ARTICULATE: SPEECH COMPANION

Submitted by

Joshwin Isac

2348523

Sai Darshan

2348548

Suhas S

2348563

MSc AIML

Department of Computer Science

for

CONTINUOUS INTERNAL ASSESSMENT

Under the guidance of

Dr. Jobin Francis
Assistant Professor

Department of Computer Science
CHRIST University Central Campus
Bangalore, Karnataka, India – 560029

December 2024

Articulation

December 11, 2024

Abstract

Articulation is an AI speech companion designed to assist users in improving their communication skills, emotional well-being, and daily productivity. This project aims to integrate speech recognition, emotion detection, and real-time feedback systems to create a multi-functional tool to reach individuals, aiding them in enhancing their speech clarity and specialized communication types and practicing general communication practices to improve themselves.

Articulate focuses on improving speech clarity, fluency, and confidence through personalized guidance and corrective feedback, further empowering users to communicate with more ease.

We also aim to integrate a voice-driven task manager, allowing users to organize their daily tasks using voice commands using NLP systems and allowing voice instructions to be converted to proper, clear actions. This functionality will be further explored to allow for good integration with other productivity apps also.

As for accessibility, we aim to provide individuals with hearing impairments and others who communicate with non-standard languages like American Sign Language (ASL), enabling the system to translate speech into sign language through on-screen avatars and offering real-time speech-to-text conversion for smoother communication in diverse settings. An additional function of analyzing pitch and tone of speech amplification is to provide users feedback on their speech style, helping them to increase their speaking abilities.

With AI and speech tech, Articulate strives to address real-world communication challenges, helping people improve their speaking abilities, manage daily tasks, and bridge the communication gap for individuals with disabilities by providing a user-centric solution approach for communicative needs.

1 Introduction

Effective communication is the foundation to personal and professional success but many people find themselves hard to express themselves fluently and confidently. *Articulate* is the innovative AI-powered speech companion with the aim of improving communication abilities, emotional well-being, and daily productivity of the users. It does all this by using state-of-the-art speech recognition and emotion detection along with giving real-time feedbacks through which it offers its personal guidance for the enhancement of speech clarity, fluency, and confidence.

Another project is its voice-driven task manager, in which one can manage day-to-day activities with most efficient usability using natural language processing technology. This particular feature of integrating productivity applications is truly seamless.

Articulate will always strive to empower people who suffer from hearing disabilities, those who use non-standard languages like Indian Sign Language. The system translates spoken language into ISL with avatars on the computer screens and gives real-time speech-to-text conversion. In addition, it will also analyze the pitch and tone of the speech and then provide constructive feedback to better speaking styles and self-expression among users.

Articulate aims at filling these gaps with regard to communication by providing a model that uses artificial intelligence and speech technology toward solution real-world issues with a focus on creating better interaction, accessibility, and productivity.

2 Related Works

Over the past decade, Speech Emotion Recognition has been an integral aspect of Human-Computer Interaction and advanced speech processing systems. SER systems are programmed to detect and classify speech patterns into their respective emotions after filtering audio signals for distinctive characteristics. Yet, there are a great number of inconsistencies in the way humans as opposed to machines perceive and draw from these emotional nuances, bringing us toward insights into speech processing and applied psychology and HCI and bridging the gap through those ends.This book combines the latest literature on methodologies for designing SER systems to present an overview of the field in a holistic manner. It identifies areas with gaps in existing research and, therefore, will form the basis for further improvements by researchers, institutions, and regulatory bodies to develop even more effective and sophisticated SER systems. [10]

Over the past decade, Speech Emotion Recognition has been an integral aspect of Human-Computer Interaction and advanced speech processing systems. SER systems are programmed to detect and classify speech patterns into their respective emotions after filtering audio signals for distinctive characteristics. Yet, there are a great number of inconsistencies in the way humans as opposed to machines perceive and draw from these emotional nuances, bringing us toward insights into speech processing and applied psychology and HCI and bridging the gap through those ends. This book combines the latest literature on methodologies for designing SER systems to present an overview of the field in a holistic manner. It identifies areas with gaps in existing research and, therefore, will form the basis for fur-

ther improvements by researchers, institutions, and regulatory bodies to develop even more effective and sophisticated SER systems.[2]
[6]

A glove-based device, which is created to make easy communication of a mute person with other peoples that, by mapping the gestures in a glove with the American Sign Language (ASL) into speech. It'll detect and convert those signs into text using flex sensors with further communication in speech. An Android application will have sent sensor data through a cloud server using a GSM module inside the system. The system used a back-propagation neural network, which is a learning device meant to continuously learn and continually monitor user behavior over time and improve the reliability. It provides an innovative and personalized solution to solving communication gaps [3].

This paper discusses the accessibility challenge of current ASR systems for PWS. It introduces a new method called Detect and Pass, using a context-aware classifier trained on lesser amounts of data to detect the presence of stuttering in acoustic frames. Such detected information is integrated into ASRs during their inference phase and was proven to improve performance in stuttered speech tasks. Experimental results also demonstrate that impressive reductions in Word Error Rate (WER), ranging from 12.18% to 71.67%, were achieved using state-of-the-art ASR systems. The approach here optimizes stutter detection and acoustic feature processing to make the technology in ASR more accessible and effective for PWS. [4]

Millions of people across the globe have speech disorders, for example, stuttering, that inhibits them from effective communication, especially nowadays with the new modern Automatic Speech Recognition systems. This paper presents a novel approach called Detect and Pass that involves context-aware classifiers trained on little data to identify stuttered segments of speech. It reduces WER by 71.67% across ASR platforms while passing this information into ASR models at time of inference, hence enhancing the overall robustness of the system. It optimizes the thresholds of acoustic feature functions and holds potential to make ASR more inclusive for PWS. The proposed Detect and Pass algorithm addresses one of the biggest problems to accessibility: filling the gap between ASR technologies and people who stutter. This paper bridges well the limitation placed by conventional ASR systems in the understanding of disfluency in speech with the emphasis on detecting disfluency acoustic frames and smoothing their inclusion into ASR. A highly impressive WER improvement certainly establishes the viability of the proposed system to be adaptive to variations in the architecture of different ASR systems. Despite the method's success in the laboratory, further research is required on its practical use and scaleability. The dataset can be expanded and tests should be conducted in different linguistic and acoustic contexts to support generalization. In a nutshell, this research provides a strong basis for enhancing speech accessibility and offers an eloquent argument for integrating this into the design of accessible technologies.[5]

The identification of emotions in speech is very important for the analysis of psychological disorders, behavioral decision-making, and human-machine interaction applications. However, a large percentage of the current techniques for speech emotion recognition rely heavily on data-driven approaches, and the lack of emotional speech datasets restricts progress in research as well as in the development of emotion recognition technologies. analysis and identification. This paper designs a new English speech corpus for specifcally solving this problem Emotion recognition and analysis. There are 5503 voices of more than 60 English speakers in different emotional states. Also, to enhance emotion analysis and recognition, fast Fourier transform (FFT), short-time Fourier transform (STFT), For feature extraction, mel-frequency cepstral coefcients (MFCCs) and continuous wavelet transform (CWT) are used from speech data. Using these algorithms, the spectrogram images of speeches are obtained, which yields four datasets that contain different speech feature images. Furthermore, to evaluate the dataset, 16 classifcation models and 19 detection algorithms are Nearly all the classification and detection models achieve very high performance as shown by the experimental results. recognition acuity on this data, confrming its efectiveness and utility. Te data proves valuable in advancing Research and development in the field of affective recognition. Key-words: data-driven; emotion analysis; speech dataset; speech emotion recognition.[6]

This work addresses the challenge of noisy real-world environments for SER systems. SE modules can improve speech quality but may alter resilient acoustic features critical for SER. The proposed approach selectively enhances weak features, degrades emotion recognition, and leaves resilient features unchanged. Weak features are identified by ranking acoustic features based on performance and robustness using models trained with clean speech. Direct improvement of those enhances the performance of SER rather than the entire signal. Significant performance improvements can be obtained, particularly at arousal at 17.7%, dominance at 21.2%, and valence at 3.3% on the MSP-Podcast corpus, under 10dB signal-to-noise ratio conditions when tested.[7]

People who stutter have problems using consumer speech recognition systems. These often stop, misunderstand, or speak for them inappropriately to express verbal thoughts and intentions. A user survey highlighted these problems; however, overall, there is a strong feeling to use these technologies. Deeper analysis has shown that disfluencies significantly degrade system performance. However, this technology has successfully reduced cutting-off errors by 79.1% and improved word-error rates from 25.4% to 9.9%, demonstrating the possibility to make speech recognition more accessible for people who stutter.[8]

Addressing the challenges with which speech-disordered people-including stutterers face when communicating with consumer-based speech recognition systems. In current days, such systems, which rely largely on data from fluent speakers, are facing major challenges in the effective handling of dysfluencies, especially sound repetitions, prolongations, and blocks. Therefore, the moderate-to-severely speech-disordered individuals perform suboptimally. The study quantifies performance degradation in consumer speech recognition systems and indicates that subjects with fluency disorders are 13.64% worse than fluent speakers on Word Error Rate (isWER). By tuning the decoding parameters of an existing hybrid system, the study achieves a relative improvement of 24% on (isWER) for users with fluency disorders. Similarly, these changes have led to increased domain and intent recognition hence forming a promising approach to enhance access for the stuttering population.[9]

Microsoft Kinect develops a system that translates sign language into speech, enabling the disabled people to communicate with others more effectively. Sign language is a form of communication that relies on visual gestures. It may not be easy for non-sign language speakers to understand the signs and phrases being used. The proposed system makes use of Kinect's motion capture to recognize human gestures such as hand movements and translate them into speech. This system is meant to make communication easier by offering a natural user interface and higher efficiency for users, making it easier for people with speech disabilities to communicate with a wider population. The system is able to trace the human skeleton and correlate gestures with predefined actions to enable gestures such as waving a hand to produce a spoken word, such as "Hello." The use of Kinect improves gesture recognition accuracy and flexibility, closing the communication gap for people suffering from speech disabilities.[10]

3 Technologies and Tools Used

- **Web Framework**: Flask (Used to create a web-based interface for audio upload and result display)
- **Python Libraries**:
  - Librosa (For audio feature extraction)
  - SoundFile (For reading and processing audio files)
  - NumPy (For numerical computations)
  - SpeechRecognition (For speech-to-text conversion using Google Web Speech API)
  - TextBlob (For sentiment analysis of text)
  - PyDub (For audio format conversion)
- **Machine Learning**:
  - MLP Classifier (A trained Multi-Layer Perceptron model for emotion classification with 83% accuracy)
  - PCA (Principal Component Analysis for dimensionality reduction of extracted features)
- **Utilities**:
  - UUID (For generating unique filenames)
  - FFmpeg (For handling audio format conversion)

Table 1: Classification Report of Multilayer Perceptron Model

Emotion	**Precision**	**Recall**	**F1-Score**	**Support**
Angry	0.94	0.84	0.89	199
Calm	0.72	0.72	0.72	86
Disgust	0.87	0.82	0.84	142
Fearful	0.86	0.83	0.85	176
Happy	0.77	0.84	0.80	186
Neutral	0.72	0.96	0.82	165
Sad	0.89	0.80	0.85	199
Surprised	0.89	0.79	0.84	160
Accuracy		0.83		1313
Macro Avg	0.83	0.83	0.83	1313
Weighted Avg	0.84	0.83	0.83	1313

4 Project Objectives

- Integrate speech recognition, emotion detection, and real-time feedback systems to assist users in improving their communication skills and emotional well-being. [10]

- Enhance speech clarity, fluency, and confidence through personalized guidance and corrective feedback.

- Develop a voice-driven task manager to help users organize daily tasks using voice commands and NLP systems, with integration to other productivity apps.

- Provide accessibility features for individuals with hearing impairments and those using non-standard languages like American Sign Language (ASL), including speech-to-sign language translation and real-time speech-to-text conversion.

- Analyze pitch and tone of speech amplification to provide feedback on speech style and help users improve their speaking abilities.

- Address real-world communication challenges using AI and speech technologies to improve speaking abilities and bridge communication gaps for individuals with disabilities.

4.1 Proposed Methodology

- **Emotion Prediction:**
  - **Feature Extraction:** Features such as MFCC (Mel Frequency Cepstral Coefficients), chroma features, and Mel spectrogram were extracted using Librosa.
  - **Dimensionality Reduction:** Principal Component Analysis (PCA) was applied to reduce the feature dimensions before feeding the data into the classifier.
  - **Classification:** A pre-trained Multi-Layer Perceptron (MLP) model was used to predict the emotion, achieving an accuracy of 83%.
- **Sentiment Analysis:**
  - **Speech-to-Text Conversion:** Audio was converted to text using Google Web Speech API.
  - **Text Analysis:** TextBlob was employed to analyze the polarity of the text, classifying the sentiment as Positive, Negative, or Neutral.
- **Audio Handling and Conversion:**
  - The application handles audio files in various formats. If the input audio is not in WAV format, it is converted using PyDub and FFmpeg.

4.2 Expected Outcomes

- **Accurate Gesture-to-Speech Conversion:** The system will accurately translate predefined gestures into corresponding speech outputs, enabling people with speech disabilities to communicate effectively with a wider audience.
- **Real-Time Gesture Recognition:** The system will integrate to allow real-time tracking of gestures, ensuring seamless and natural communication without noticeable delays.
- **Improved Accessibility:** The system will offer an accessible communication method for non-verbal persons or those who rely on sign language, reducing the need for interpreters or written language.
- **Natural User Interface (NUI):** The system will provide a user-friendly interface by using a gesture-based design, simplifying the complexity of traditional communication aids.
- **Facilitating Public Communication:** Users will be able to interact with groups conveniently, as the system will convert gestures into clear and loud speech, making it suitable for public settings.

- **Adjustable Gesture Recognition:** The system will allow customization of gestures, enabling users to define personalized actions for specific speech outputs.
- **Sign Language Detection:** The project will demonstrate the potential of combining motion sensing and gesture recognition for human-machine interaction, showcasing its utility beyond accessibility purposes.

5 Conclusions

The Articulate initiative is an all-rounded way of improving communication skills, emotional well-being, and productivity in cases of speech difficulties. Speech recognition, emotion detection, sentiment analysis, and real-time feedback mechanisms incorporated into the initiative will enable users to upgrade their speech clarity, fluency, and self-confidence. Additionally, a voice-operated task management system can be integrated to help organize tasks for the day through voice commands, thereby increasing productivity overall.

Thus, accessibility features of automatic speech-to-sign language translation and real-time speech-to-text conversion provide ease to deaf persons or even those who use non-standard languages like American Sign Language, etc, so that it can be engaged for interaction purposes. Further, it comprises pitch and tone analysis. This provides improvement in the speech style and enhances the feedback provided for continuous improvements. Articulate tries to break the barriers in actual life situations for people with disabilities, aided by sophisticated artificial intelligence and speech technologies. This, therefore, now ends the communication gap of people with disabilities as the product is both accessible and user needs-centric. In constant development, it will evolve more toward meeting the broader spectrum of user needs, increasing effectiveness and impact.

Table 2: Summary of Key Findings from the Model

Metric	**Value**
Overall Accuracy	83%
Highest Precision (Class 0 - Angry)	94%
Highest Recall (Class 5 - Neutral)	96%
Lowest F1-Score (Class 1 - Calm)	72%
Overall Weighted F1-Score	83%

6 Features

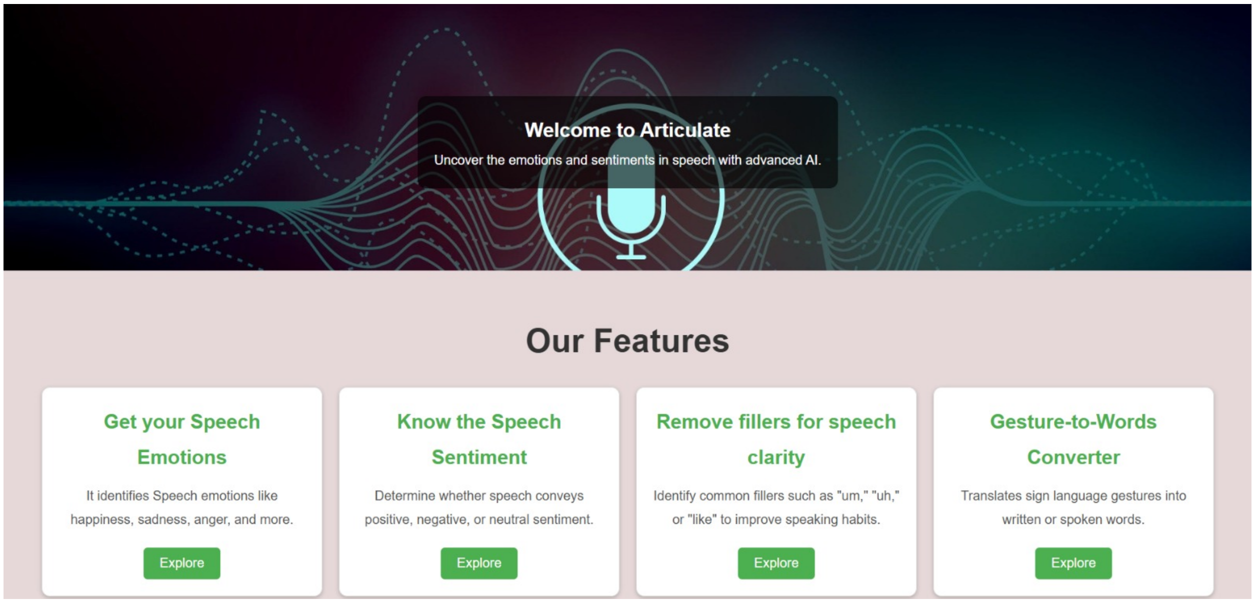
- Supports multiple audio formats for input.
- Real-time audio processing and predictions.
- Combines emotion detection with sentiment analysis for comprehensive information.
- Automatically handles file conversions and cleanup to prevent storage problems.

- Also includes Sign Language for Detection in Real time

7 Findings

- Successfully implemented a robust emotion prediction system with 83% precision using an MLP classifier.
- Enabled end-to-end sentiment analysis by integrating speech recognition and NLP techniques.
- Developed a user-friendly web interface for seamless audio analysis.

8 Project Images



Welcome to Articulate

Uncover the emotions and sentiments in speech with advanced AI.

Our Features

Get your Speech Emotions

It identifies Speech emotions like happiness, sadness, anger, and more.

Know the Speech Sentiment

Determine whether speech conveys positive, negative, or neutral sentiment.

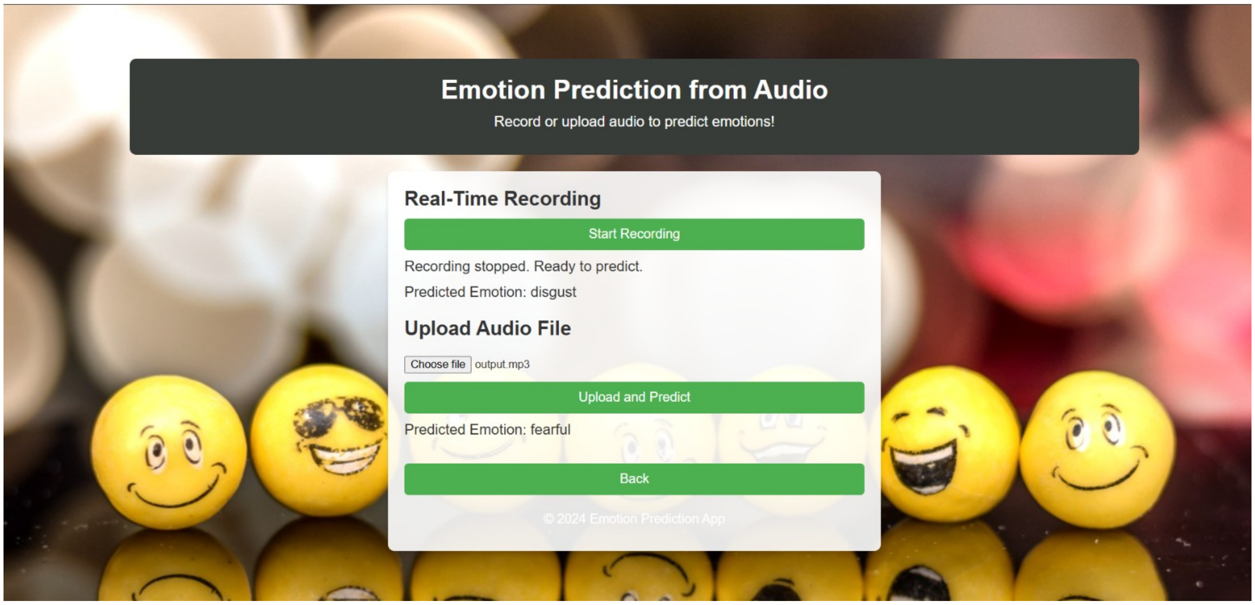
Remove fillers for speech clarity

Identify common fillers such as "um," "uh," or "like" to improve speaking habits.

Gesture-to-Words Converter

Translates sign language gestures into written or spoken words.

Figure 1: Home Page of the website



Emotion Prediction from Audio

Record or upload audio to predict emotions!

Real-Time Recording

Recording stopped. Ready to predict.

Predicted Emotion: disgust

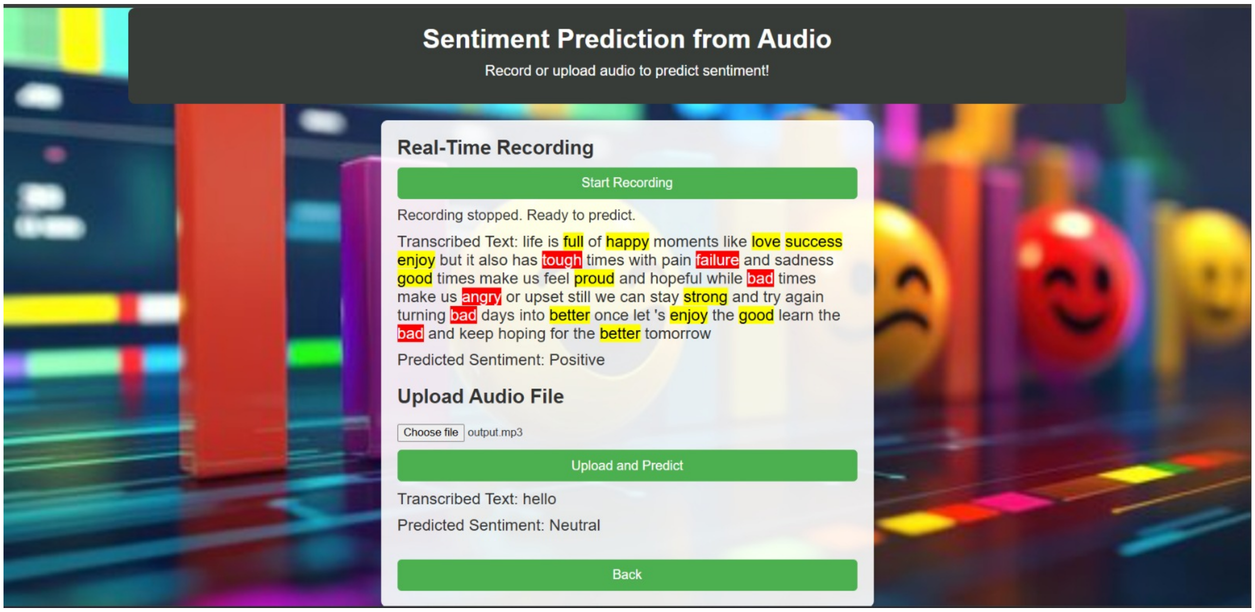
Upload Audio File

output.mp3

Predicted Emotion: fearful

© 2024 Emotion Prediction App

Figure 2: Emotion Prediction Page



Sentiment Prediction from Audio

Record or upload audio to predict sentiment!

Real-Time Recording

Recording stopped. Ready to predict.

Transcribed Text: life is full of happy moments like love success enjoy but it also has tough times with pain failure and sadness good times make us feel proud and hopeful while bad times make us angry or upset still we can stay strong and try again turning bad days into better once let 's enjoy the good learn the bad and keep hoping for the better tomorrow

Predicted Sentiment: Positive

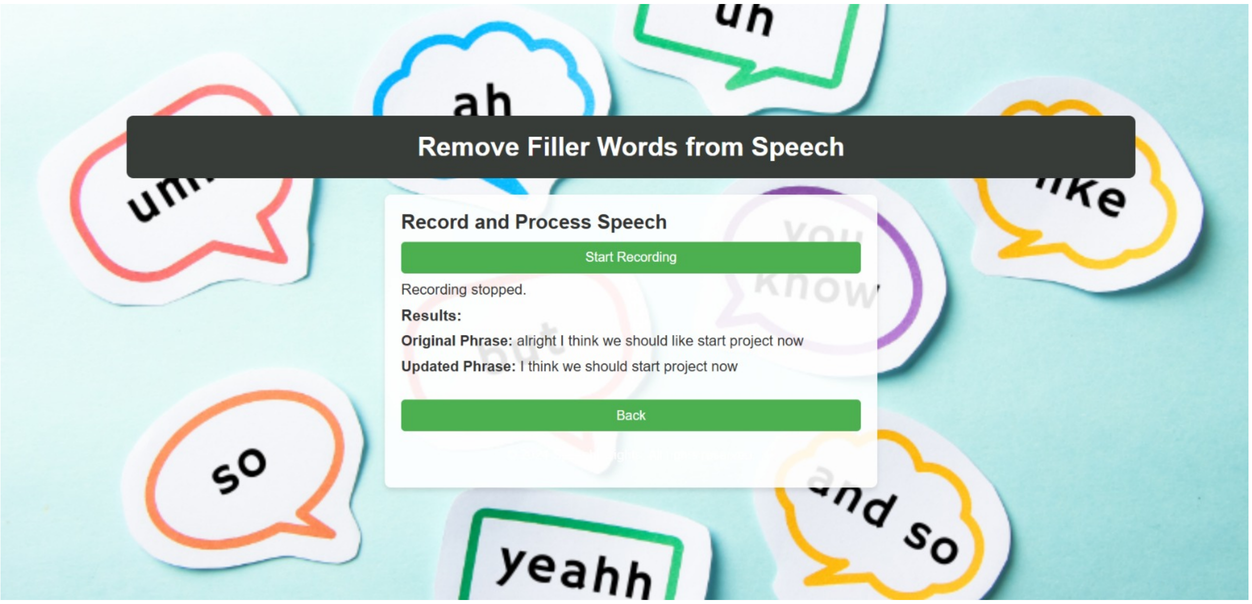
Upload Audio File

output.mp3

Transcribed Text: hello

Predicted Sentiment: Neutral

Figure 3: Sentiment Prediction from Speech



Remove Filler Words from Speech

Record and Process Speech

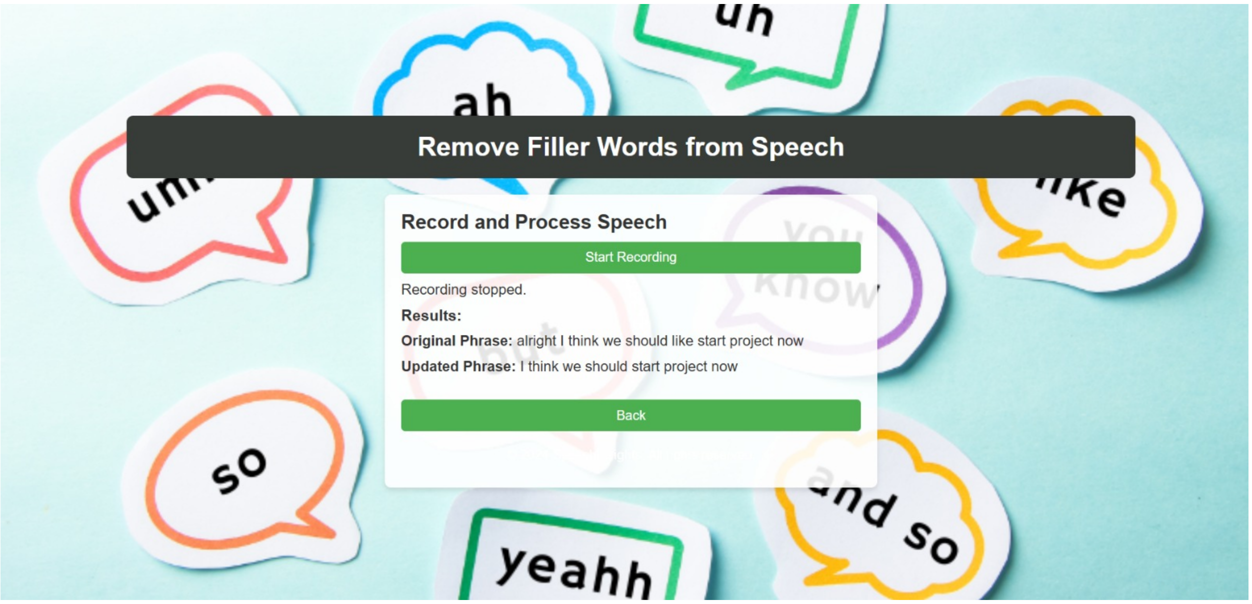
Recording stopped.

Results:

Original Phrase: alright I think we should like start project now

Updated Phrase: I think we should start project now

Figure 4: Speech Improvement by Filler words removal suggestions



Remove Filler Words from Speech

Record and Process Speech

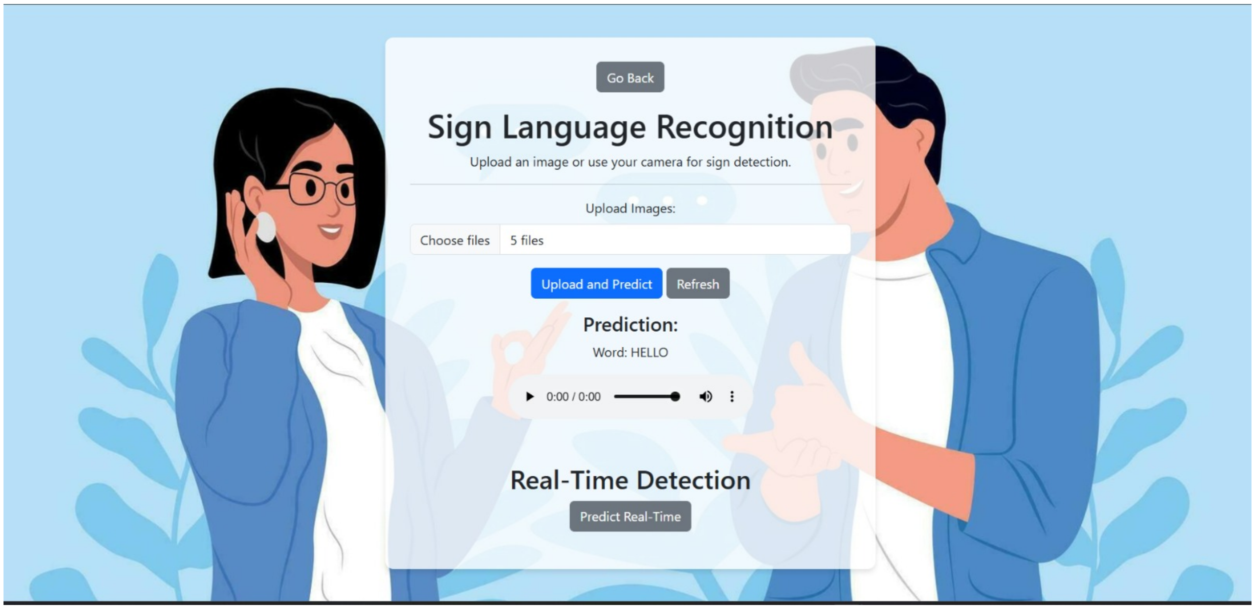
Recording stopped.

Results:

Original Phrase: alright I think we should like start project now

Updated Phrase: I think we should start project now

Figure 5: Filler Word corrected output



Sign Language Recognition

Upload an image or use your camera for sign detection.

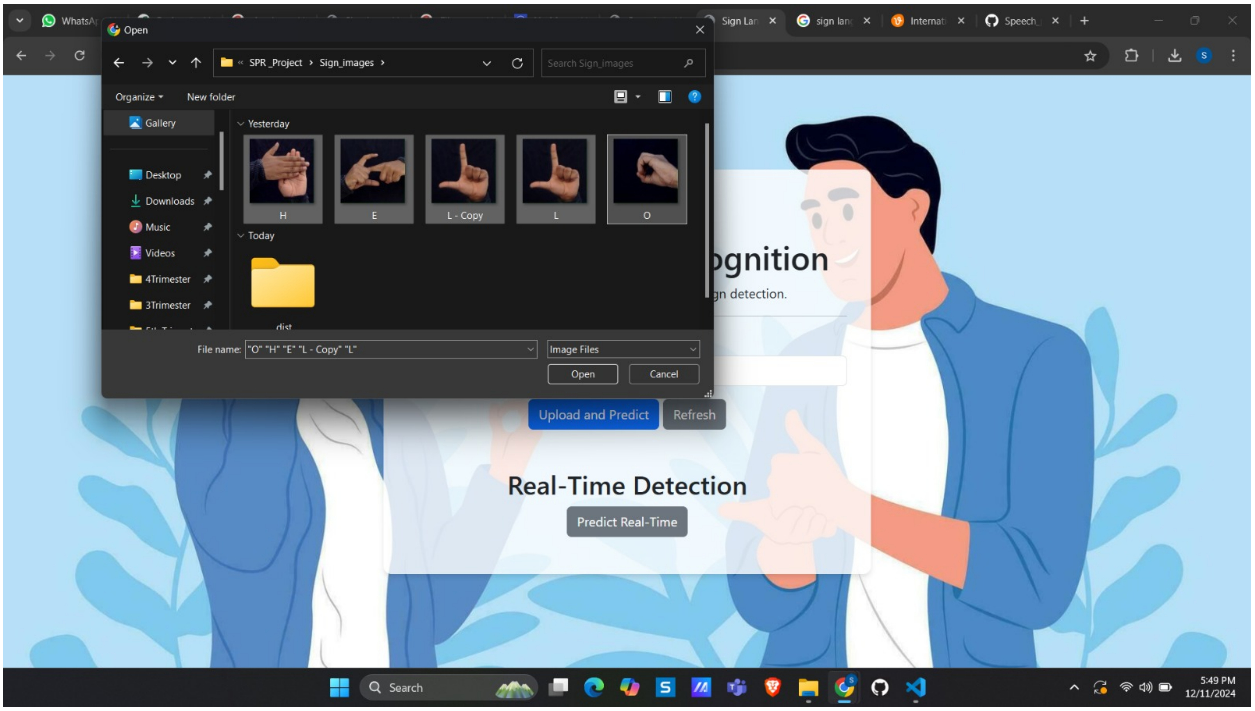
Upload Images: 5 files

Prediction:

Word: HELLO

Real-Time Detection

Figure 6: Sign Language to Speech conversion

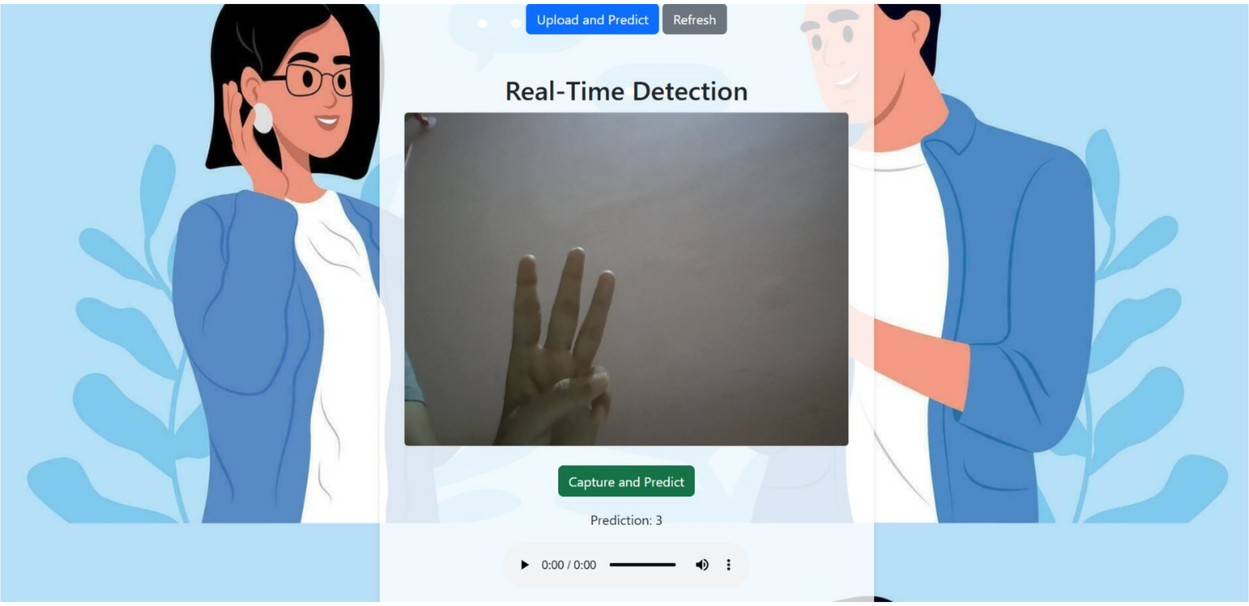


File name: "O" "H" "E" "L - Copy" "L"

Image Files

Real-Time Detection

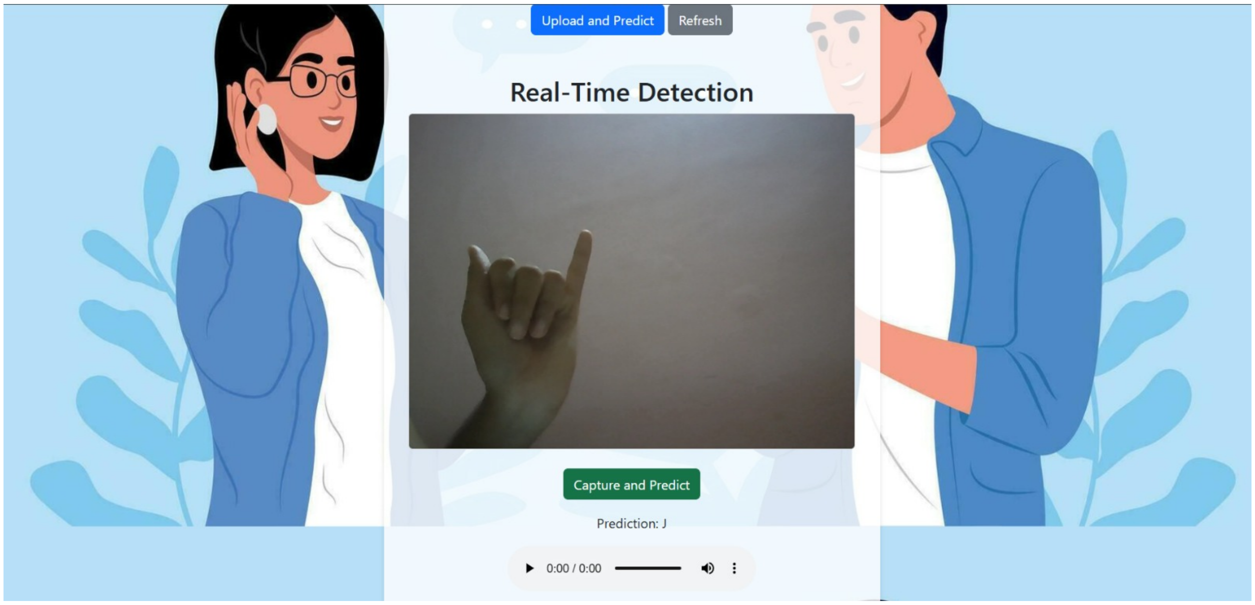
Figure 7: Choosing the images



Real-Time Detection

Prediction: 3

Figure 8: Real Time Utilization of the Functionality



Real-Time Detection

Prediction: J

Figure 9: Another Example of Sign to Speech function

References

[1] Abey Abraham and V Rohini. Real time conversion of sign language to speech and prediction of gestures using artificial neural network. *Procedia Computer Science*, 143:587–594, January 2018.

[2] D G Childers. Speech processing and synthesis for assessing vocal disorders. *IEEE Engineering in Medicine and Biology Magazine*, 9(1):69–71, March 1990.

[3] Tedd Kourkounakis, Amirhossein Hajavi, and Ali Etemad. Fluentnet: End-to-end detection of stuttered speech disfluencies with deep learning. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 29:2986–2999, January 2021.

[4] Colin Lea, Zifang Huang, Jaya Narain, Lauren Tooley, Dianna Yee, Dung Tien Tran, Panayiotis Georgiou, Jeffrey P Bigham, and Leah Findlater. From user perceptions to technical improvement: Enabling people who stutter to better use speech recognition. *From User Perceptions to Technical Improvement: Enabling People Who Stutter to Better Use Speech Recognition*, page 1–16, April 2023.

[5] Seong-Gyun Leem, Daniel Fulford, Jukka-Pekka Onnela, David Gard, and Carlos Busso. Selective acoustic feature enhancement for speech emotion recognition with noisy speech. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 32:917–929, December 2023.

[6] Yuanchao Li, Zeyu Zhao, Ondřej Klejch, Peter Bell, and Catherine Lai. Asr and emotional speech: A word-level investigation of the mutual impact of speech and emotion recognition. *INTERSPEECH 2023*, page 1449–1453, Aug 2023.

[7] Wenjin Liu, Jiaqi Shi, Shudong Zhang, Lijuan Zhou, and Haoming Liu. E-speech: Development of a dataset for speech emotion recognition and analysis. *International Journal of Intelligent Systems*, 2024(1), January 2024.

[8] Vikramjit Mitra, Zifang Huang, Colin Lea, Lauren Tooley, Sarah Wu, Darren Botten, Ashwini Palekar, Shrinath Thelapurath, Panayiotis Georgiou, Sachin Kajarekar, and Jefferey Bigham. Analysis and tuning of a voice assistant system for dysfluent speech. *Interspeech 2022*, August 2021.

[9] S. Rajaganapathy, B. Aravind, B. Keerthana, and M. Sivagami. Conversation of sign language to speech with human gestures. *Procedia Computer Science*, 50:10–15, January 2015.

[10] Taiba Majid Wani, Teddy Surya Gunawan, Syed Asif Qadri, Mira Kartiwi, and Eliathamby Ambikairajah. A comprehensive review of speech emotion recognition systems. *IEEE Access*, 9:47795–47814, 2021.