Smart Speech Detection Technique for Behaviour Analysis

Mr. Sahil V. Thakare[1], Mr. Tanishq V. Sahu[2], Mr. Dhairya I. Wadhwa[3],
Miss. Samiksha V. Katyarmal[4], Miss. Khusbhu N. Gulhane[5], Miss. Sakshi G. Bobade[6],
Miss. Shruti S. Deshmukh[7], Prof. U. V. Ramekar[8]
UG Students, Department of Electronics & Telecommunication Engineering[1,2,3,4,5,6,7]
Assistant Professor, Department of Electronics & Telecommunication Engineering[8]
SIPNA College of Engineering & Technology, Amravati, Maharashtra, India

Abstract: *The integration of speech recognition with artificial intelligence (AI) has transformed how humans interact with technology. This paper presents the development and implementation of a Smart Speech Recognition System using a mobile-based approach built on the Flutter framework.*

The system is divided into three primary modules: (1) real-time speech-to-text conversion, (2) emotion identification from the spoken input, and (3) AI-powered conversational responses using Google's Gemini API.

The project aims to deliver a smart, intuitive, and emotionally aware interface where user voice commands are not only understood but also responded to with contextual and emotionally relevant answers. With a focus on enhancing user interaction, this system serves as a foundational model for future emotionally intelligent virtual assistants and AI-driven human-computer interfaces.

This project focuses on the development of an Android application using Flutter, designed to convert speech into text, integrate with Gemine AI for advanced text analysis, and detect emotions within the converted speech or text. The application aims to provide users with a seamless and intelligent interface to transcribe spoken words, analyze them using natural language processing (NLP), and determine the underlying emotional tone. The core functionalities include real-time speech-to-text conversion, enhanced by Gemine AI's ability to extract meaningful insights, along with emotion detection that can assess sentiments such as happiness, sadness, anger, or neutrality.

The app leverages Flutter's cross-platform framework, ensuring an efficient and responsive user experience on Android devices. With the speech-to-text functionality, users can generate accurate text from their spoken input, making it useful for communication, transcription, and productivity purposes. Gemine AI is integrated to perform advanced analysis, offering features such as summarization, content recommendations, and predictive analytics. Emotion detection adds another layer of intelligence, providing feedback on the emotional context of conversations.

This application has broad potential use cases, from mental health monitoring and customer service to personal communication tools and educational resources. It represents a blend of mobile development, AI-driven natural language processing, and emotion analysis, contributing to the growing demand for emotionally aware AI systems and enhancing user interaction through innovative technology.

Keywords: Smart Speech Detection Technique, Emotion Identification, Speech-to-text conversion, AI-powered Conversational responses

I. INTRODUCTION

Speech is the most natural form of human communication. In recent years, the convergence of speech processing and AI technologies has revolutionized digital assistants, customer service bots, and accessibility tools. However, most existing speech recognition systems lack emotional awareness and struggle to deliver context-sensitive responses.

DOI: 10.48175/IJARSCT-25297

Our Smart Speech Recognition System aims to bridge this gap by integrating three critical capabilities: voice-to-text recognition, emotion detection, and AI-powered responses—all through a user-friendly mobile application

The primary motivation for developing this app is to enhance user interaction by making communication more accessible and emotionally aware. Speech-to-text technology allows users to interact with the app through voice commands, eliminating the need for typing and improving accessibility for individuals with disabilities. Integrating AI to analyze the transcribed text further enables the app to deliver intelligent responses, suggestions, or insights. By detecting emotions, the app can provide more personalized and context-aware experiences, which can be beneficial in multiple domains like customer service, mental health, personal communication, and education.

The use of Flutter, a cross-platform mobile development framework, ensures that the app can be developed efficiently for both Android and iOS, making it accessible to a broader audience. Additionally, the project leverages APIs and AI models to process speech and text, enabling real-time processing and delivering fast results.

This project bridge the gap between human interaction and technology by creating an app that can understand not only what is said but also how it is said, thus enhancing the overall user experience.

The evolution of mobile technologies and cloud-based AI services enables real-time speech analysis and interaction without significant computational overhead. This project utilizes Flutter for cross-platform development, Google's Gemini API for natural language understanding and emotion detection, and speech recognition libraries to handle voice input. This synergy of technologies enhances the interactive potential of AI systems, making them more human-like in conversation and responsiveness.

II. LITERATURE REVIEW

Title	Author	AI-Methodology
Deep Neural Networks for Acoustic Modeling in Speech Recognition.	**Hinton, G., et al. (2012)**	This seminal paper discusses the use of deep neural network (DNNs) in improving acoustic modelling for speech recognition systems. The author demonstrate that (DNNs) significantly outperform traditional methods, such as Gaussian Mixture Models, leading to advancements in accuracy and robustness in noisy environments.
Advances in large Vocabulary continuous speech recognition.	**Huang, X., et al. (2014)**	This article reviews the developments in large vocabulary continuous speech recognition (LVCSR) systems, highlighting the integration of deep learning techniques. The authors provide insights into challenges such as speaker variability and environmental noise, and propose solutions including end-to-end training approaches.
Out of the Box: Unsupervised Learning of speech features.	**Chung, J. S. & Zisserman, A. (2016)**	Chung and Zisserman propose an unsupervised learning framework for extracting speech features without labeled data. This approach shows promise in reducing the need for extensive annotated datasets, enabling more scalable solutions in speech detection tasks.
Robust Speech Recognition using deep learning technique	**Wang, Y., & Makkone, T. (2017)**	This paper explores robust speech recognition methods leveraging deep learning architectures, specifically convolutional neural networks (CNNs) and recurrent neural networks (RNNs). The study demonstrates how these techniques can enhance performance in adverse conditions, making speech detection more reliable in practical applications.

DOI: 10.48175/IJARSCT-25297

Artificial Intelligence in Speech Recognition: A Review	**Vasilakos, A. V., et al. (2019)**	This review surveys various AI methodologies employed in speech recognition, including machine learning and deep learning techniques. The authors analyze the effectiveness of different algorithms, discuss current trends, and identify areas for future research, particularly in enhancing multilingual recognition.
End-to-End Speech Recognition	**Zhang, Y., et al. (2012)**	The authors provide a comprehensive overview of end- to-end speech recognition systems that streamline the process from audio input to text output. They compare various architectures, such as sequence-tosequence models and transformers, underscoring their efficiency and potential for real-time applications.

The literature on speech detection algorithms in AI reveals significant advancements, particularly through deep learning techniques. Key contributions include improved acoustic modeling, robustness in noisy environments, and the potential of unsupervised learning. Ongoing research continues to refine these algorithms, aiming for greater accuracy, reduced data requirements, and enhanced multilingual capabilities, paving the way for more intuitive and reliable speech recognition systems.

III. PROPOSED METHODOLOGY

The development of this project involves a modular and layered approach that ensures scalability, maintainability, and efficient processing of voice data. The application is developed using the Flutter framework, chosen for its cross-platform compatibility and expressive UI capabilities. The system is divided into three primary modules: Speech-to-Text Conversion, Speech Emotion Identification, and Speech-Based AI Query Response. Each module follows a distinct flow but remains interconnected within the application architecture. The methodology encompasses both the technical implementation and logical workflow of each module. The methodology is divided into the following stages: (1) Requirement *Analysis Objective: Clearly define the project goals, target user group, and application use cases.* (2) Design & Planning *Objective: Develop a structured design for the application architecture and user interface.* (3) Technology Selection *Objective: Choose the most suitable tools, APIs, and libraries to implement the required functionalities.* (4) Development Phase *Objective: Build and integrate the core functionalities of the application*

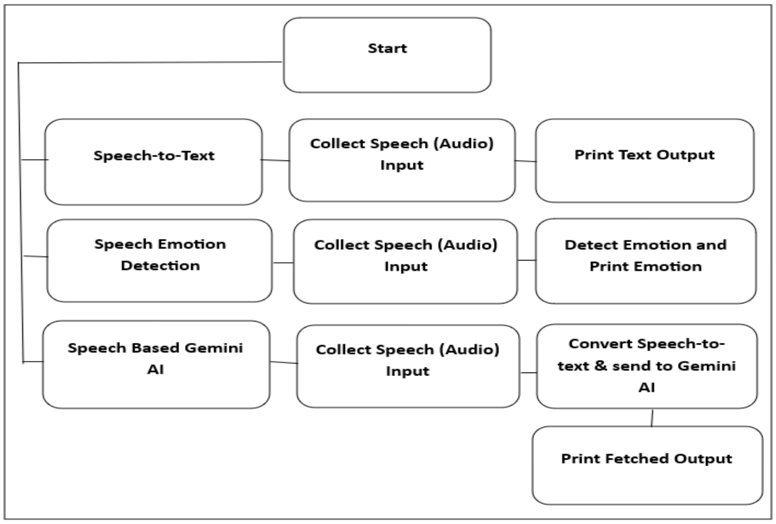
Start
Speech-to-Text
Collect Speech (Audio) Input
Print Text Output
Speech Emotion Detection
Collect Speech (Audio) Input
Detect Emotion and Print Emotion
Speech Based Gemini AI
Collect Speech (Audio) Input
Convert Speech-to-text & send to Gemini AI
Print Fetched Output

Fig. 1 Smart Speech Recognition System

DOI: 10.48175/IJARSCT-25297

Basically there are 3 types of uses of this technique: (1) First one is speech to text conversion, in this the input is provided as a audio in speech format and the speech present in the audio is converted into text. (2) The second use is to detect the behaviour or we can say that detect the emotion present in the audio (speech) given as a input by the help of algorithm. (3) The third one is “speech based AI”. In this, we get heavy data as a audio input we covert it into text format and send it to a Gemini a AI and then Gemini AI print the fetched output. (4) Basically Gemini is Google's AI-powered assistant that can help with a variety of tasks, including: Research, Sales, Marketing etc

IV. CONCLUSION

The Smart Speech Recognition System successfully integrates speech-to-text conversion, emotion detection, and conversational AI to provide a more human-centric voice interface. By understanding not only the content but also the emotional intent behind user speech, the system offers more meaningful and engaging interactions.

The development of an Android application using Flutter for speech-to-text conversion, AI integration via Gemine AI, and emotion detection presents an innovative and promising solution with diverse use cases in personal productivity, communication, education, healthcare, and customer service. The project capitalizes on cutting-edge technologies to provide users with real-time transcription, deep text analysis, and an understanding of emotional context, making the app interactive, intelligent, and highly usercentric

It bridges the gap between traditional voice assistants and emotionally intelligent AI, paving the way for future innovations in smart personal assistants, teletherapy tools, and AI-driven customer service applications.

In conclusion, the integration of speech-to-text and emotion detection technologies, powered by AI, offers vast potential to enhance user experiences, but its success depends on balancing the benefits with the technical, ethical, and operational challenges. With a strategic approach to overcoming these hurdles, this application can become a highly valuable tool across multiple industries, providing both individuals and businesses with deeper insights into communication and emotional intelligence.

The results demonstrate that such an application is feasible and effective on modern mobile platforms. With enhancements like multi-language support, offline AI capabilities, and advanced emotion analytics, this system has the potential to evolve into a powerful tool for both commercial and personal use.

REFERENCES

[1]. Huang, X., et al. (2014). "Advances in Large Vocabulary Continuous Speech Recognition." IEEE Signal Processing Magazine.

[2]. Wang, Y., & Makkonen, T. (2017). "Robust Speech Recognition using Deep Learning Techniques." IEEE Transactions on Audio, Speech, and Language Processing.

[3]. Chung, J. S., & Zisserman, A. (2016). "Out of the Box: Unsupervised Learning of Speech Features." arXiv preprint arXiv:1601.03588.

[4]. Vasilakos, A. V., et al. (2019). "Artificial Intelligence in Speech Recognition: A Review." Future Generation Computer Systems.

[5]. Zhang, Y., et al. (2021). "End-to-End Speech Recognition: An Overview." IEEE Transactions on Audio, Speech, and Language Processing.

[6]. Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. ICASSP.

[7]. Zhao, S., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep CNN and LSTM. Biomedical Signal Processing and Control, 47, 312–323.

[8]. Binali, H., Wu, C., & Potdar, V. (2010). A new significant area: Emotion detection in text. IEEE International Conference on Industrial Technology.

[9]. Radford, A., et al. (2019). Language Models are Unsupervised Multitask Learners. OpenAI.

[10]. Brown, T. B., et al. (2020). Language Models are Few-Shot Learners. arXiv:2005.14165.

[11]. Google Developers (2023). Speech-to-text API https//cloud.google.com/speech-to-text

[12]. Flutter.dev (2024). Build beautiful apps for any screen. https://flutter.dev/

DOI: 10.48175/IJARSCT-25297