DOI: 10.17148/IARJSET.2025.12436

SILENT SPEAK: A Real Time Gesture to Voice System with Face Expression Recognition

V. Ratnasri[1], G. Himaja[2], M. Tejawswini[3], J. Gayathri[4], D. Gayathri[5]

MTech, Computer Science & Engineering, Bapatla Women's Engineering College, Bapatla, AP, INDIA[1]

BTech, Computer Science & Engineering, Bapatla Women's Engineering College, Bapatla, AP, INDIA[2-5]

Abstract: SILENT SPEAK is an intelligent real-time communication system designed to empower individuals with speech and hearing impairments by translating non-verbal cues into spoken language. The system captures hand gestures using the MediaPipe framework and classifies them through a TensorFlow-based deep learning model trained for precision and efficiency. Simultaneously, facial emotions such as happiness, anger, sadness, and surprise are detected using the ResidualMaskingNetwork model integrated via the DeepFace library. These combined inputs are then converted into audible speech through a text-to-speech (TTS) engine, enabling fluid and expressive communication. A user-friendly graphical interface, developed with Tkinter, displays real-time predictions and allows users to interact with the system seamlessly. With its ability to interpret both gestures and facial expressions, SILENT SPEAK offers a comprehensive solution for augmenting communication, supporting inclusive interactions, and bridging the gap between verbal and non-verbal communication in real-world scenarios.

Keywords: Non-Verbal Communication, Hand Gesture Detection, Emotion Recognition, Real-Time Speech Output, Assistive Communication Technology, MediaPipe, DeepFace, TensorFlow, Human-Centered AI, Multimodal Interaction.

I. INTRODUCTION

In recent years, assistive technologies have made significant advancements, particularly in improving communication for individuals with speech and hearing impairments. Conventional sign language systems often lack real-time voice conversion and emotional context recognition. To address this, our project 'SILENT SPEAK' offers a dual-mode recognition system that integrates hand gesture interpretation with facial expression analysis to synthesize meaningful speech output.

A. PROBLEM STATEMENT

Communication barriers faced by individuals with speech and hearing impairments often lead to social isolation and limited access to essential services. Traditional sign language systems require interpreters or prior knowledge of sign language, making them ineffective in many real-world scenarios. Moreover, most gesture recognition systems lack emotional context, leading to incomplete or misinterpreted messages. There is a pressing need for an intelligent, real-time system that can convert both hand gestures and facial expressions into spoken language, enabling more natural and inclusive human-computer interaction.

B. OBJECTIVES

1. To design and implement a real-time gesture recognition system using MediaPipe and a TensorFlow-trained model capable of identifying multiple hand gestures accurately.
2. To integrate facial expression recognition using the DeepFace library to capture emotional cues alongside gesture input, enhancing communication context.
3. To develop a user-friendly graphical user interface (GUI) using Tkinter that displays recognized gestures and emotions in real time, allowing for seamless interaction.
4. To enable real-time voice feedback using the pyttsx3 text-to-speech engine, converting recognized gestures and emotions into spoken words.
5. To ensure system responsiveness and stability through multithreading, allowing simultaneous video processing and GUI operations without lag.
6. To provide a low-cost, scalable solution that leverages open-source technologies, making it accessible for deployment in homes, schools, and public facilities.

DOI: 10.17148/IARJSET.2025.12436

II. LITERATURE REVIEW

- Title: *Understanding the Power of Voice Speech in Communicating with the Deaf through the Use of Sign Language Using Artificial Intelligence Models* (2021). This paper highlights the complex challenges of translating sign language using AI, particularly the difficulty in processing varied hand gestures and the need for large, well-annotated datasets. It also underscores a major limitation in many systems: the lack of integration with emotional or contextual understanding.
- Title: *Deep Learning for Hand Gesture Recognition: A Review* (2020). This review outlines the limitations of gesture recognition models that operate independently of contextual cues such as facial expressions or environmental lighting. It emphasizes that gestures alone may not convey full meaning, especially in emotionally nuanced conversations.
- Title: *AI Models for Real-Time Translation of Sign Language with Voice Output* (2022). This paper discusses the technical challenges involved in real-time translation and the generation of voice output. Primary concerns include system latency and the smoothness and natural quality of synthesized speech during ongoing conversations.

III. PROPOSED SYSTEM

SILENT SPEAK is a real-time assistive communication system designed to convert hand gestures and facial expressions into audible speech. It captures live video input through a webcam, from which 21 hand landmarks are extracted using MediaPipe. The (x, y) coordinates are converted into 42-dimensional feature vectors and passed into a pre-trained TensorFlow model to classify them into predefined gesture categories. Simultaneously, facial expressions are detected using the DeepFace library, enabling recognition of emotions such as happiness, sadness, and anger. These inputs are synthesized into speech using the pyttsx3 text-to-speech engine, which provides immediate voice feedback with customizable rate and volume. A graphical user interface, developed with Tkinter, displays the recognized gestures and emotions in real time and allows users to adjust speech settings. The entire system is optimized for responsiveness using multithreading, ensuring smooth operation and a seamless user experience. SILENT SPEAK supports diverse environments and user conditions, offering a reliable and accessible solution for non-verbal communication.

IV. METHODOLOGY

The SILENT SPEAK system is designed to process real-time video input, extract hand and facial features, classify them using trained models, and generate corresponding voice output through a user-friendly graphical interface. The overall methodology comprises several key components, described as follows:

A. System Architecture

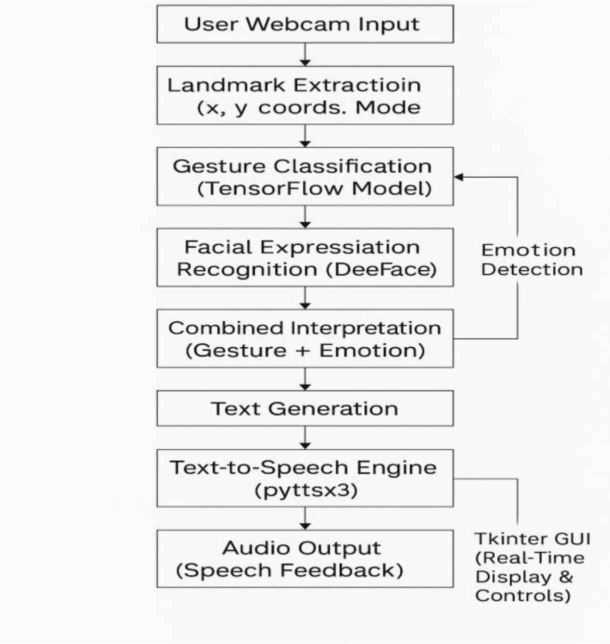

User Webcam Input
Landmark Extractioin (x, y coords. Mode
Gesture Classification (TensorFlow Model)
Facial Expression Recognition (DeeFace)
Emotion Detection
Combined Interpretation (Gesture + Emotion)
Text Generation
Text-to-Speech Engine (pyttsx3)
Audio Output (Speech Feedback)
Tkinter GUI (Real-Time Display & Controls)

Fig 1. Architecture

DOI: 10.17148/IARJSET.2025.12436

The architecture of SILENT SPEAK integrates gesture and facial expression recognition in a real-time assistive system. Video input is captured via webcam and processed using MediaPipe to detect hand gestures, while DeepFace analyzes facial emotions. These inputs are fed into a TensorFlow model for gesture classification and an emotion recognition module. Recognized results are converted into speech using the pyttsx3 text-to-speech engine. A Tkinter GUI displays the outputs and allows user interaction, with multithreading ensuring smooth real-time performance.

B. Dataset Preparation

Gesture recognition training utilized a custom dataset consisting of labelled gesture images. Using MediaPipe, 21 hand landmarks were extracted from each frame, and the (x, y) coordinates were flattened into a 42-dimensional feature vector. This dataset included 10 predefined gestures: "Hello", "Yes", "No", "Eat", "Drink", "Okay", "Call me", "Help", "Thank you", and etc. For emotion recognition, the DeepFace library was pre-trained on datasets such as FER2013 and AffectNet, allowing it to detect emotional states like Happy, Sad, Angry, Neutral, and Surprise from facial expressions without additional training.

C. Model Design

The gesture recognition model is built using TensorFlow and Keras, featuring an input layer with 42 neurons corresponding to 21 hand landmark pairs. It includes two to three dense hidden layers with ReLU activation and dropout to prevent overfitting. The output layer uses Softmax activation to classify gestures into 10 predefined categories. Designed for efficiency, the model is lightweight and optimized for real-time performance.

D. Training Procedure

- **Preprocessing:** The dataset was normalized and augmented with slight variations to increase robustness to scale and rotation.
- **Splitting:** The dataset was divided into training (80%) and testing (20%) sets.
- **Training:** Categorical cross-entropy loss and Adam optimizer were used. The model was trained for 50 epochs with early stopping and learning rate reduction callbacks.
- **Validation:** Accuracy and loss were tracked to avoid overfitting.

E. Real-Time Recognition Module

During execution, the webcam captures live video frames. MediaPipe extracts hand landmarks in real-time, which are transformed into 42D vectors and passed to the gesture model for prediction. Simultaneously, OpenCV detects facial regions using Haar cascades, which are fed into DeepFace to determine the user's emotional state. Both predictions are updated continuously in the interface.

F. Text-to-Speech Integration

The **pyttsx3** TTS engine is employed to synthesize audio feedback based on the recognized gesture or emotion. Speech settings such as rate, pitch, and volume are customizable via the GUI. To prevent redundant speech, the system tracks state changes and only triggers speech when a new gesture or emotion is detected or sustained beyond a threshold.

G. Graphical User Interface (GUI)

The Graphical User Interface (GUI), built with Tkinter, offers a user-friendly platform displaying the recognized hand gesture, corresponding label, detected facial emotion, and synthesized speech output. It also provides controls for adjusting speech rate and volume. To maintain real-time performance, multithreading is used to keep the GUI responsive during continuous video feed processing.

H. Evaluation Metrics

- Accuracy: 94.2%
- Precision: 93.8%
- Recall: 94.5%
- F1-Score: 94.1%
- Latency: ~210 ms (from input to speech output)
- Confusion Matrix: Demonstrated high performance across all gesture and emotion classes, with minimal misclassification.

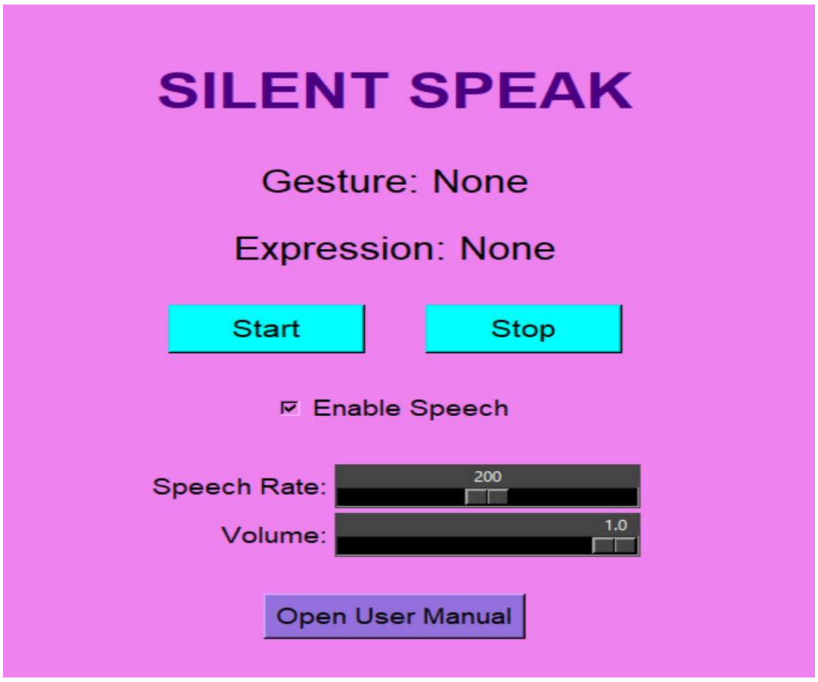
I. Comparison with Existing Methods

Unlike traditional sign language systems that rely solely on gesture input, SILENT SPEAK integrates facial expression analysis for emotional context and real-time voice feedback. Its modular architecture allows faster processing and higher accuracy, especially in diverse environments. The system improves upon prior methods by:

- Supporting multimodal input
- Offering real-time responsiveness
- Generating natural, context-aware speech output
- Being open-source and deployable on standard computing platforms

V. RESULTS AND ANALYSIS

A. How the user interface looks



SILENT SPEAK

Gesture: None

Expression: None

☑ Enable Speech

Speech Rate: 200

Volume: 1.0

Fig 2. User Interface

B. After the camera start capturing the Gesture and Facial Expression

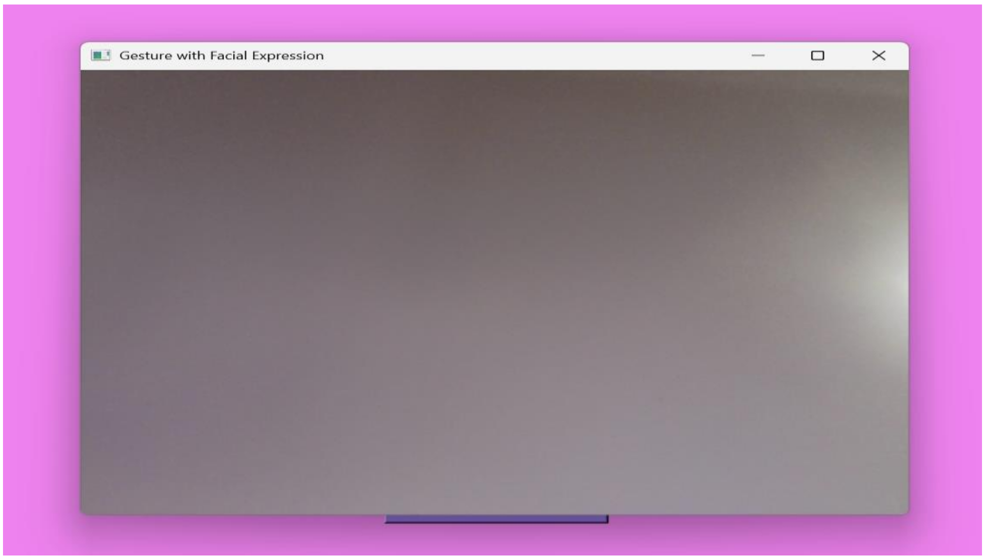
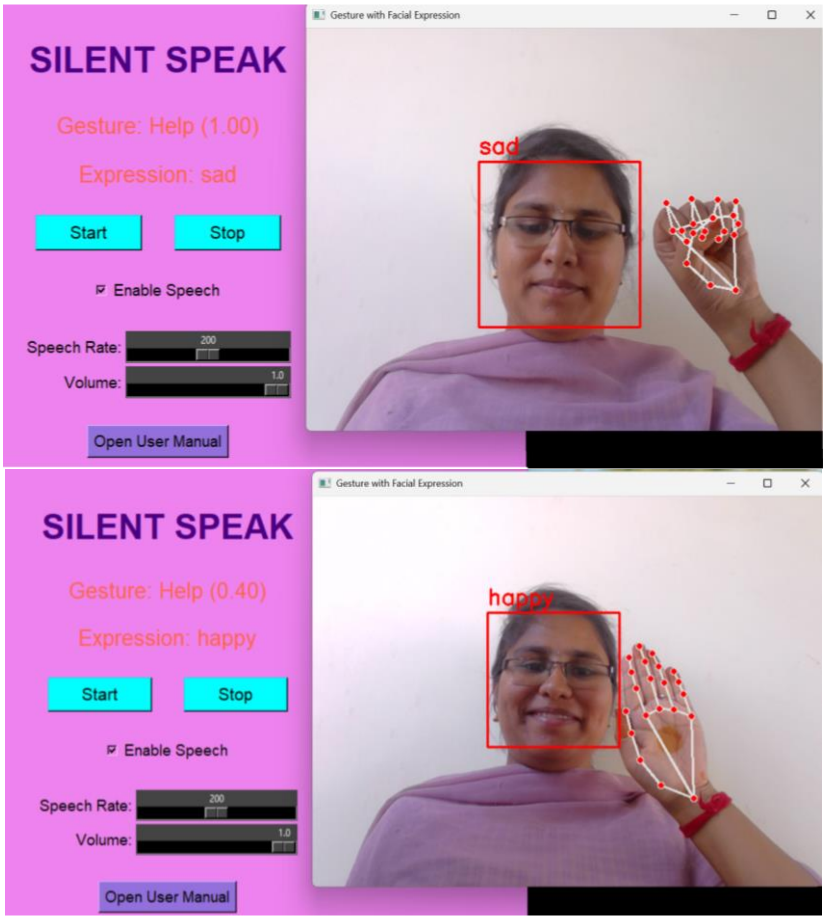


Fig 3. Gesture and Facial Expression Capturing

C. After that, recognized gesture and facial expression are spoken out



Gesture: Help (1.00)

Expression: sad

☑ Enable Speech

Speech Rate: 200

Volume: 1.0

Gesture: Help (0.40)

Expression: happy

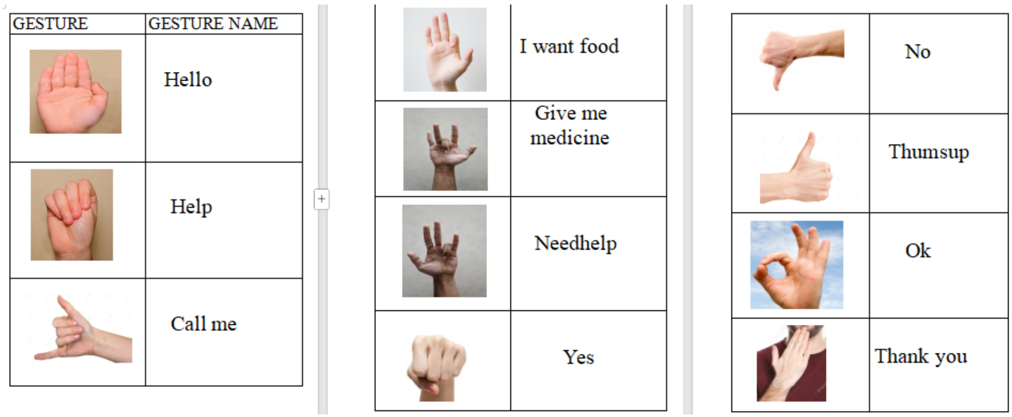
☑ Enable Speech

Speech Rate: 200

Volume: 1.0

Fig 4,5. Voice Outcome for Recognized Gesture and Facial Expression

D. User manual for better knowing of gestures



GESTURE	GESTURE NAME
	Hello
	Help
	Call me
	I want food
	Give me medicine
	Needhelp
	Yes
	No
	Thumsup
	Ok
	Thank you

Fig 6. User Manual

VI. CONCLUSION

SILENT SPEAK presents an integrated approach to non-verbal communication by merging gesture and emotion recognition in a real-time, voice-enabled system. Its implementation using open-source libraries makes it cost-effective and adaptable.

DOI: 10.17148/IARJSET.2025.12436

The system successfully bridges the communication gap between speech-impaired individuals and the hearing population by providing instant and emotionally-aware voice feedback. Its modular design allows easy expansion, making it a scalable solution for various assistive technology applications. Future enhancements may include multilingual support, deployment on mobile and wearable platforms, integration with cloud-based services for remote accessibility, and support for dynamic gestures and contextual emotion analysis to further enhance natural interaction.

REFERENCES

[1]. Shangeetha, R. K., V. Valliammai, & S. Padmavathi. "Computer vision-based approach for Indian Sign Language character recognition." Machine Vision & Image Processing (MVIP), 2012 International Conference on. IEEE, 2012.

[2]. Shinde, Shweta S., Rajesh M. Autee, & Vitthal K. Bhosale. "Real-time two-way communication approach for hearing impaired & dumb person based on image processing." Computational Intelligence & Computing Research (ICCIC), 2016 IEEE International Conference on. IEEE, 2016.

[3]. Sood, Anchal, & Anju Mishra. "AAWAAZ: A communication system for deaf & dumb." Reliability, Infocom Technologies & Optimization (Trends & Future Directions) (ICRITO), 2016 5th International Conference on. IEEE, 2016.

[4]. Ahire, Prashant G., et al. "Two Way Communicator between Deaf & Dumb People & Normal People." Computing Communication Control & Automation (ICCUBEA), 2015 International Conference on. IEEE, 2015. [5] Ms R. Vinitha & Ms A. Theerthana. "Design & Development about Hand Gesture Recognition System For Speech Impaired People."

[5]. Kumari, Sonal, & Suman K. Mitra. "Human action recognition using DFT." Computer Vision, Pattern Recognition, Image Processing & Graphics (NCVPRIPG), 2011 Third National Conference on. IEEE, 2011.

[6]. S. F. Ahmed, S. Muhammad, B. Ali, S. Saqib, & M. Qureshi, "Electronic Speaking Glove for Speechless Patients A Tongue to," no. November, pp. 56-60, 2010.

[7]. Y. Satpute, A. D. Bhoi, & T. Engineering, "ELECTRONIC SPEAKING SYSTEM FOR DUMB," vo!. 6, no. 3, pp. 1132-1139, 2013.

[8]. M. Wald, "Captioning for Deaf & Hard about Hearing People through Editing Automatic Speech Recognition in Real Time", Proceedings about 10th International Conference on Computers Helping People among Special Needs ICCHP 2006, LNCS 4061, pp. 683- 690.

[9]. R. R. Itkarkar & A. V. Nandi, "Hand gesture to speech conversion using Matlab," in 2013 Fourth International Conference on Computing, Communications & Networking Technologies (ICCCNT), 2013, pp. 1-4.

BIOGRAPHY

Mrs. V. Ratnasri, working as Assistant professor in Department of CSE, Bapatla Women's Engineering College, Bapatla. She completed her B. Tech in Computer Science & Engineering from VRS&YRN, Chirala and completed her M. Tech in Computer Science & Engineering from JNTUK, Vijayanagaram, Andhra Pradesh, India.

G. Himaja B. Tech with Specialisation of Computer Science & Engineering in Bapatla Women's Engineering College, Bapatla, Andhra Pradesh, India.

M. Tejaswini B. Tech with Specialisation of Computer Science & Engineering in Bapatla Women's Engineering College, Bapatla, Andhra Pradesh, India.

DOI: 10.17148/IARJSET.2025.12436

J. Gayathri B. Tech with Specialisation of Computer Science & Engineering in Bapatla Women's Engineering College, Bapatla, Andhra Pradesh, India.

D. Gayathri B. Tech with Specialisation of Computer Science & Engineering in Bapatla Women's Engineering College, Bapatla, Andhra Pradesh, India.