Grenze International Journal of Engineering and Technology, January Issue

Speak-Sure Interview Assistant: Application of Speech and Gesture Recognition

Sridhar H. S[1], Sanket Raj[2], Sakshi Priya[3] and Atik Kumar Singh[4]
[1-4]Siddaganga Institute of Technology, Department of Electrical and Electronics Engineering, Tumakuru, Karnataka, India
Email: sri054@sit.ac.in, sanketraj526@gmail.com, sakshi62ss@gmail.com, atik.kumar960@gmail.com

***Abstract*— Communication has always been a crucial part of everyday life and is fundamental to profes-sional success. In recent years, it has become a key indicator of a person's achievements in pro-fessional settings and is often considered an essential skill for any potential candidate in an or-ganization. Recognizing this trend, this paper aims to develop interview assistant software by in-tegrating speech and gesture recognition technologies, focusing on both verbal and non-verbal aspects of communication. A website is developed using programming languages and frame-works to accurately capture and detect the user's speech and gestures. The proposed software employs API endpoints for speech recognition and utilizes TensorFlow, an open-source machine learning framework, for gesture recognition. TensorFlow is trained using datasets created with Teachable Machine, a web-based tool.**

***Index Terms*— HCI-Human Computer Interaction, TensorFlow, API- Application Programming Interface, Speech Detection, Gesture Detection.**

I. INTRODUCTION

Language proficiency and communication skills have emerged as significant challenges for the youth in today's world. With the increasing emphasis on global connectivity and digital communication, the ability to effectively articulate thoughts and ideas has become more crucial than ever. Young people face various hurdles in this domain, including limited vocabulary, difficulties with grammar and challenges in conveying their messages clearly and persuasively. Addressing these challenges requires practical training to enhance language skills and communication abilities. Communication can be broadly categorized into two aspects: verbal and non-verbal. Verbal communication includes clarity, speed, and the meaningfulness of speech, all of which are essential for accurately conveying messages. Non-verbal communication, encompassing gestures and body language, plays a crucial role in reinforcing spoken words and expressing emotions. This project seeks to address these challenges by integrating both Speech and Gesture recognition technologies, creating a comprehensive solution designed to enhance these communication elements and ultimately improve one's ability to deliver messages effectively.
Speech recognition aims to assess the verbal aspect of one’s speech that primarily converts speech to text, and later assess this text according to the data that needs to be fetched. Human Gestures will make the communication more comprehensible and to express ourselves humans are habituated by using gesture, not just in our daily lives but also gained interest in the area of HCI and computer vision. Ma-chine learning and deep learning algorithms increase the level of abstraction and improvise the recognition and classification process. The objective of this system is to develop a speech recognition model to analyze the vocal aspect of communication and a gesture and body language recognition model to assess the non-vocal aspects.

Grenze ID: 01.GIJET.11.1.558_1

An integrator module will be devised using Flask, TensorFlow, and JavaScript to combine the data from both machine learning models and provide feedback based on the analysis. The system will be tested on real-time participants to evaluate its performance. Personalized feedback will be given to each par-ticipant to help improve their communication skills and provide users with a seamless experience akin to an interview scenario.

II. Literature Survey

Kooragama K.G.C.M, Jayashanka L.R.W.D., Munasinghe J.A., Jayawardana K.W, Muditha Tissera, Thilini Buddhika [1] In this paper, a novel approach is proposed to analyze English speech delivering skills by implementing an online tool, “Speech Master”. This tool analyzes English speech delivering skills in many aspects including the content, grammar, and grammatical efficiency, expressions, and flow of the speech. For such components, separate algorithms were built mainly using NLP techniques, deep learning, and machine learning algorithms.

A. Deshpande, R. Pandharkar and S. Deolekar [2] This paper introduces "Speech Coach," a web-based framework. It offers a platform for users to practice speeches, build confidence, and enhance oratory skills. The framework provides tools for visualizing speech elements and machine-generated transcrip-tions, allowing users to self-assess their performance and compare it with other speeches. Developed with open-source software, it focuses on the English-speaking population in India and employs both visual and quantitative methods for speech analysis and improvement.

Zheng Zhang [3] The article presents an English teaching system that leverages artificial intelligence and deep learning models. It includes three modules—speech recognition, machine translation, and speech synthesis—to offer personalized learning experiences and enhance efficiency and quality. The paper details data preprocessing, system architecture, implementation, and testing, demonstrating the system's superior performance in all three areas.

Sagnik Das, Nisha Gandhi, Tejas Naik, Roy Shilkrot [5] The paper proposes a speech stream manipula-tion system designed to help non-professional speakers produce fluent, professional-like speech by addressing disfluencies such as "uh," "um," filler words, and long pauses. The system uses a predictive model trained on professional speech data to segment and adjust these disfluencies, resulting in smoother and more confident speech. Quantitative evaluations show a significant improvement in flu-ency by reducing pauses and fillers.

Ainampudi Kumari Sirivarshitha,Koneru Lakshmaiah Education Foundation, Kothamasu Santhi Priya, Vasantha Bhavani [9] This study explores a face recognition technology that converts facial features into a unique face print for identification, comparing new images to those stored in a database using deep learning. It emphasizes image quality and uses Python with libraries like OpenCV and Face Recognition. The system evaluates facial feature vectors against data collection to identify individuals, demonstrating that OpenCV excels in locating and recognizing faces. The study develops algorithms for face detection and recognition, leveraging high image resolution to improve accuracy and speed in face identification.

Xuefeng Chen, Xin Luo, Xingyao Liu, Jikang Fang [10] This paper proposes an improved eyes localiza-tion algorithm to address inaccuracies in MTCNN’s eye position regression at low image resolutions. The method starts by segmenting the eyebrow area based on face detection and pupil positioning from the MTCNN network. It then calculates the gray gradient integral projection horizontally and vertically within the eyebrow area. By combining the MTCNN key points with these projections, the algorithm accurately locates the pupil. Experimental results demonstrate a detection accuracy of 95.02% and strong robustness for eye detection across various image qualities.

III. Methodology

The proposed methodology involves the implementation and integration of both speech recognition and gesture recognition systems into a unified framework, which is developed as a comprehensive website. The website provides a framework for users to interact with the speech recognition system through their system microphone, while gesture detection is achieved via the system's webcam.

The overall functionality and workflow of this integrated system are visually represented through a block diagram, as shown in Fig. 1. This diagram illustrates the flow of data from the user's input devices (mi-crophone and webcam) to the respective recognition modules, and finally to the user interface, where the processed data is used to generate real-time feedback.

This methodology ensures a seamless inter-action experience, leveraging the combined capabilities of speech and gesture recognition to create a versatile and user-friendly system.

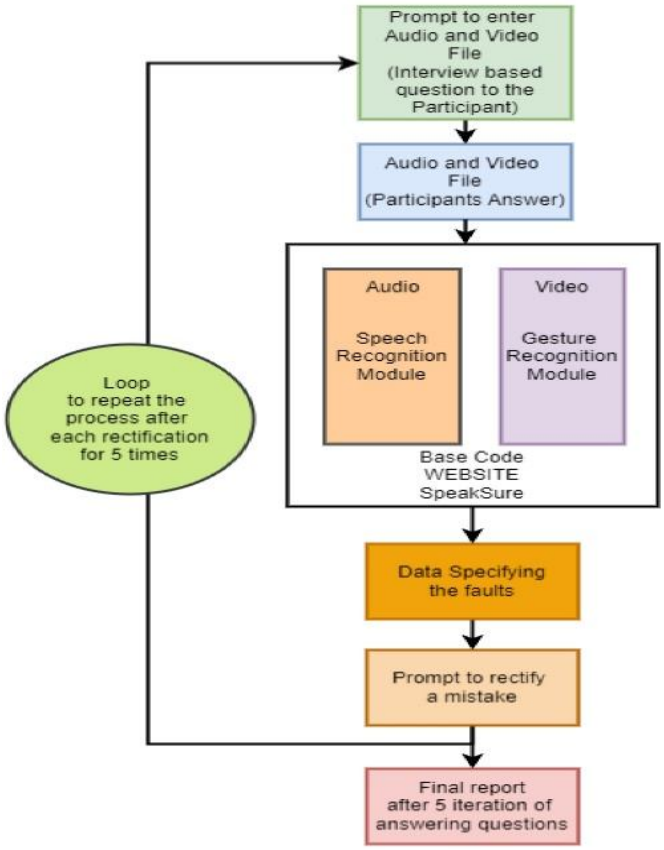

Fig.1 Block diagram of communication system

A. Proposed Modules

The proposed modules include a speech recognition module, a gesture recognition module, and a web-site that serves as the integrator module.

The speech recognition process is depicted in a flowchart in Fig. 2. In this process, input from the mi-crophone is used to make an API call that converts speech to text. The resulting data is then sent back to the module to provide appropriate feedback.

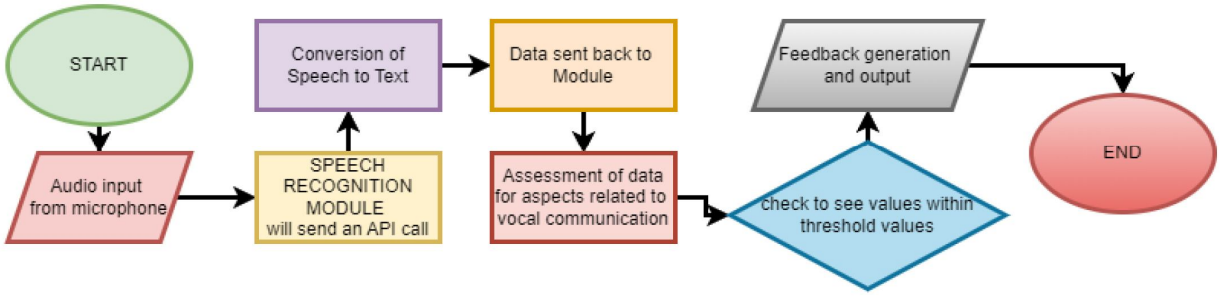

Fig.2 Flowchart of Speech recognition module

In the gesture recognition module, the video recognition module utilizes TensorFlow, a machine learn-ing component with a pre-trained dataset.

Tensor flow is trained using a web-based tool, Teachable Machine. It provides an easy way to train models for tasks like image recognition by creating datasets in the form of different classes for easy recognition of various features. The flowchart of the working of this module is represented in Fig.3 and the results are represented as varied numerical values.

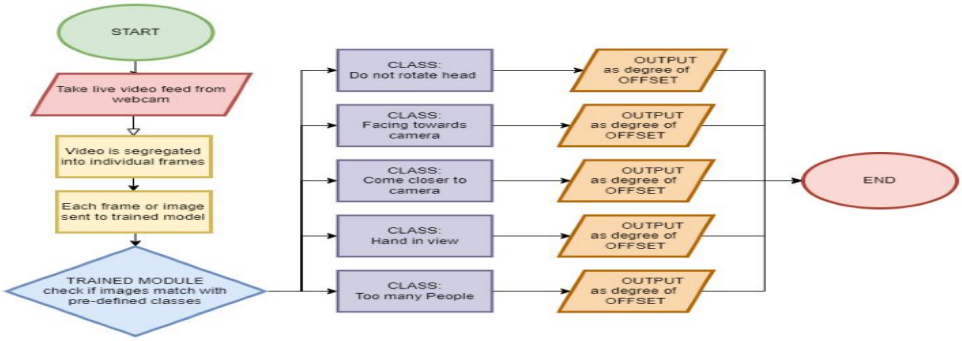

Fig.3 Flowchart of Gesture recognition module

The integrated module has been developed in the form of a user-friendly website. This website features a dedicated page with three key attributes: audio data, video data, and a question-and-answer section. Each attribute allows users to interact with the content easily, making the module accessible and intuitive for all users. The system characteristics is shown as a flowchart in Fig.4

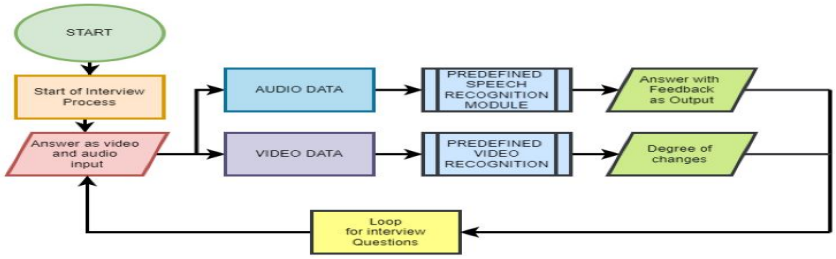

Fig.4 Flowchart of Integrator module

B. Dataset Collection

This paper presents a training module developed using TensorFlow, a widely used framework for image classification, object detection, and video analysis. The model is trained using Google's Teachable Machine, a web-based tool designed to simplify the creation of machine learning models. Teachable Machine offers an intuitive drag-and-drop interface that allows users to train models for tasks such as image recognition, sound classification, and pose detection. The process consists of three main steps: first, users collect data by uploading files or capturing input directly through their device's camera or microphone to represent different categories; next, the tool trains a machine learning model on this data in real-time; finally, the trained model can be exported and deployed in various applications, in-cluding integration into websites, mobile apps, or physical devices using platforms such as Tensor-Flow.js.

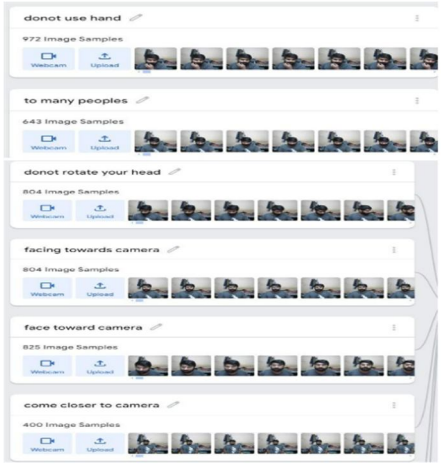

Fig.5 Datasets created at Teachable Machine Platform

The datasets created for the proposed system are shown in Fig. 5. Considering the different features to be recognized by the system, various classes were created and trained for specific gestures such as “Do Not Use Hand,” “Do Not Rotate Head,” “Face Towards Camera,” “Come Closer to Camera,” “Many People Detected,” and “Facing Towards Camera.” The selection of these features was informed by the specific requirements of an interview setting, where maintaining eye contact and appropriate body language such as refraining from touching the face or looking to the side are essential. For each fea-ture, 800-900 datasets were generated to achieve a high level of accuracy in the system, ultimately im-proving the user experience by ensuring that the results are accessible and easy to interpret.

C. Data Preprocessing

The data processing in the proposed system involves several key components:
Image Classification with TensorFlow.js: The process starts with loading and initializing a pre-trained image classification model from Teachable Machine via a URL. The webcam is set up to display the live feed in a <canvas> element. During real-time processing, the webcam feed is continuously cap-tured and analyzed using the loop() function, with predictions dynamically updated on the page. The webcam feed can be stopped and cleared using the stop() function.
Question Display: The system includes a dynamic question updating feature, where a predefined list of questions is stored. Each time the user clicks the "next" button, the question displayed on the page up-dates, cycling through the list.
Speech Recording and Processing: Audio recording is triggered by the "Answer" button, using the Mic Recorder library to capture audio. This file is sent to an API for speech-to-text conversion, and the re-sulting transcription is displayed on the page.

D. Implementation of the Proposed Model

A dedicated website has been developed to run both speech and gesture recognition systems locally, ensuring real-time processing with minimal latency. The home page, shown in Fig.5, features three main sections: video/gesture recognition, audio recognition, and a question-and-answer interface for user interaction. The backend integrates web development frameworks and machine learning libraries to support these tasks. The speech recognition system uses a specialized algorithm and a speech-to-text API to process audio inputs, detect filler words, and evaluate speech fluency, providing timely feedback. The gesture recognition system, powered by a TensorFlow model trained on a large dataset, performs real-time inference to accurately interpret gestures. The web application serves as a crucial link between the user interface and server-side processing, facilitating smooth interaction and efficient data handling. This architecture ensures that both speech and gesture recognition systems function seamlessly and effectively, aiming to deliver precise results within a user-friendly environment.

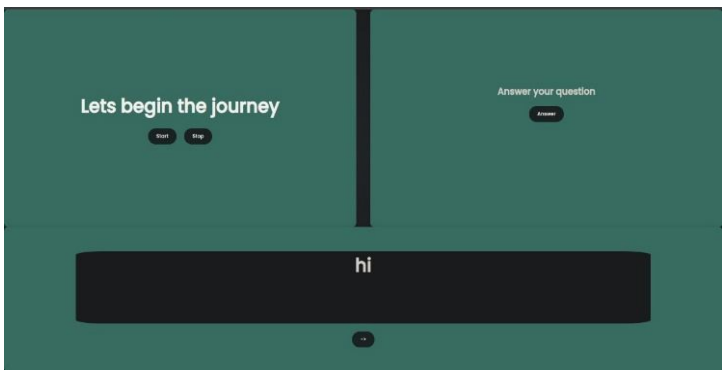

Fig.6 Home page of the website

IV. RESULTS

The data flow in the proposed system starts with the user activating the webcam and optionally record-ing their voice in response to a question. During data collection, images are captured and analyzed in real-time by TensorFlow.js, while audio is recorded and sent to an API for transcription. In the data processing stage, image predictions and text transcriptions are generated. Finally, results are displayed on the webpage, showing real-time image predictions and transcribed text with filler word analysis, ensuring smooth and effective user interaction and feedback.

A. Audio Implementation

Implementing a speech recognition function, as shown in Fig.7, that processes uploaded audio data, converts it to text using API

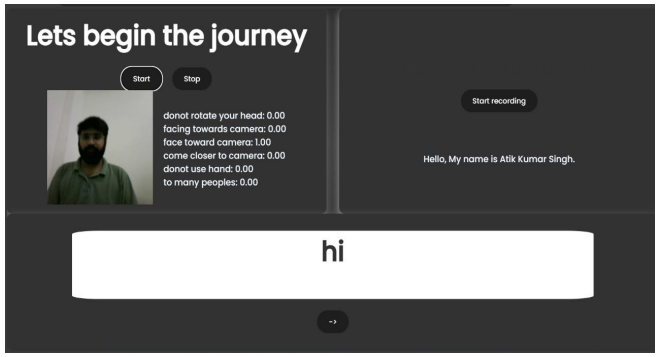

Fig.7 Audio analysis

B. Video Implementation

For gesture detection, a TensorFlow model was trained using Teachable Machine, where various clas-ses were created to represent different features. During the implementation of this module, we ob-served that the values sometimes deviated from the set standard. When such deviations occur, they indicate to the user that their posture is incorrect, providing an opportunity for correction. This func-tionality is illustrated in Fig. 8, which clearly show instances of incorrect posture and the correspond-ing deviations in values.

The website also features a loop of questions as shown in Fig.9 designed to simulate a professional interview experience, allowing users to practice their responses while simultaneously monitoring their gestures. The pre-set questions include prompts such as “What are your strengths?”, “Why are you interested in this job?”, and “Where do you see yourself in five years?” Users also have the option to create their own custom questions, enabling them to tailor the practice session to their specific needs and scenarios. This dual functionality helps users refine both their verbal answers and non-verbal communication skills.

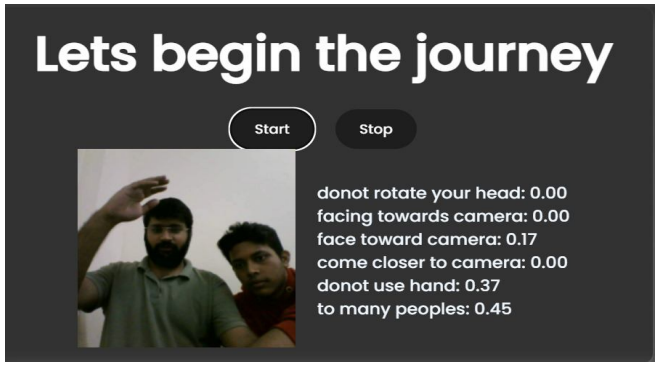

Fig.8 Gesture detection of “do not use your hand” and “many people present” feature

Fig.9 Questions Section

V. Conclusion

This paper details the development of the "Speak-Sure Interview Assistant" project, which focuses on methods for speech and gesture detection through preprocessing and machine learning models, while also integrating API configurations and user interface design using HTML, CSS, and JavaScript. The system design involves creating a website with comprehensive blueprints for key pages (home, video upload, live webcam), backend integration, and gesture detection through machine learning.

The overall aim of the project is to integrate gesture and speech recognition technologies to enhance communication skills, demonstrating the benefits of technological advancements. The system's archi-tecture outlines the interactions between the user interface, web server, application logic, and data management, utilizing appropriate tools and technologies.

Issues and Future Scope

The communication analysis modules currently require both formatting and training to deliver detailed and effective user feedback. Identifying key values and data essential for robust communication anal-ysis is a critical aspect of this process. Additionally, the integrator module needs further development to ensure it can consolidate data and generate accurate, actionable feedback smoothly. The main code driving the continuous processing loop also requires refinement to enhance efficiency, support real-time processing, and integrate advanced algorithms for more precise feedback.

In the future, optimizing performance will involve running the program on the onboard computer (Jet-son Nano) with upgraded camera and microphone modules. This enhancement will improve input qual-ity and allow for a larger dataset, ultimately increasing the system's accuracy and effectiveness.

References

[1] Kooragama K.G.C.M, Jayashanka L.R.W.D, Munasinghe J.A, Jayawardana K.W, Muditha Tissera, Thilini Bud-dhika, "Speech Master: Natural Language Processing and Deep Learning Approach for Automated Speech Evaluation", 2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Confer-ence (IEMCON)

[2] A. Deshpande, R. Pandharkar and S. Deolekar, "Speech Coach: A framework to evaluate and improve speech delivery," 2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP), 2020.

[3] Zheng Zhang*, "Construction of English Teaching System Based on Deep Learning Models and Artificial Intelligence Technology", 2023 2nd International Conference on Artificial Intelligence and Computer Information Technology (AICIT).

[4] Amitrajit Sarkar, Surajit Dasgupta, Sudip Kumar Naskar, Sivaji Bandyopadhyay, "Says Who? Deep Learning Models for Joint Speech Recognition, Segmentation And Diarization", 978-1-5386-4658- 8/18/$31.00 ©2018 IEEE— ICASSP 2018.

[5] S. Das, N. Gandhi, T. Naik and R. Shilkrot, "Increase Apparent Public Speaking Fluency by Speech Augmenta-tion," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 6890-6894, doi: 10.1109/ICASSP.2019.8682937.

[6] Yao Qian, Ximo Bian,Yu Shi, Naoyuki Kanda, Leo Shen, Zhen Xiao and Michael Zeng, "Speech Language Pre-Training For End-To-End Spoken Language Understanding", ICASSP 2021 - 2021 IEEE International Confer-ence on Acoustics, Speech and Signal Processing (ICASSP) — 978-1-7281-7605-5/20/$31.00 ©2021 IEEE DOI:10.1109/ ICASSP39728.2021.9414900.

[7] Jhuma Sunuwar, Samarjeet Borah, Ratika Pradhan, "Gesture Recognition Approaches and its Applicability: A Study", Fourth International Conference on Electronics, Communication and Aerospace Technology (ICECA-2020), IEEE Xplore Part Number: CFP20J88- ART; ISBN: 978-1- 7281-6387-1.

[8] Jen-Hang Wang, Yen-Hua Chen, Shao-Yun Yu, Yu-Ling Huang, Gwo-Dong Chen, "Digital Learning Theater with Automatic Instant Assessment of Body Language and Oral Language Learning", 2020 IEEE 20th Interna-tional Conference on Advanced Learning Technologies (ICALT) 2161-377X/20/$31.00 ©2020 IEEE,DOI 10.1109/ICALT.

[9] Ainampudi Kumari Sirivarshitha, Kadavakollu Sravani, Kothamasu Santhi Priya, Vasantha Bhavani: "An ap-proach for Face Detection and Face Recognition using OpenCV and Face Recognition Libraries in Py-thon",2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS) 979-8-3503-9737-6/23/$31.00 © 2023 IEEE.

[10] Xuefeng Chen, Xin Luo, Xingyao Liu, Jikang Fang: "Eyes Localization Algorithm Based on Prior MTCNN Face Detection",2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC 2019), 978-1-5386-8178- 7/19/$31.00 ©2019 IEEE.

[11] D. Sunitha, Vidya C Patil, Manjula H N, Sheba Jebakani, "Digital notice board using Smart Phones- Speech Recognition Voice command", Proceeding of 2018 IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore, India.

[12] Phyu Myo Thwe, Nwe Ni Kyaw, Moe Moe Htay, G R Sihna, Phyu Phyu Khaing, “A Critical Survey On Static And Dynamic Hand Gesture Recognition”, 2019 Joint International Conference on Sci-ence, Technology and Innovation, Mandalay by IEEE.
[13] Umberto Maniscalco And Antonio Messina And Pietro Storniolo, “A Hybrid Model For Gesture Recognition And Speech Synchronization”, 2023 Ieee International Conference On Advanced Ro-botics And Its Social Impacts (Arso) Berlin, Germany, June 5-7, 2023.
[14] S. V. Tathe, A. S. Narote, S. P. Narote: “Human Face Detection and Recognition in Videos”, 2016 Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI), Sept. 21-24, 2016, Jaipur, India, 978-1-5090-2029-4/16/$31.00 @2016 IEEE.
[15] Ning Zhang, Junmin Luo, Wuqi Gao. “Research on Face Detection Technology Based on MTCNN”, 2020 International Conference on Computer Network, Electronic and Automation (ICCNEA),978-1-7281-7083-1/20/$31.00 ©2020 IEEE DOI 10.1109 / ICCNEA 50255.2020.000404