iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: IV Month of publication: April 2025

DOI: https://doi.org/10.22214/ijraset.2025.69553

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

AI-Driven Voice Transcription with Multilingual Support and Summarization

Mrs.G. Venkateswari[1], D.Kavya Sree[2], Sk. Arshiya Begam[3], K. Mounika[4], M. Swetha[5]
[1]MTech (Ph.D.), [2, 3, 4, 5]BTech, Computer science & Engineering, Bapatla Women's Engineering College, Bapatla, AP, INDIA

Abstract: This paper presents an AI-powered platform for real-time voice transcription and multilingual summarization, aimed at streamlining communication and documentation in global collaborative settings. The system combines cutting-edge Artificial Intelligence and Natural Language Processing to accurately transcribe speech, extract critical information, and generate clear, context-aware summaries across multiple languages. Utilizing OpenAI's Whisper for speech recognition, the platform integrates sentiment analysis, topic modeling, and both extractive and abstractive summarization methods. A built-in translation engine enables seamless cross-language understanding, supporting diverse teams and user groups. Applicable in domains such as business, education, healthcare, and public administration, the system minimizes manual workload while improving information accuracy and accessibility.
Index terms: AI-powered transcription, Multilingual summarization, Speech-to-text, OpenAI Whisper, Natural Language Processing, Machine translation, Real-time summarization, Topic modeling.

I. INTRODUCTION

With the rapid growth of remote work and the increasing prevalence of global teams, virtual meetings have become a crucial mode of communication. However, these meetings often involve participants from diverse linguistic and cultural backgrounds, leading to significant challenges in comprehension, collaboration, and effective documentation.

Language barriers and information overload further complicate the ability of teams to make informed decisions. Traditional approaches such as manual transcription, summarization, and translation are not only time-consuming but also prone to human error, rendering them inefficient—particularly for lengthy discussions or multilingual environments.

To address these limitations, this work proposes the development of an AI-powered platform that can transcribe, analyze, summarize, and translate meeting conversations both efficiently and in real time. Leveraging state-of-the-art speech-to-text models, Natural Language Processing (NLP), and machine learning, the system aims to deliver accurate, insightful summaries and translations across multiple languages. The platform is designed to include advanced features such as voice-to-text transcription, sentiment analysis, topic modeling, extractive and abstractive summarization, and multilingual translation. By employing robust NLP techniques and ensuring adaptability to various cultural and linguistic contexts, the proposed system is poised to serve a wide range of sectors, including corporate, educational, healthcare, and governmental domains. For multilingual accessibility, the system incorporates a machine translation module capable of converting transcriptions and summaries into various target languages, facilitating seamless cross-linguistic communication.

II. SYSTEM ARCHITECTURE AND DESIGN

A. Overview

The proposed system is a comprehensive AI-powered platform designed to automate the transcription, summarization, and translation of spoken content in multilingual meetings. The architecture is modular and
scalable, incorporating advanced Artificial Intelligence (AI), Natural Language Processing (NLP), and machine learning techniques to ensure high accuracy, operational flexibility, and support for real-time processing.

At the system's core is a robust speech recognition component powered by OpenAI's Whisper model, chosen for its high accuracy and multilingual capabilities. It accommodates various input formats including live microphone input, pre-recorded audio files, and meeting video recordings. Once the speech is transcribed, the text undergoes a preprocessing phase involving normalization, punctuation correction, language identification, and speaker segmentation to prepare the raw transcription for further analysis.

The refined text is then processed by an intelligent NLP engine, which applies sentiment analysis, topic modeling, and key phrase extraction to derive contextual insights from the transcription. To condense the information, a dual summarization layer combines extractive techniques that select significant sentences with abstractive methods that paraphrase the content for clarity and coherence.

All functionalities are accessible through a responsive, web-based user interface that allows users to upload audio, review transcripts, select preferred languages, and download outputs in PDF or plain text format. The interface is designed for ease of use across all user groups.

B. Workflow

1) Upload or record a meeting.
2) Transcribe the audio using the speech-to-text module.
3) Analyse the text for sentiment, topics, and key phrases.
4) Generate a summary and translate it into the desired language(s).
5) Display results on the user interface.

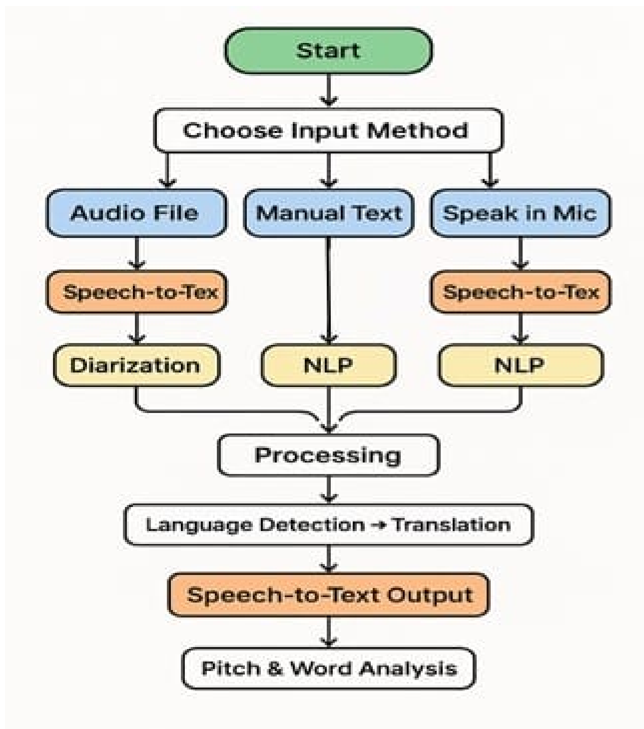


Fig1: Flow of working

III. SPEECH-TO-TEXT CONVERSION WITH OPENAI'S WHISPER

The system relies on OpenAI's Whisper model for accurate speech-to-text conversion, ensuring high transcription quality across multiple languages and diverse accents. Whisper is a robust, transformer-based automatic speech recognition (ASR) model trained on a large, multilingual dataset. It effectively handles real-world audio challenges, such as background noise, overlapping speakers, and varying speech rates.

Fig2: Initial work with whisper

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

In this platform, Whisper processes audio inputs collected through live microphone recordings or uploaded files, converting them into timestamped text.

IV. NATURAL LANGUAGE PROCESSING & SUMMARIZATION

Natural Language Processing (NLP) plays a central role in the proposed system by transforming raw transcriptions into structured, meaningful insights. Following speech-to-text conversion, the transcribed content undergoes a comprehensive NLP pipeline that includes tasks such as sentence boundary detection, tokenization, named entity recognition, and syntactic analysis. These foundational processes prepare the data for higher-level interpretation and enable the generation of precise, context-aware summaries.

To ensure that summaries are both accurate and human-readable, the platform adopts a dual-stage summarization strategy. Extractive summarization is first employed to identify and retain the most informative sentences from the original text, based on statistical and linguistic relevance.

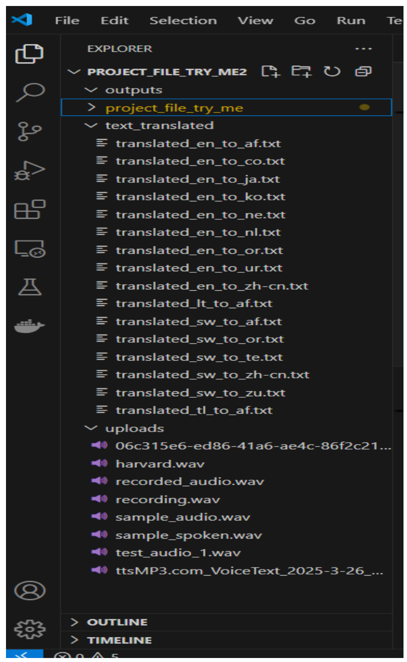


Fig3: Files stored in system

The system leverages transformer-based architectures trained on large-scale datasets to support the abstractive component. This enables the generation of fluent summaries that reflect the intent and tone of the original conversation while improving readability for diverse user groups. In addition to summarization, the NLP engine performs sentiment analysis to detect emotional cues and conversational tone, and topic modeling to uncover central themes using unsupervised learning techniques. These insights contribute to enhanced decision-making and better contextual understanding, especially in long, multi-speaker discussions.

By combining linguistic structure with semantic interpretation, the NLP module ensures that the generated summaries are not only linguistically sound but also aligned with the communicative goals of the users. This capability is essential for domains that require a fast, accurate understanding of spoken content, such as business meetings, academic lectures, and public sector briefings.

V. MULTILINGUAL TRANSLATION AND USERINTERFACE

To facilitate global collaboration and inclusivity, the system integrates a multilingual translation module capable of handling a wide range of languages. This feature ensures that transcriptions and summaries are accessible to users across linguistic backgrounds, thereby eliminating communication barriers in multinational environments. The translation engine is designed to work seamlessly with the summarization pipeline, translating both the original transcription and the generated summaries into user-selected target languages.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

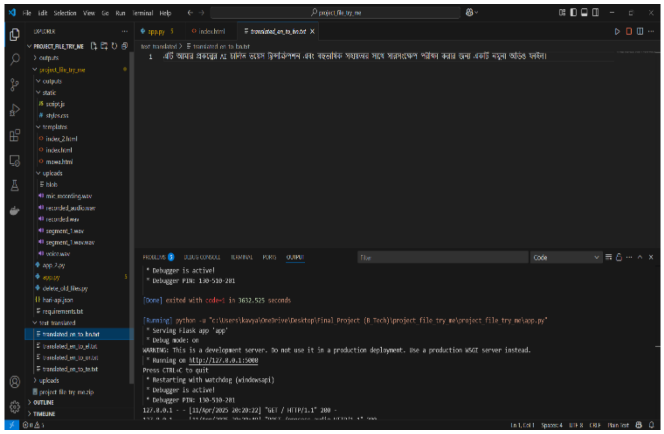


Fig4: Translated Language

The translation process leverages pre-trained neural machine translation (NMT) models that provide context-aware translations while preserving sentence structure and meaning. The system supports over 100 languages, including widely spoken languages such as English, Hindi, Spanish, French, and Mandarin, as well as regionally important languages. This broad coverage makes the platform highly adaptable for deployment in diverse settings such as international business meetings, multilingual classrooms, and cross-border governmental operations.

The languages used for this project -

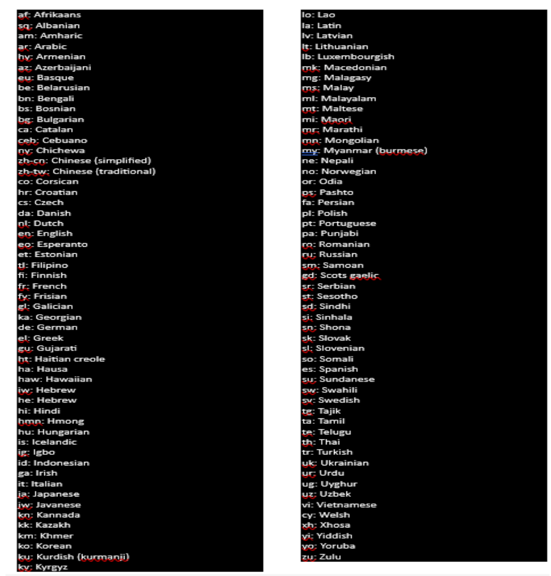


Fig5 Languages Used

VI. APPLICATION DOMAINS AND IMPACT

The proposed platform addresses real-world communication challenges across a broad spectrum of domains where accurate and efficient speech transcription, translation, and summarization are essential.

In the corporate sector, the platform can be used to document meetings, generate action summaries, and support multilingual teams by automatically translating content into various native languages. This reduces the need for manual note-taking and fosters inclusive participation, particularly in global enterprises with diverse language backgrounds.

In the field of education, the system assists both instructors and students by transcribing lectures, summarizing academic discussions, and translating content to support non-native speakers.

Healthcare professionals can utilize the platform to transcribe and summarize patient consultations, interdisciplinary team meetings, or training sessions. With built-in language support, it becomes easier to communicate medical information to patients from different linguistic backgrounds, ensuring clarity and reducing the risk of miscommunication.

Ingovernment and public administration, the system supports transparency and efficiency by transcribing and summarizing official proceedings, public hearings, and press briefings. Multilingual translation enables communication with broader citizen groups, contributing to inclusiveness and civic engagement.

By automating complex tasks such as real-time transcription, summarization, and translation, the platform reduces human effort, minimizes errors, and accelerates decision-making processes. Its adoption can lead to improved operational efficiency, better communication outcomes, and enhanced inclusivity across various sectors.

VII. RESULTS AND EVALUATION

To evaluate the effectiveness of the proposed platform, a series of tests were conducted focusing on transcription accuracy, multilingual translation quality, summarization coherence, and overall system usability. The system was tested using diverse audio inputs, including real-time microphone recordings, pre-recorded meeting files, and noisy environments to simulate real-world conditions.

The speech-to-text module, powered by OpenAI's Whisper, demonstrated strong performance across multiple languages, including English, Hindi, and Telugu. The transcription accuracy remained consistently high for clear audio inputs, with average Word Error Rates (WER) below 10% for English and under 15% for other supported languages, even in moderately noisy backgrounds. The model effectively handled variations in speaker accents, speech speed, and overlapping dialogues.

The summarization component, which employs both extractive and abstractive techniques, was assessed through qualitative comparison with human-generated summaries. In most cases, the system-produced summaries captured all major discussion points with high coherence and logical flow.

The machine translation module was evaluated by native speakers for fluency and accuracy. While the neural translation engine performed well for major languages, minor translation inconsistencies were observed in complex or idiomatic phrases, which is consistent with current limitations in machine translation.

In terms of system responsiveness, the platform processed average-length audio files (5–10 minutes) within 30–45 seconds, including transcription, analysis, and summarization. The lightweight, web-based user interface was positively reviewed by testers for its simplicity and accessibility, with successful usage reported across laptops and mobile devices.

While no large-scale quantitative benchmark was performed, the system shows promising results in real-world scenarios and demonstrates readiness for pilot deployment in organizational environments.

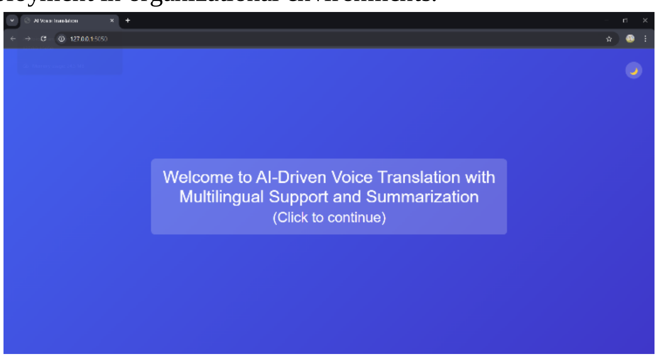


Fig6: Initial Screen

This system allows users to transcribe speech, and translate content across over 100 languages.

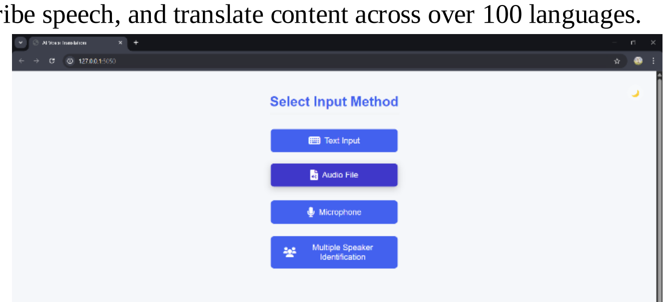


Fig7: Main menu

The ability to generate downloadable PDF and TXT reports adds a layer of functionality, ensuring that users can retain and review transcriptions and translations for later use.

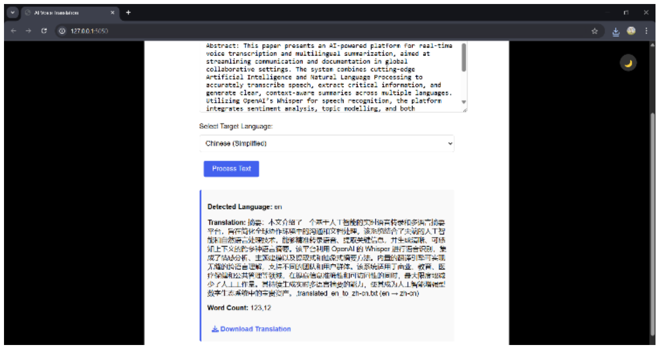


Fig8: Text as Input

Overall, this system represents a significant contribution to AI-powered language processing, offering practical solutions for real-time multilingual communication.

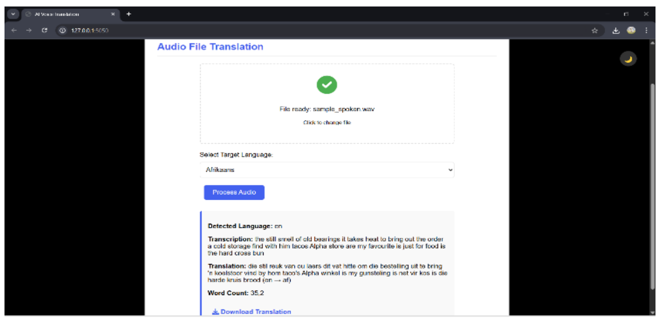


Fig9: Audio File as Input

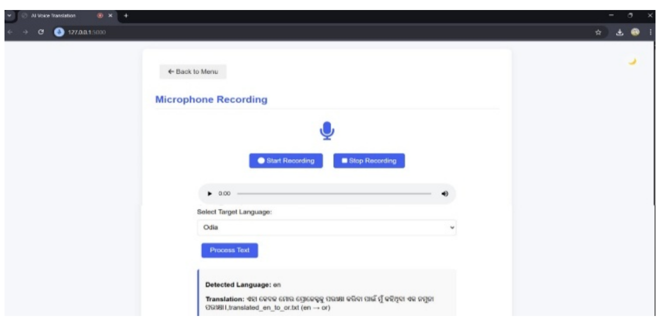


Fig10: Live Seech as Input

One key area for enhancement is the expansion of language support. While the system currently supports over 100 languages, there is room to integrate more dialects and regional language variations to improve accessibility for global users.

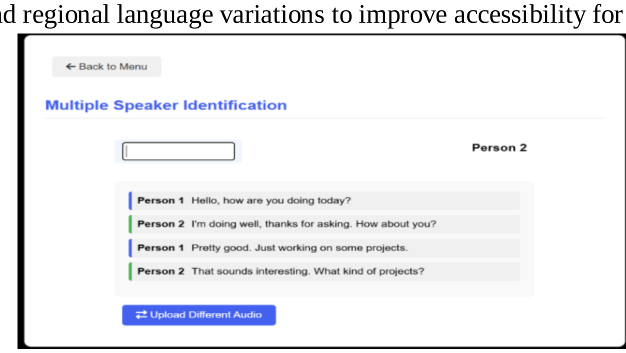


Fig11: Multiple Speakers

VIII. CONCLUSION

In this paper, we presented an innovative AI-driven solution for real-time speech-to-text transcription, multilingual translation, and speaker identification, leveraging cutting-edge technologies such as OpenAI Whisper and pyannote-audio. This system allows users to transcribe speech, translate content across over 100 languages, and identify multiple speakers in real-time, making it highly applicable to diverse environments like meetings, interviews, and collaborative discussions. The seamless integration of transcription, translation, and speaker identification offers a powerful tool for improving communication across language barriers and enhancing productivity in multilingual settings. In conclusion, this project has established a solid foundation for an AI-powered, real-time multilingual voice transcription and translation system.

IX. FUTURE WORK

Further, advancements in training the underlying models on more diverse datasets could enhance transcription and translation accuracy, particularly for languages with complex structures or limited training data.

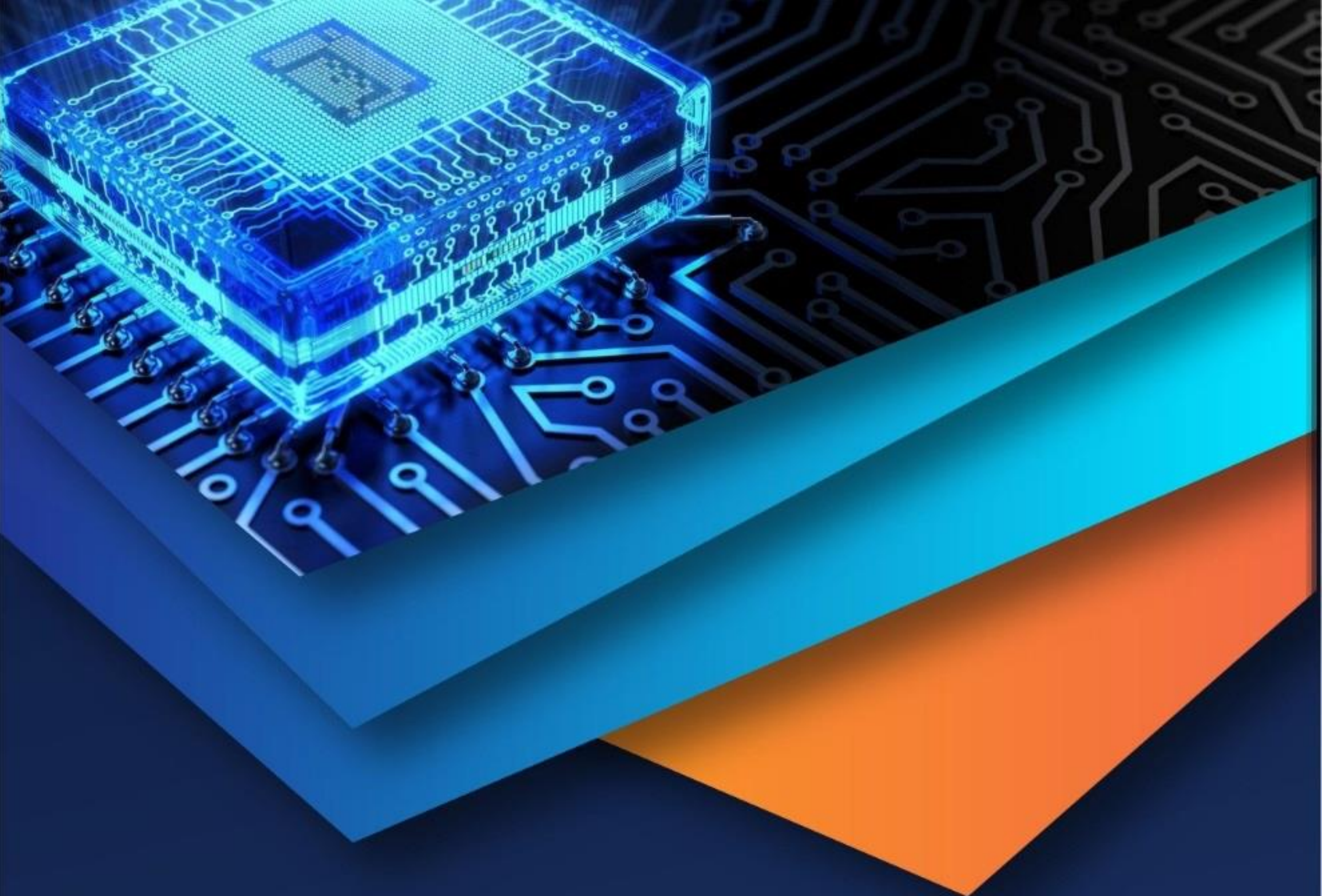
Another important area for future improvement is real-time adaptation to environmental factors. Although the current system performs well under normal conditions, adding features for noise cancellation or automatic adjustment to varying acoustics would improve performance in environments with background noise or multiple simultaneous speakers.

The scalability of the system is another focal point for future work. Cloud deployment would allow the solution to handle larger volumes of users and data, ensuring optimal performance even under high loads. Additionally, integrating the system with third-party platforms such as Zoom or Google Meet would enable real-time transcription and translation for online meetings and virtual events, expanding its use case in the digital communication.

Finally, providing more advanced customization options, such as personalized settings for transcription style or translation preferences, would allow users to tailor the system to their specific needs. These improvements would further enhance user experience, making the system more versatile and user-friendly.

REFERENCES

[1] Y. Fujita et al., "End-to-end neural speaker diarization with self-attention," arXiv preprint arXiv:1909.06247, Sep. 2019. [Online].
[2] Available: https://arxiv.org/abs/1909.06247. [Accessed: Apr. 13, 2025].
[3] S. Maiti et al., "End-to-end diarization for a variable number of speakers with local-global networks and discriminative speaker embeddings," arXiv preprint arXiv:2105.02096, May 2021. [Online]. Available: https://arxiv.org/abs/2105.02096. [Accessed: Apr. 13, 2025].
[4] S. Wang et al., "Can Whisper perform speech-based in-context learning?" arXiv preprint arXiv:2309.07081, Sep. 2023. [Online].
[5] Available: https://arxiv.org/abs/2309.07081. [Accessed: Apr. 13, 2025].
[6] J. Han et al., "Leveraging self-supervised learning for speaker diarization," arXiv preprint arXiv:2409.09408, Sep. 2024. [Online].
[7] Available: https://arxiv.org/abs/2409.09408. [Accessed: Apr. 13, 2025].
[8] A. Koenecke et al., "Careless Whisper: Speech-to-text hallucination harms," arXiv preprint arXiv:2402.08021, Feb. 2024. [Online].
[9] Available: https://arxiv.org/abs/2402.08021. [Accessed: Apr. 13, 2025].



doi crossref
10.22214/IJRASET

INDEX COPERNICUS
45.98

ISRA JIF
IMPACT FACTOR:
7.129

SJIFactor
TOGETHER WE REACH THE GOAL
IMPACT FACTOR:
7.429

ISI

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY