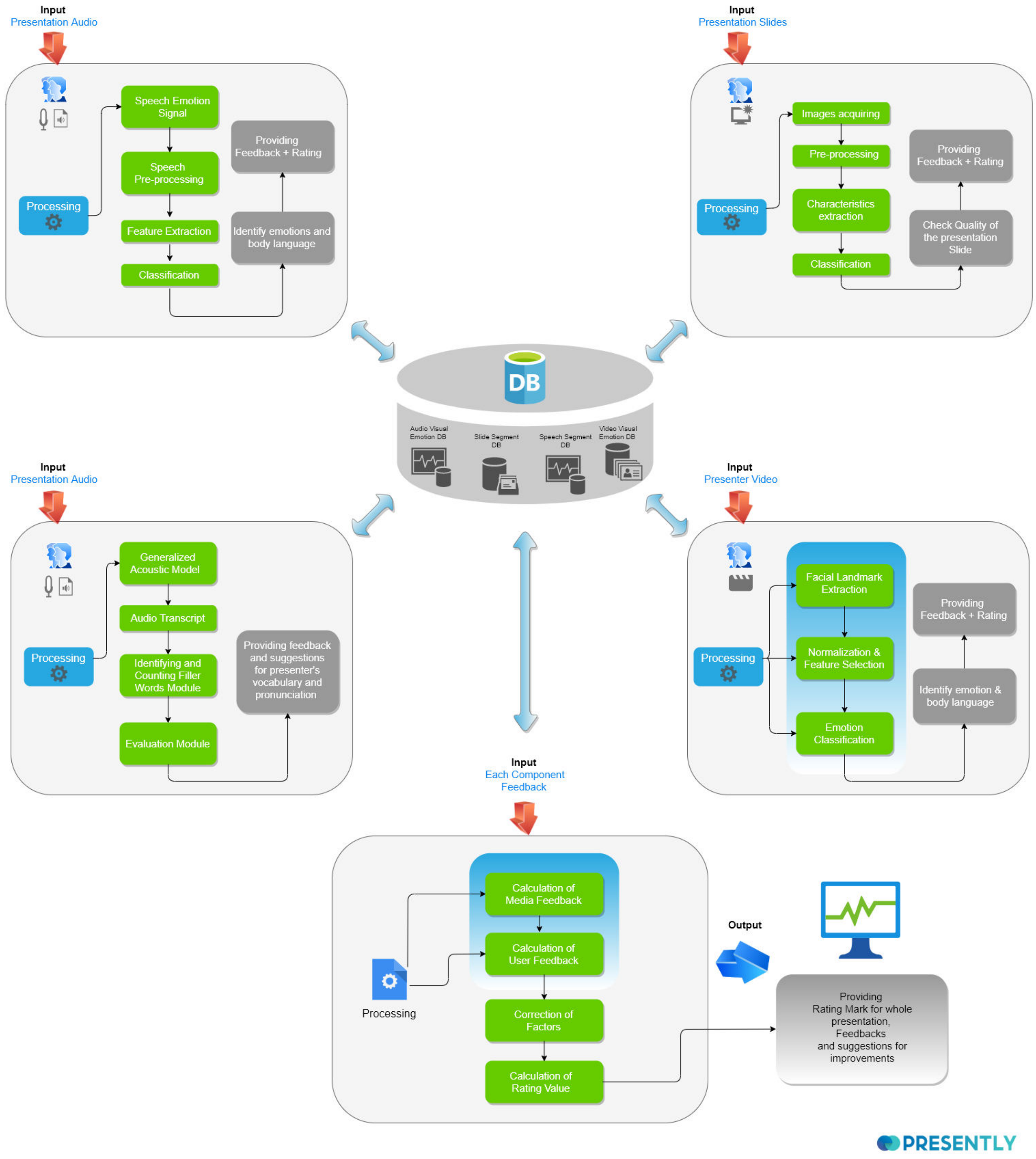

Input
Presentation Audio
Speech Emotion Signal
Speech Pre-processing
Feature Extraction
Classification
Processing
Providing Feedback + Rating
Identify emotions and body language
Input
Presentation Slides
Images acquiring
Pre-processing
Characteristics extraction
Classification
Processing
Providing Feedback + Rating
Check Quality of the presentation Slide
DB
Audio Visual Emotion DB
Slide Segment DB
Speech Segment DB
Video Visual Emotion DB
Input
Presentation Audio
Generalized Acoustic Model
Audio Transcript
Identifying and Counting Filler Words Module
Evaluation Module
Processing
Providing feedback and suggestions for presenter's vocabulary and pronunciation
Input
Presenter Video
Facial Landmark Extraction
Normalization & Feature Selection
Emotion Classification
Processing
Providing Feedback + Rating
Identify emotion & body language
Input
Each Component Feedback
Calculation of Media Feedback
Calculation of User Feedback
Correction of Factors
Calculation of Rating Value
Processing
Output
Providing Rating Mark for whole presentation, Feedbacks and suggestions for improvements
PRESENTLY

Figure 6 Project overview diagram

3.3 System Overview

1) **When delivering a presentation to provide a user with incorrect pronunciation and vocabulary mistakes**

   Communication is the most important component in our daily life. Our society moves on the wheels of communication. So, for better communication, we need correct pronunciation, because pronunciation affects very much the understanding of the meanings of the words. For effective communication, everyone should have a good vocabulary, the knowledge of grammar rules.

   As English is not the native language in Sri Lanka, people tend to do more pronunciation and vocabulary errors when they speak. When doing a presentation, people always had to face lots of difficulties because of these. So, practicing a presentation beforehand will give many chances to mend these pronunciation and vocabulary mistakes.

   The system we designed can suggest the wrong pronunciations and vocabulary mistakes done by the presenters when they upload the audio file, video file, or presentation slide. If a video clip is uploaded, the system we designed, will extract the audio clip from it and analyze it separately for the possible wrong pronunciation and possible vocabulary errors. After the analysis is completed, the system will give a rating and feedback on the performance of the presenter which is useful when doing the actual presentation of the speech by improving their skills. The audio analysis and checking for pronunciation errors will be done using Natural Language Processing (NLP). At end of the day, anyone who needs assistance when practicing a speech or a presentation beforehand will get help from the system for the successful delivery of the speech.

2) **Suggest to the user what emotions and enhancements are used to present the story using video analysis.**

   Many people had to face some difficulties when doing a presentation or a speech in front of an audience. Not practicing enough, avoiding eye contact, failing to engage emotionally are some of the main difficulties that people had to face when doing a presentation. Therefore, practicing beforehand to the presentation helps to stable the emotions while doing the presentation.

The presenter can upload the video to the proposed mobile responsive web application. From the system, only the video clip will be extracted from the uploaded video. Then the system will identify presenters' emotions, body language, and analyze those emotions separately in the video using video analysis and emotion analysis using Computer Vision. At the end of the process, the proposed system will give a rating and feedback using Machine Learning. This rating and feedback can help presenters to get an idea about their presentation skills. By getting feedback on the speech can improve how to stable their emotions while doing the presentation.

3) Suggest the user how to attract an audience effectively by analyzing slides for accuracy of content and aesthetics using computer vision and rules of design-best-practices

A slide is a single screen of a presentation, and every presentation is composed of several slides. Slides keep an audience's attention during a presentation and provide additional supporting information in textual or graphic format. Every presenter knows that simplicity is a virtue that wins business and academic presentations. But when it comes to presentation slides, they make the message difficult to understand by others. To make more sense to the audience the presentation is about, the slides should be more attractive. This will help to maximize the productivity and performance of the presenter during the presentation.

By analyzing the uploaded presentation slides, the system will detect the accuracy and attractiveness using text, color themes, etc. This will be done using computer vision and NLP techniques. This will help to design a more attractive and more accurate set of slides for the final presentation. With the feedback and the rating given by the system by analyzing the uploaded slides, the presenter can improve the quality of the presentation aesthetically.

3.4 Suggest user, the match or mismatch between topic tone and emotions used to present the story using audio analysis - Individual Component

Architectural Design

Emotional prosody is defined as a person's voice tone in a speech that is expressed through changes in pitch, loudness, timbre, speech rate, and pauses. It interacts with verbal content and can be separated from linguistics (e.g., Sarcasm). For the analysis prosody, I will use speech pitch and rhythm to demonstrate how these characteristics contribute to meaning. It focuses on features of speech that often apply at a higher level than the individual phoneme, and frequently to word sequences (in prosodic phrases). Emotions can be conveyed on three levels. [4]:

1) Physiological level (e.g., describing the basic structures involved in the voice-production process' nerve impulses or muscle innervation patterns)
2) Phonatory-articulatory level (e.g., describing the position or movement of the major structures such as the vocal folds)
3) Acoustic level (e.g., indicating the elements of the mouth-produced speech waveform)

Our application proposes to train the software using normal features and then allow the software to detect abnormalities in new data if it deviates too much from the given data by running it through multiple anomaly detectors simultaneously that check for different emotions.

The vocal tract system, excitation source, duration of sound units (syllables, words, and phrases), and intonation are the four components that play a key role in carrying the information of the speech signal. The vocal tract system and source excitation mainly reflect the inherent characteristics of the speech production mechanism. The suprasegmental or prosody features are mostly determined by duration and intonation. When a speaker's utterances of the same sentence in emotional and neutral modes are compared, differences in one or more components of speech are noted.

For my research component, I have decided to use a python-based frontend and backend. And backend contains a modern text-to-text speech framework (TTS) and uses a rule-based method. In additionally I have decided to use common audio analysis tools.

All the users can use the frontend side to upload their presentations. Next, the backend captures the audio and analyzes emotions and prosody. Finally, the results push into the front end. From that user can get feedback on the presentation detected by the system (identified emotions and prosody during presenting).

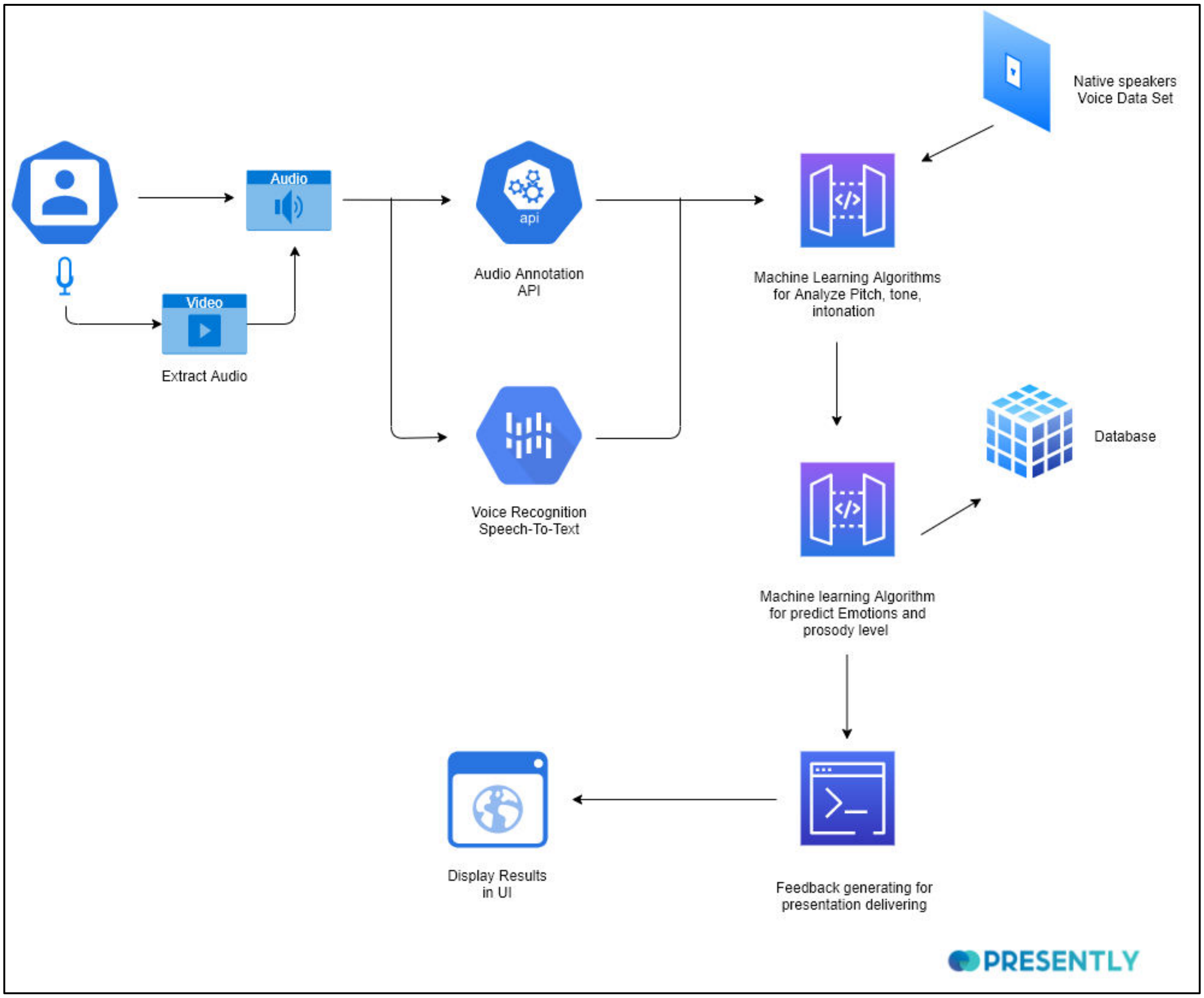

Native speakers
Voice Data Set
Audio
Video
Extract Audio
api
Audio Annotation API
Voice Recognition Speech-To-Text
Machine Learning Algorithms for Analyze Pitch, tone, intonation
Database
Machine learning Algorithm for predict Emotions and prosody level
Display Results in UI
Feedback generating for presentation delivering
PRESENTLY

Figure 7 Architectural Diagram – individual component

Architectural Diagram

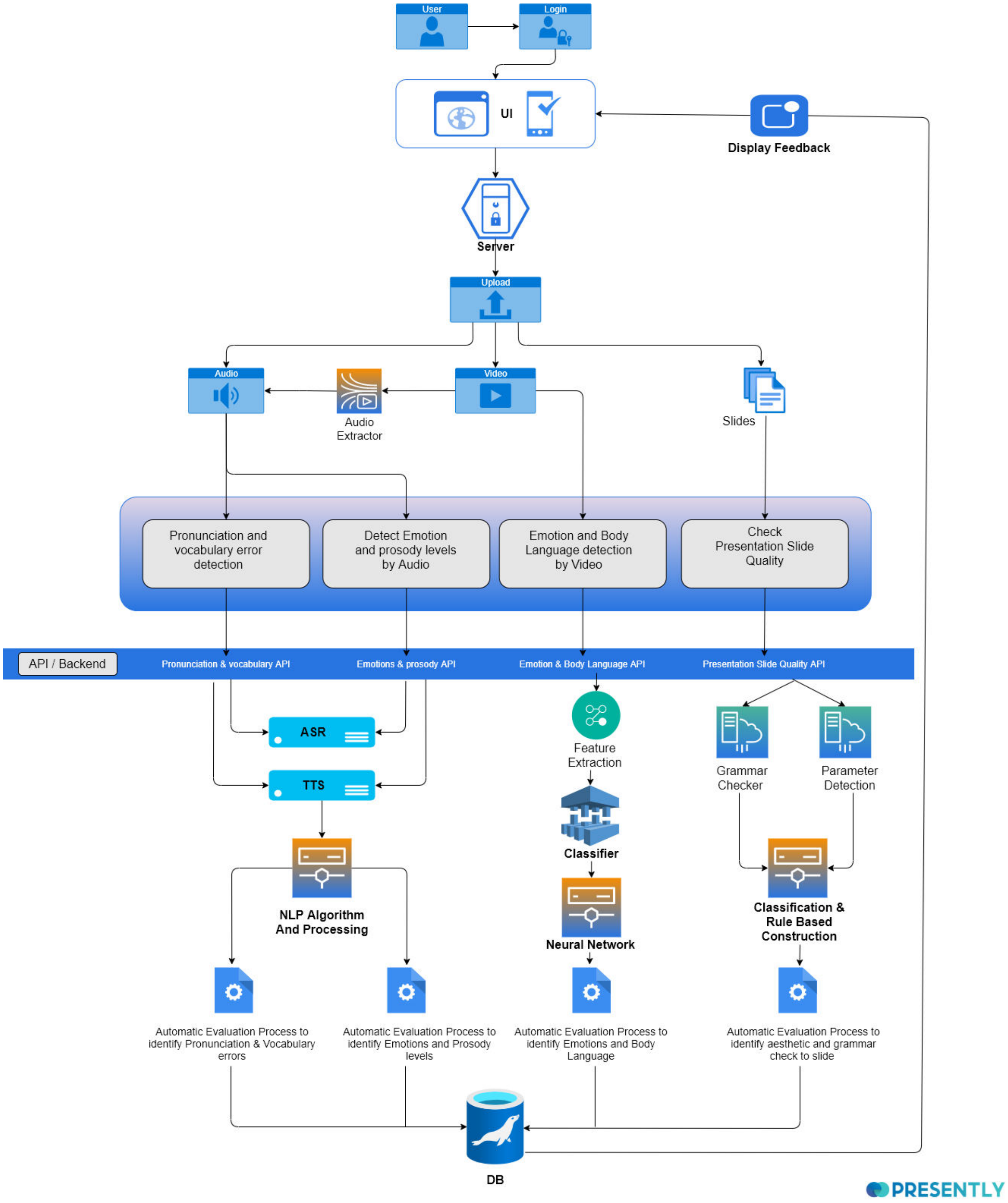

User
Login
UI
Display Feedback
Server
Upload
Audio
Audio Extractor
Video
Slides
Pronunciation and vocabulary error detection
Detect Emotion and prosody levels by Audio
Emotion and Body Language detection by Video
Check Presentation Slide Quality
API / Backend
Pronunciation & vocabulary API
Emotions & prosody API
Emotion & Body Language API
Presentation Slide Quality API
ASR
TTS
Feature Extraction
Grammar Checker
Parameter Detection
Classifier
NLP Algorithm And Processing
Neural Network
Classification & Rule Based Construction
Automatic Evaluation Process to identify Pronunciation & Vocabulary errors
Automatic Evaluation Process to identify Emotions and Prosody levels
Automatic Evaluation Process to identify Emotions and Body Language
Automatic Evaluation Process to identify aesthetic and grammar check to slide
DB
PRESENTLY

Figure 8 System Architectural Diagram

3.4.1 The flow of the system – Individual Component

Speech is a basic medium for communicating not only words but also a wide range of human emotions. As a counter-speech, speech processing applications such as the human-machine interface and speech recognition can benefit from the introduction of a reliable method for the automatic detection of human emotions through speech. Speaker recognition is a method that recognizes a speaker from the characteristics of a voice. Speech recognition technology is widely used in many domains. The functionality of these speech recognition systems is degraded when recognizing emotionally charged speech.

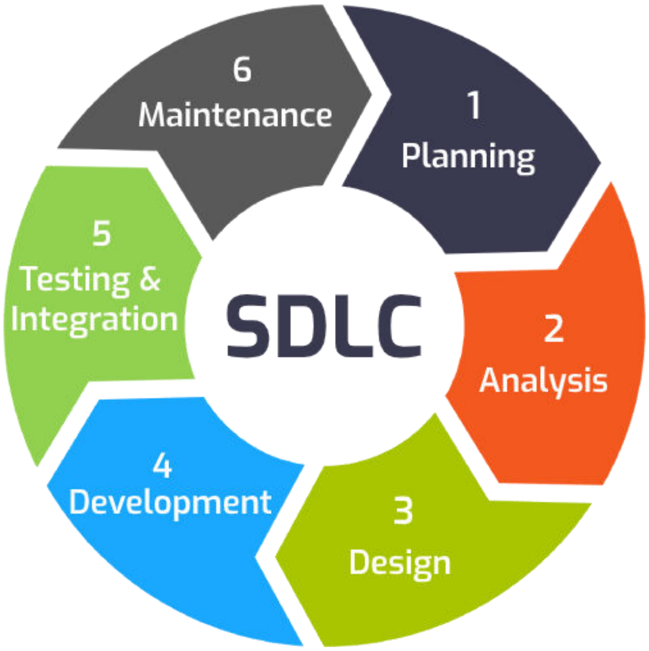


Figure 9 Software development life cycle

Since this is a one-year research project with tight deadlines, the Agile model will be used as an SDLC methodology (Figure 5-Software development life cycle) because the Agile model quickly delivers a working product and is considered a very realistic development approach. Git and Git lab will be used as version control and microservice as the software architecture because the proposed solution is research and doing with new technologies. So, I have to do more testing rather than an ordinary software solution and because of the maintainability. While Django is used to develop the frontend, python will be used in the backend and ML algorithms. Google ASR will be used to capture the voice input, Librosa libraries will be used to processing tasks, and Google Colab, visual studio code will be used as IDEs.

Requirement Gather and analysis

Gathering requirements is an essential stage when presenting a new idea. With the help of gathered requirements, we were able to get an idea about the need for a presentation skills prediction platform, what people think about the presentation skills personal trainer methods, and how it needs to improve when it comes to a web application platform and problems presenters' facing when delivering a presentation. The questionnaire is used in the survey and gets feedback from several university students and industry people to get more information.

Feasibility Study

The feasibility study is an analysis that considers all the relevant factors of a project, including economic, technical, and scheduling factors, to determine the probability of successful completion of the project.

- Economic Feasibility

  The focus is on economic feasibility, project cost, and marketing value. Collected survey data clearly show the market value and demand of this project and it will further discuss the research problem.

- Technical Feasibility

To archive, the goal of the project must need to learn everything from project planning to project deployment. During that journey, group members should need to be a specialist in some technologies. For example, to analyze the data, and predictions, one needs to learn machine learning and deep learning concepts. For data visualization, students need to learn frontend technologies while learning backend technologies to process the data. Knowledge of version control will add more advantages while implementing the project.

- Scheduling Factors

The proposed project should be completed within the stipulated time frame Complete each stage with reliable rewards while maintaining a time frame. Submit the final results with the product on the scheduled due date.

- **Implementation**

The following activities are performed during the implementation phase,

1. Identify user's voice input

By using automatic speech recognition, the system will capture the user's audio input.

2. Analyze the audio input

PRAAT helps to annotate the audio signal file and by using machine learning algorithms such as the gaussian mixture model.

3. Detect Stress level, pitch, tone, and intonation of the speech

To detect stress levels, the audio fill brake in into words and an added time frame into all words by using an ML algorithm.

4. Detect emotions and prosody levels of the speak

Since the system has a text format of the audio file, the system will be able to capture the prosody and by analyzing the data gathered in the above steps, the system will predict the emotions of the speaker.

5. Provide feedbacks according to the mistakes.

To provide feedback, mistakes are done by the user, and to train the user, I hope to develop a user-friendly modern UI by using frontend technologies such as Django. It is among the best python frameworks and is used for the quick development of APIs and web applications.

3.4.2 The flow of the prosody modeling – Individual Component

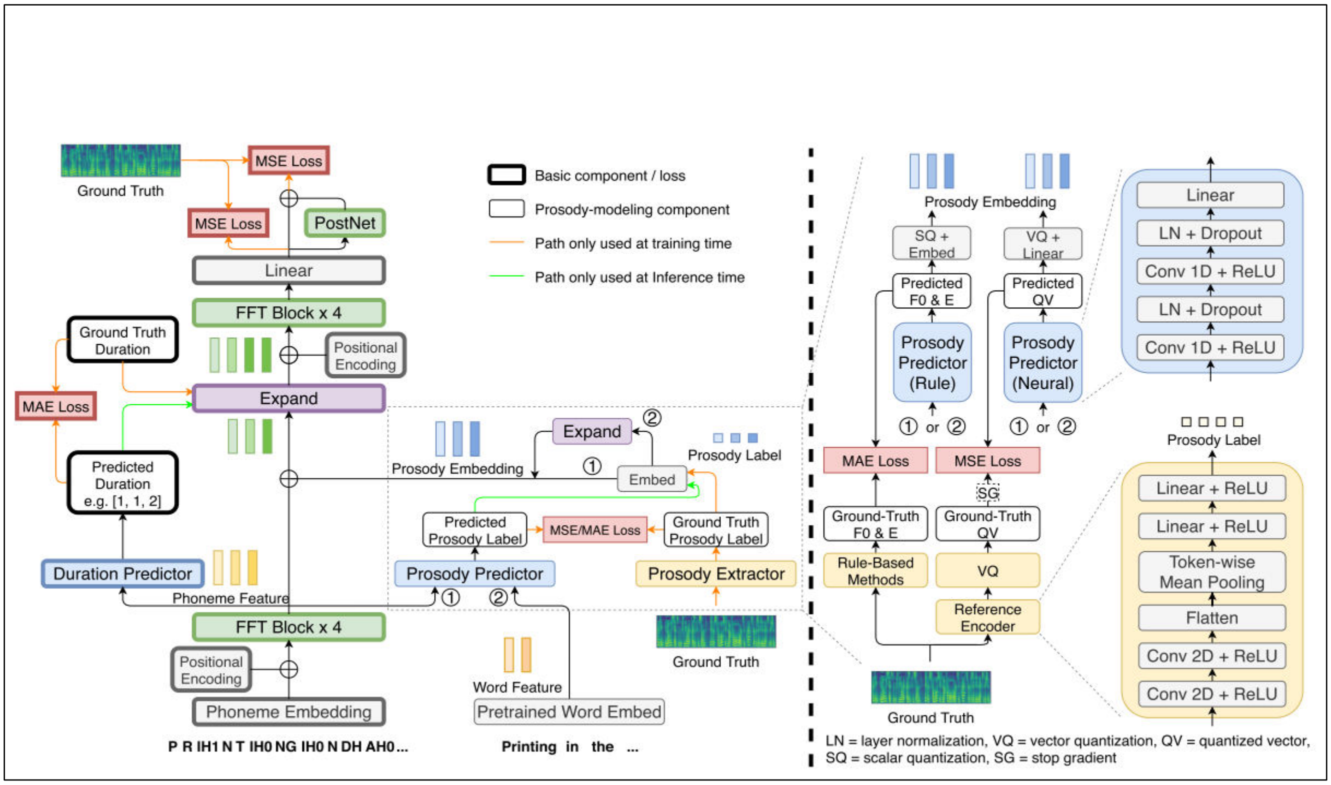


Figure 10 Overview of model architecture and prosody modeling components [7]

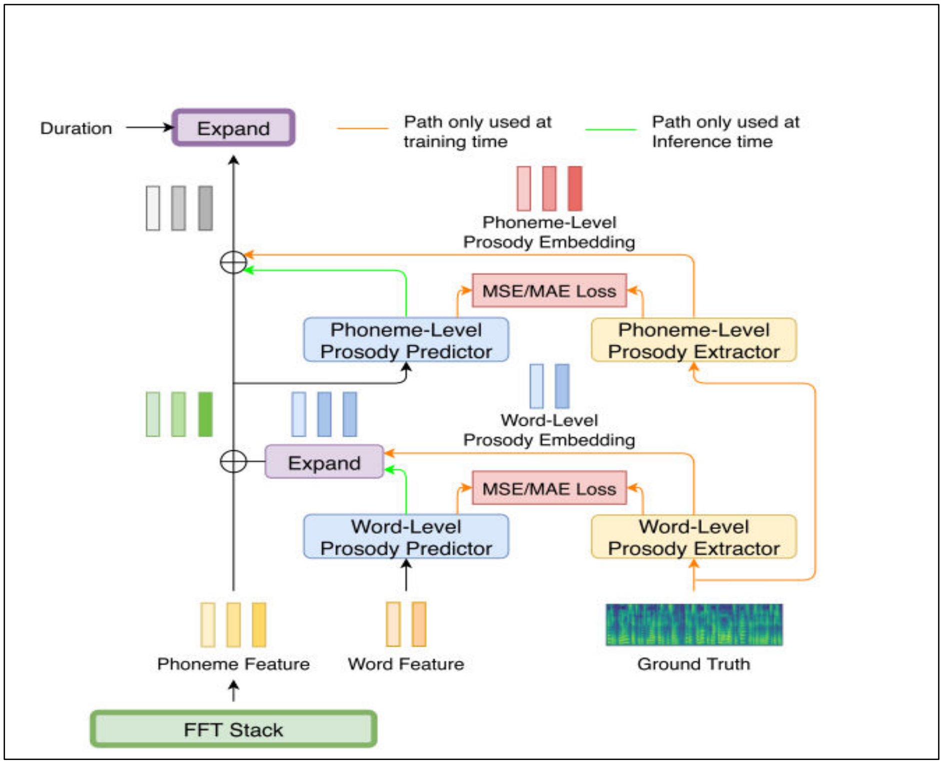

Duration
Expand
Path only used at training time
Path only used at Inference time
Phoneme-Level Prosody Embedding
MSE/MAE Loss
Phoneme-Level Prosody Predictor
Phoneme-Level Prosody Extractor
Word-Level Prosody Embedding
Expand
MSE/MAE Loss
Word-Level Prosody Predictor
Word-Level Prosody Extractor
Phoneme Feature
Word Feature
Ground Truth
FFT Stack

Figure 11 Prosody modeling architecture [7]

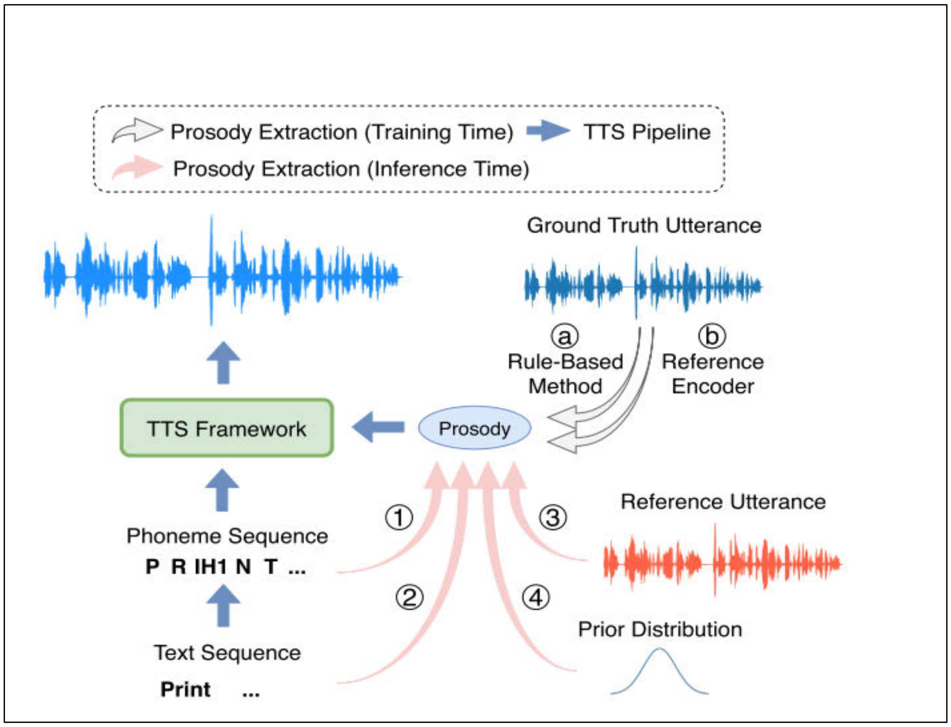


Figure 12 General pipeline for TTS systems with prosody modeling [7]

3.5 Wireframes for proposed System

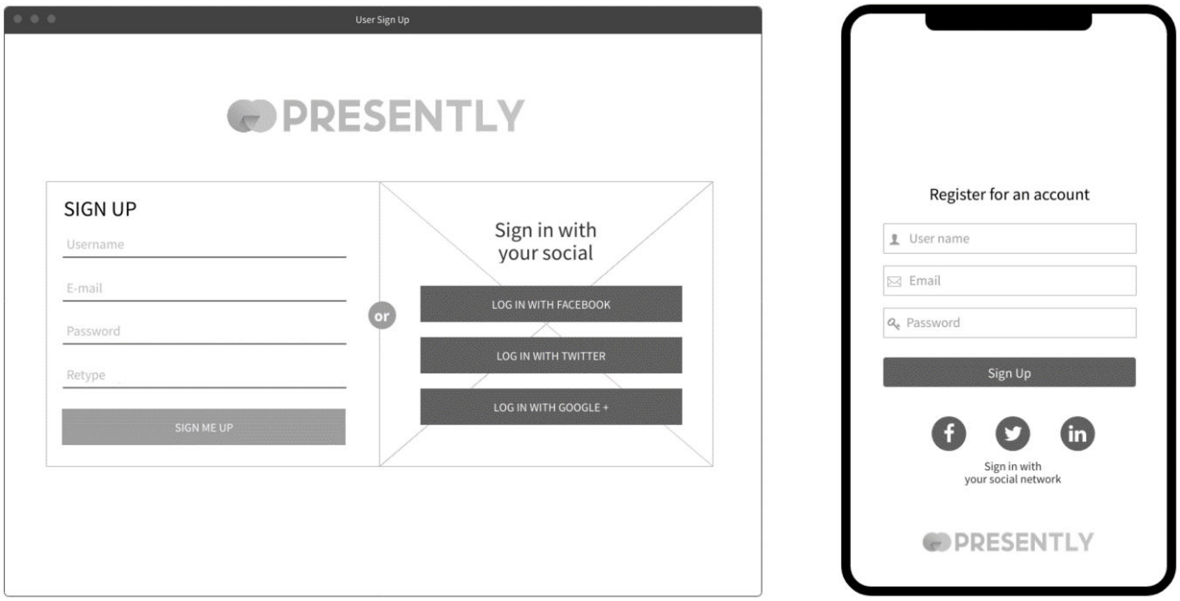


Figure 13 Sign-up for the System - Wireframe

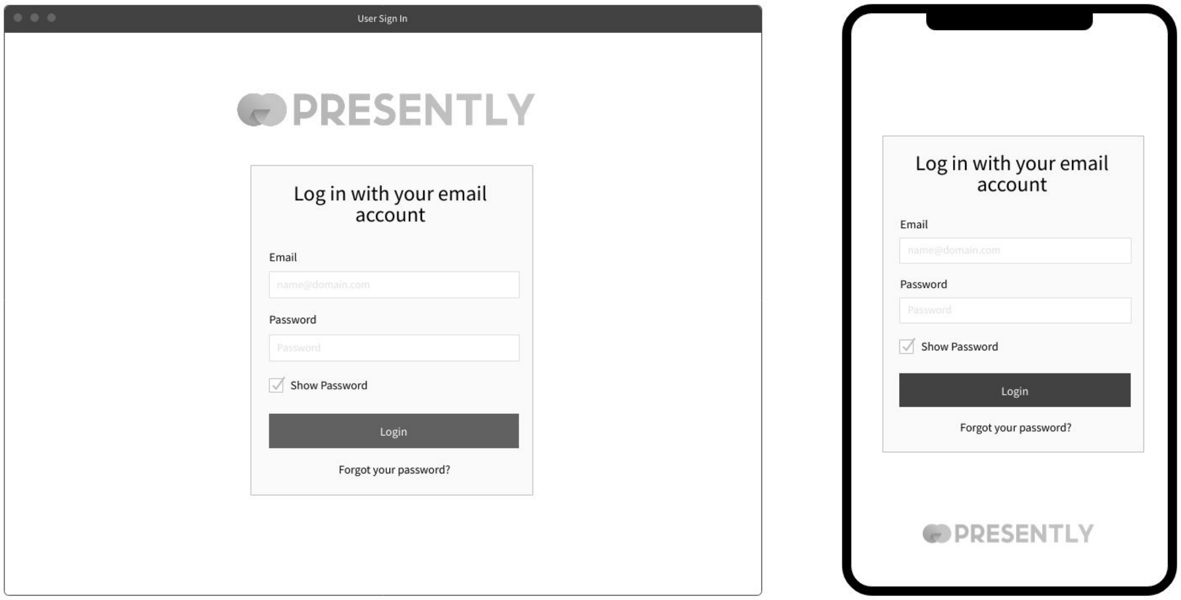
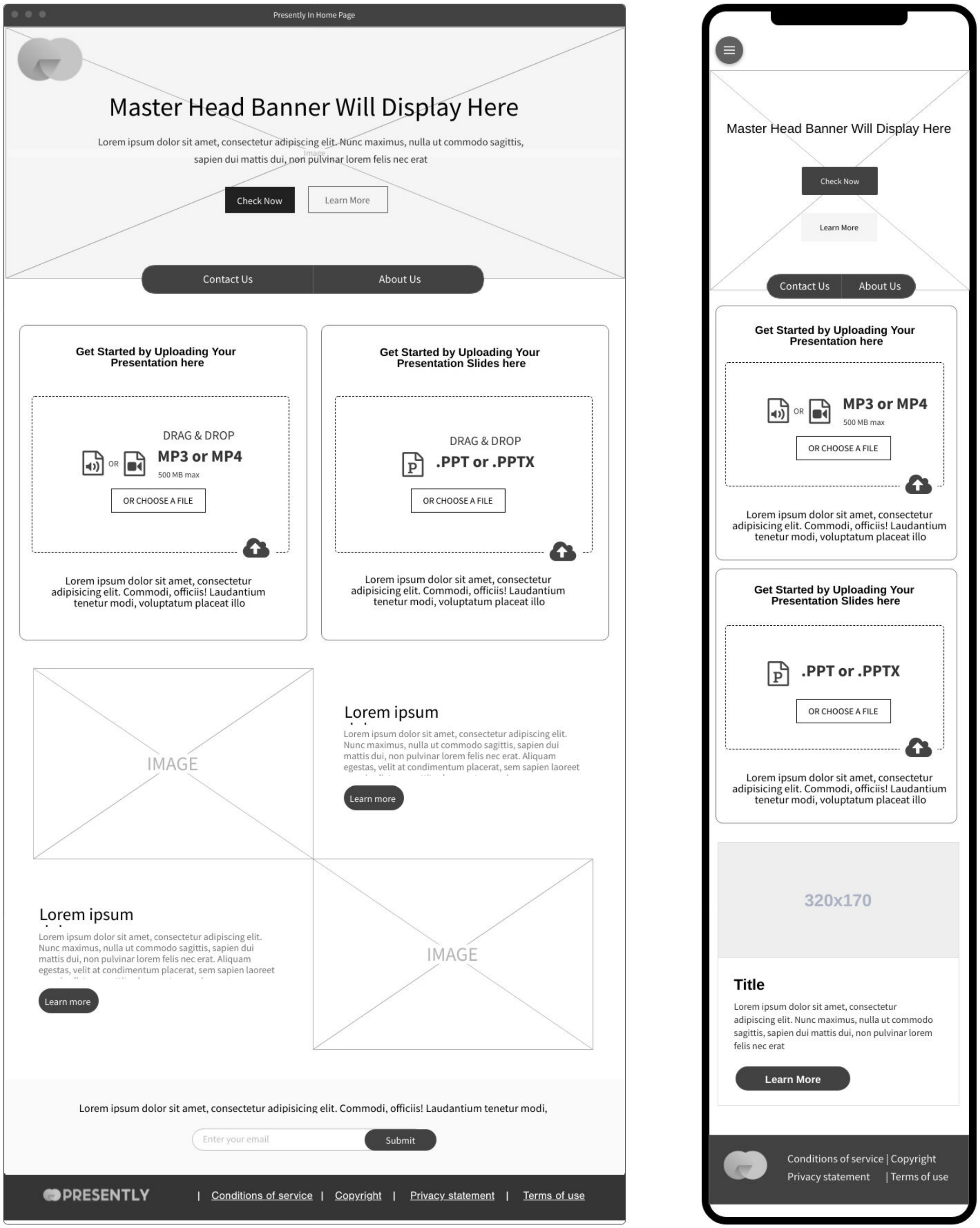


Figure 14 Sign-in for the system - Wireframe



Master Head Banner Will Display Here

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nunc maximus, nulla ut commodo sagittis, sapien dui mattis dui, non pulvinar lorem felis nec erat

Get Started by Uploading Your Presentation here

Lorem ipsum dolor sit amet, consectetur adipisicing elit. Commodi, officiis! Laudantium tenetur modi, voluptatum placeat illo

Get Started by Uploading Your Presentation Slides here

Lorem ipsum dolor sit amet, consectetur adipisicing elit. Commodi, officiis! Laudantium tenetur modi, voluptatum placeat illo

Lorem ipsum

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nunc maximus, nulla ut commodo sagittis, sapien dui mattis dui, non pulvinar lorem felis nec erat. Aliquam egestas, velit at condimentum placerat, sem sapien laoreet

Lorem ipsum

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nunc maximus, nulla ut commodo sagittis, sapien dui mattis dui, non pulvinar lorem felis nec erat. Aliquam egestas, velit at condimentum placerat, sem sapien laoreet

Lorem ipsum dolor sit amet, consectetur adipisicing elit. Commodi, officiis! Laudantium tenetur modi,

Master Head Banner Will Display Here

Get Started by Uploading Your Presentation here

Lorem ipsum dolor sit amet, consectetur adipisicing elit. Commodi, officiis! Laudantium tenetur modi, voluptatum placeat illo

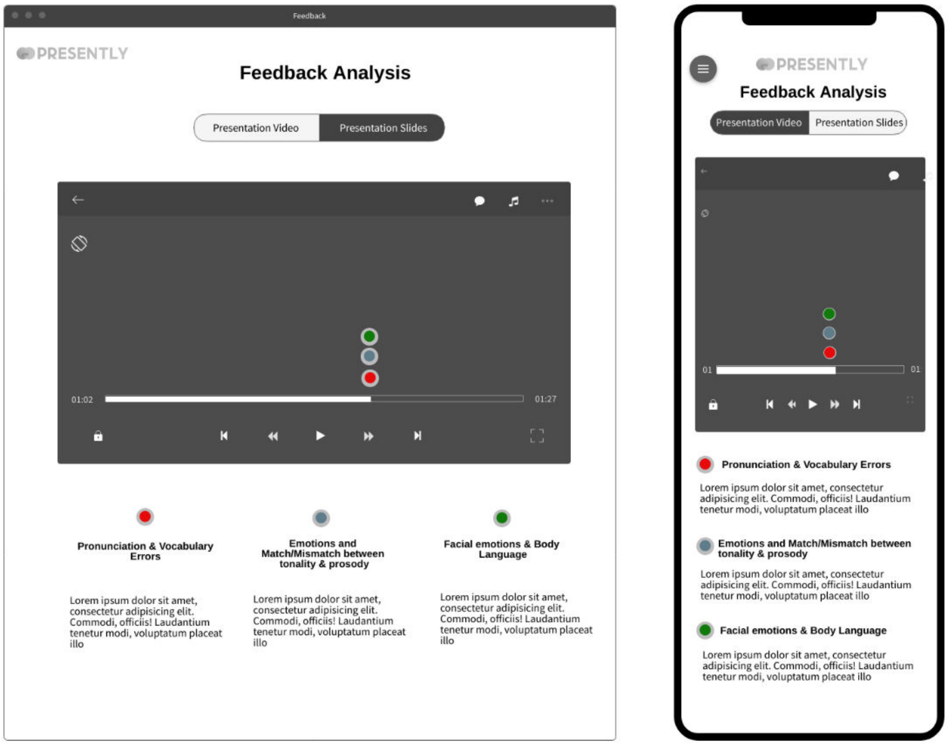
Get Started by Uploading Your Presentation Slides here

Lorem ipsum dolor sit amet, consectetur adipisicing elit. Commodi, officiis! Laudantium tenetur modi, voluptatum placeat illo

Title

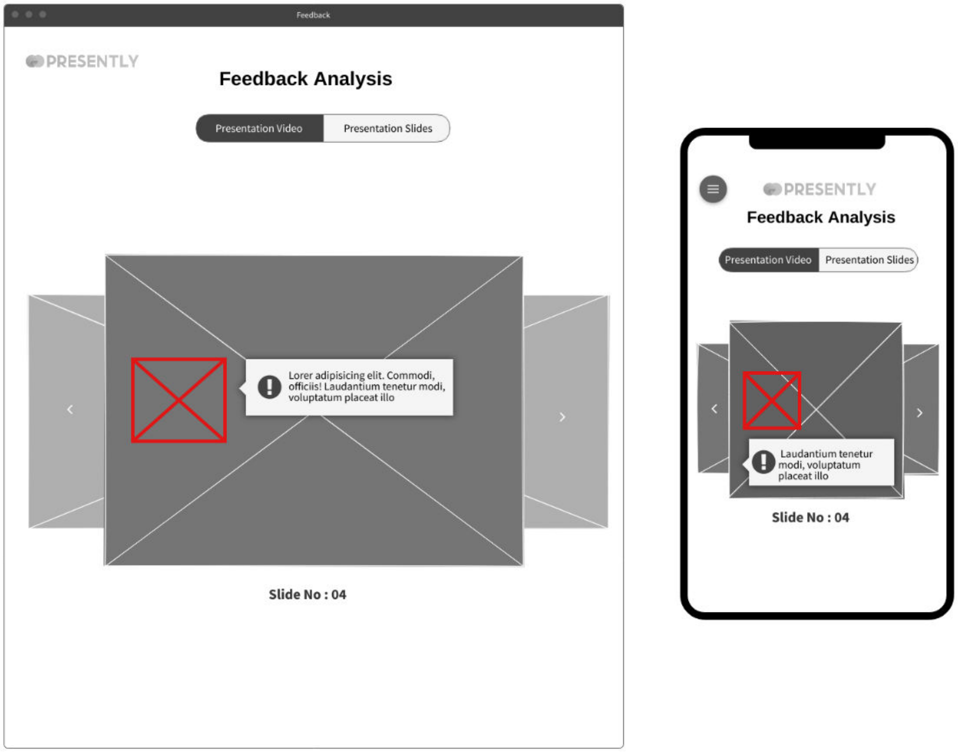
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nunc maximus, nulla ut commodo sagittis, sapien dui mattis dui, non pulvinar lorem felis nec erat

Figure 15 - Main Home page of the system - Wireframe


Feedback
PRESENTLY
Feedback Analysis
Presentation Video
Presentation Slides
01:02
01:27
Pronunciation & Vocabulary Errors
Emotions and Match/Mismatch between tonality & prosody
Facial emotions & Body Language
PRESENTLY
Feedback Analysis
Presentation Video
Presentation Slides
01
01
Pronunciation & Vocabulary Errors
Emotions and Match/Mismatch between tonality & prosody
Facial emotions & Body Language

Lorem ipsum dolor sit amet, consectetur adipisicing elit. Commodi, officiis! Laudantium tenetur modi, voluptatum placeat illo

Figure 16 - Output of the system | Video Analysis – Wireframe



Lorer adipisicing elit. Commodi, officiis! Laudantium tenetur modi, voluptatum placeat illo

Laudantium tenetur modi, voluptatum placeat illo

Figure 17 - Output of the system | Slide Analysis - Wireframe

As shown in Figures 13 and 14, user can create their account in the system and log in to the system. Then User will be redirected to the main home page. As shown in Figure 15, presentation video or audio and presentation slides can be uploaded to the system in the user's favor. These input uploading columns can be easily identified on the home page. From minimum steps, the user will be able to get feedback for their presentation video and slides.

As shown in Figures 16 and 17, all the outputs are displayed. This is designed in a more user-friendly way. From 2 different windows, feedbacks are displayed to the user. For the presentation video, identified pronunciation and vocabulary errors, emotions, and prosody levels, facial expressions are displayed in the 3 different color bullets when the video is playing. So, the user will be able to get feedback from time to time. And for the presentation slides also identified errors and mistakes are display as shown in Figure 17.

3.6 Work Break Down Chart and Gantt Chart

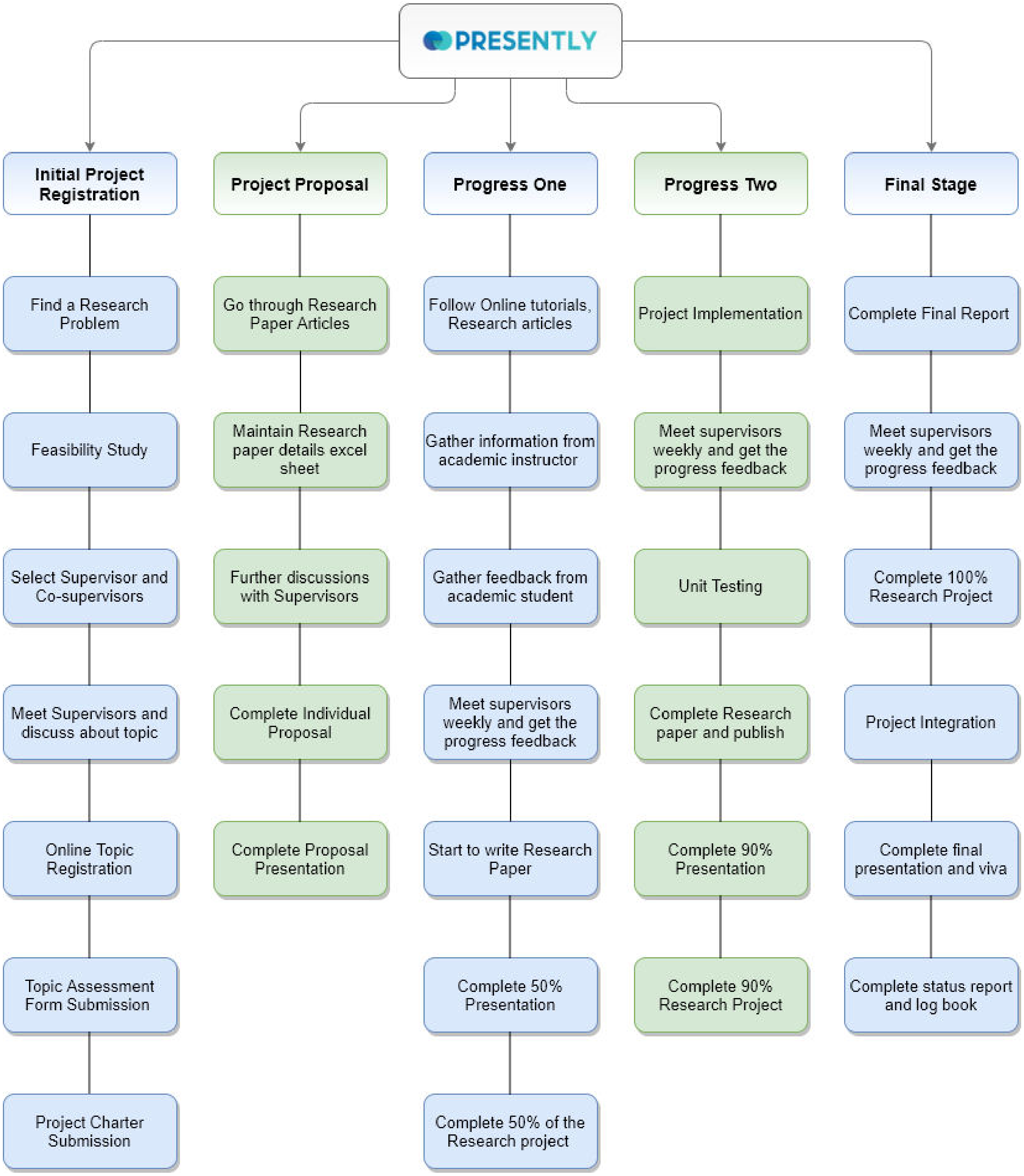

PRESENTLY
Initial Project Registration
Find a Research Problem
Feasibility Study
Select Supervisor and Co-supervisors
Meet Supervisors and discuss about topic
Online Topic Registration
Topic Assessment Form Submission
Project Charter Submission
Project Proposal
Go through Research Paper Articles
Maintain Research paper details excel sheet
Further discussions with Supervisors
Complete Individual Proposal
Complete Proposal Presentation
Progress One
Follow Online tutorials, Research articles
Gather information from academic instructor
Gather feedback from academic student
Meet supervisors weekly and get the progress feedback
Start to write Research Paper
Complete 50% Presentation
Complete 50% of the Research project
Progress Two
Project Implementation
Meet supervisors weekly and get the progress feedback
Unit Testing
Complete Research paper and publish
Complete 90% Presentation
Complete 90% Research Project
Final Stage
Complete Final Report
Meet supervisors weekly and get the progress feedback
Complete 100% Research Project
Project Integration
Complete final presentation and viva
Complete status report and log book

Figure 18 Work breakdown chart

No	PRESENTLY Assessment / Milestone	Start Date	End Date	2021-2022													
				April	May	June	July	August	September	October	November	December	January	February	March	April	May
1	**Project discussion workshop**	23-Apr-21	23-Apr-21	🟩													
2	**Topic evaluation**	15-May-21	30-Jul-21		🟩	🟩	🟩										
2a	Select a topic	15-May-21	20-May-20		🟦												
2b	Select a supervisor	20-May-21	23-May-21		🟦												
2c	Topic Evaluation form submission	23-May-21	25-Jun-21		🟦	🟦											
2d	Project charter submission	20-Jun-21	30-Jul-21			🟦	🟦										
3	**Project proposal report**	15-Jun-21	10-Aug-21			🟩	🟩	🟩									
3a	Create Project Proposal - individual	15-Jun-21	15-Jul-21			🟦	🟦										
3b	Create Project Proposal - group	15-Jul-21	06-Aug-21				🟦	🟦									
3c	Project proposal presentation	01-Aug-21	10-Aug-21					🟦									
4	**Develop the system**	06-Aug-21	20-Feb-22					🟩	🟩	🟩	🟩	🟩	🟩	🟩			
4a	Identifying functions	06-Aug-21	20-Aug-21					🟦									
4b	Database designing	20-Aug-21	12-Sep-21					🟦	🟦								
4c	Implementation	12-Sep-21	30-Dec-21						🟦	🟦	🟦	🟦					
4d	Unit testing	01-Jan-22	30-Jan-22										🟦				
4e	Integration testing	30-Jan-22	20-Feb-22										🟦	🟦			
5	**Progress Presentation - I**	01-Jan-22	06-Jan-22										🟩				
5a	Project Status document	01-Jan-22	06-Jan-22										🟦				
5b	Create presentation document	01-Jan-22	06-Jan-22										🟦				
5c	Progress Presentation – I (50%)	06-Jan-22	06-Jan-22										🟦				
6	**Research Paper**	18-Oct-21	18-Mar-22							🟩	🟩	🟩	🟩	🟩	🟩		
6a	Create the Research Paper	18-Oct-21	18-Mar-22							🟦	🟦	🟦	🟦	🟦	🟦		
7	**Progress Presentation - II**	22-Mar-22	29-Apr-22												🟩	🟩	
7a	Create presentation document	22-Mar-22	29-Apr-22												🟦	🟦	
7b	Progress presentation – II (90%)	29-Apr-22	29-Apr-22													🟦	
8	**Final Report Submission**	14-Apr-22	14-May-22													🟩	🟩
8a	Final Report Submission	14-Apr-22	14-May-22													🟦	🟦
8b	Application assessment	01-May-22	14-May-22														🟦
8c	Project status document	14-May-22	14-May-22														🟦
8d	Student logbook	14-May-22	14-May-22														🟦
9	**Final Presentation & Viva**	14-Apr-22	25-May-22													🟩	🟩
9a	Create final presentation	01-May-22	25-May-22													🟦	🟦
9b	Final report submission	25-May-22	25-May-22														🟦

Figure 19 Project Gantt chart

3.7 Requirement Analysis

Functional Requirements

- Either Student, employee, or any user should create their account in “Presently” and log from their account to the system.
- The user can easily record their presentation as a video or audio-only.
- The system should be able to upload recorded video/audio by itself.
- Obtain a presentation slide to get aesthetic analysis.
- Received final feedback (output) for the uploaded presentation and slides.

Non-Functional Requirements

1. Performance

- Should be able to access all users who need to predict presentation skills.
- Should be able to upload recorded video presentation or audio of speech get feedback for skills.
- Should be able to upload presentation slides (ppt) to get aesthetic analysis feedback.
- Should be able to save all feedbacks in the user account for overall feedback.

2. Security

- Need to create an account and correct credentials for login to the profile.
- Not allowed to watch feedbacks which are given by system for outsiders.
- Not allowed to break rules and regulations given by the system.

3. Reliability

- The system should be available all the time.
- The system should be run on any device.

4. Correctness

- Should get final feedback from the highly accurate level.
- The application should be error-free.

4 Budget and Budget Justification

Component	Amount (Rs.)
Internet	5000.00
Stationery	2000.00
Documentation and printing cost	3000.00
Server cost	4000.00
Educational survey cost (online payments)	1000.00
Electricity	1000.00
Transport	2500.00
Total	17500.00

Table 2 Budget and budget justification

5 References

[1] V. V. V. Raju, H. K. Vydana, S. V. Gangashetty and A. K. Vuppala, "Importance of non-uniform prosody modification for speech recognition in emotion conditions," 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017, pp. 573-576, doi: 10.1109/APSIPA.2017.8282109.

[2] B. Sisman, G. Lee, H. Li and K. C. Tan, "On the analysis and evaluation of prosody conversion techniques," 2017 International Conference on Asian Language Processing (IALP), 2017, pp. 44-47, doi: 10.1109/IALP.2017.8300542.

[3] V. V. Raju V., K. Gurugubelli, K. N. R. K. R. Alluri and A. Kumar Vuppala, "Differenced Prosody Features from Normal and Stressed Regions for Emotion Recognition," 2018 5th International Conference on Signal Processing and Integrated Networks (SPIN), 2018, pp. 821-825, doi: 10.1109/SPIN.2018.8474265.

[4] N. Ang, D. Bein, D. Dao, L. Sanchez, J. Tran and N. Vurdien, "Emotional prosody analysis on human voices," 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), 2018, pp. 737-741, doi: 10.1109/CCWC.2018.8301691.

[5] V. V. V. Raju, P. Gangamohan, S. V. Gangashetty and A. k. Vuppala, "Application of prosody modification for Speech Recognition in different Emotion conditions," 2016 IEEE Region 10 Conference (TENCON), 2016, pp. 951-954, doi: 10.1109/TENCON.2016.7848145.

[6] R. Fu, J. Tao, Z. Wen, J. Yi, T. Wang and C. Qiang, "Bi-Level Style and Prosody Decoupling Modeling for Personalized End-to-End Speech Synthesis," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 6568-6572, doi: 10.1109/ICASSP39728.2021.9414422.

[7] C. -M. Chien and H. -y. Lee, "Hierarchical Prosody Modeling for Non-Autoregressive Speech Synthesis," 2021 IEEE Spoken Language Technology Workshop (SLT), 2021, pp. 446-453, doi: 10.1109/SLT48900.2021.9383629.

[8] H. Prafianto, T. Nose and A. Ito, "A precise evaluation method of prosodic quality of non-native speakers using average voice and prosody substitution," 2016 International Conference on Audio, Language and Image Processing (ICALIP), 2016, pp. 208-212, doi: 10.1109/ICALIP.2016.7846620.

[9] A. Kamat, A. Krishnamurthy, D. N. Krishna and V. Ramasubramanian, "Prosodic differential for narrow-focus word-stress in speech synthesis," 2017 Twenty-third National Conference on Communications (NCC), 2017, pp. 1-6, doi: 10.1109/NCC.2017.8077132.

[10] L. Matsane, A. Jadhav and R. Ajoodha, "The use of Automatic Speech Recognition in Education for Identifying Attitudes of the Speakers," 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2020, pp. 1-7, doi: 10.1109/CSDE50874.2020.9411528.

[11] Z. HAN and J. WANG, "Speech Emotion Recognition Based on Deep Learning and Kernel Nonlinear PSVM," 2019 Chinese Control and Decision Conference (CCDC), 2019, pp. 1426-1430, doi: 10.1109/CCDC.2019.8832414.

[12] H. Koo, S. Jeong, S. Yoon and W. Kim, "Development of Speech Emotion Recognition Algorithm using MFCC and Prosody," 2020 International Conference on Electronics, Information, and Communication (ICEIC), 2020, pp. 1-4, doi: 10.1109/ICEIC49074.2020.9051281.

[13] M. G. K. Noor Fathima et al., "Phonetically conditioned prosody transplantation for TTS: Unit granularity, context and prosody styles," 2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), 2016, pp. 99-104, doi: 10.1109/ICSDA.2016.7918992.

[14] S. Suganya and E. Y. A. Charles, "Speech Emotion Recognition Using Deep Learning on audio recordings," 2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer), 2019, pp. 1-6, doi: 10.1109/ICTer48817.2019.9023737.

[15] P. Robitaille, S. Trempe, P. Gournay and R. Lefebvre, "On the influence of quantization on the identifiability of emotions from voice coding parameters," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5950-5954, doi: 10.1109/ICASSP.2016.7472819.

[16] （Audio-Based Granularity-Adapted Emotion Classificatio，2018）