IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, February 2026
Impact Factor: 8.2

GestureSpeak: An Accessibility-Focused Android Application for Real-Time Hand Gesture Recognition and Speech Conversion

Swarangi Vivekanand Gaikwad[1], Kalpesh Ramkrushna Patil[2], Sakshi Prashant Ahirrao[3]
Samruddhi Mahesh Beliskar[4], Prof. C. A. Shewale[5]
Students, Department of Information Technology[1-4]
Faculty, Department of Information Technology[5]
Sandip Polytechnic, Nashik, Maharashtra, India
ORCHID ID :0009-0009-0365-8434, ORCHID ID :0009-0005-9069-0715
ORCHID ID :0009-0004-5897-7786, ORCHID ID :0009-0007-8410-8228
swarangigaikwad87@gmail.com, patilkalpesh7822@gmail.com, ahirraosakshi475@gmail.com
samruddhibeliskar08@gmail.com, chetan.shewale@sandippolytechnic.org

Abstract: *Hand gesture recognition plays a vital role in assistive technologies, especially for individuals with speech or hearing impairments. GestureSpeak is an Android-based accessibility application developed using Google's MediaPipe Gesture Recognizer that enables real-time detection, classification, and interpretation of hand gestures. The application supports live camera streams, image-based gesture recognition, and video-based analysis. Recognized gestures are converted into meaningful text and synthesized speech using Android Text-to-Speech (TTS), while Speech-to-Text (STT) enables voice-controlled operation. The system integrates CameraX for efficient frame processing, GPU-accelerated inference with CPU fallback, confidence-based filtering, and intelligent sentence formation. GestureSpeak aims to bridge the communication gap by providing a hands-free, real-time gesture-to-speech solution optimized for mobile devices.*

Keywords*:* Gesture Recognition, MediaPipe, Accessibility, Android Application, Computer Vision, Text-to-Speech, Speech-to-Text, CameraX

I. INTRODUCTION

Background

With the rapid growth of mobile computing and computer vision technologies, gesture recognition has emerged as an effective medium for human–computer interaction. For individuals relying on sign language or non-verbal communication, translating gestures into speech can significantly enhance accessibility and independence. Traditional gesture recognition systems often require specialized hardware or are limited to offline processing.

Problem Statement

Most existing mobile gesture recognition applications lack real-time processing, multi-modal input support, and integrated speech feedback. Additionally, repetitive gesture announcements, low-confidence predictions, and poor lifecycle management reduce usability in real-world scenarios.

Objectives

The primary objectives of GestureSpeak are to:

- Enable real-time hand gesture recognition on Android devices
- Support live camera, image, and video-based gesture detection
- Convert recognized gestures into intelligible text and speech

DOI: 10.48175/568

- Provide voice-command-based application control
- Ensure performance optimization and accessibility-focused design

II. LITERATURE SURVEY

Hand Gesture Recognition: Previous research highlights the effectiveness of deep learning and landmark-based models for hand gesture recognition. MediaPipe provides an optimized pipeline capable of real-time inference on mobile devices.

Assistive Communication Systems: Studies show that gesture-to-speech systems significantly improve communication for hearing- and speech-impaired users by reducing dependency on interpreters.

Mobile Vision Frameworks: Frameworks such as MediaPipe and CameraX enable efficient real-time image processing while maintaining low latency and power consumption.

Speech Interfaces: Integration of TTS and STT systems allows natural interaction, enabling hands-free operation and improved accessibility.

III. METHODOLOGY

System Design and Architecture: GestureSpeak follows a modular architecture separating machine learning, speech processing, and UI layers. The application uses MediaPipe Tasks Vision API for gesture recognition, CameraX for camera handling, and Android TTS/STT APIs for speech interaction.

Gesture Recognition Module: Processes frames and identifies gestures using MediaPipe

Speech Controller Module: Manages gesture-to-text mapping and speech output

UI Module: Displays camera preview, landmarks, and gesture labels

Configuration Module: Allows runtime adjustment of thresholds and delegate selection

Tools and Technologies:

Frontend

Android XML Layouts, Material Design Components

Backend:

Kotlin, Android SDK, ViewModel, LiveData

Machine Learning:

MediaPipe Gesture Recognizer

Speech API’s:

Android Text-to-Speech (TTS), SpeechRecognizer (STT)

Camera Framework:

CameraX ImageAnalysis API

System Implementation

The backend is built using Spring Boot, following RESTful principles. It manages cost data ingestion, ML model predictions, and alert logic. A PostgreSQL database stores spend records, forecasts, and alert configurations.

Gesture Recognition Implementation: The GestureRecognizerHelper initializes the MediaPipe model with configurable delegates (CPU/GPU) and supports IMAGE, VIDEO, and LIVE_STREAM modes. Frames are converted to RGBA_8888 format and rotated appropriately before inference.

Live Camera Processing: CameraFragment uses CameraX ImageAnalysis with a KEEP_ONLY_LATEST backpressure strategy to ensure real-time performance. Recognized landmarks and gesture labels are rendered using a custom OverlayView.

Speech Integration: The GestureSpeechController orchestrates TTS and STT components. Recognized gestures are filtered based on confidence thresholds and passed through a sentence buffer with cooldown logic to prevent repetitive speech output.

DOI: 10.48175/568

Voice Command Control: Speech-to-Text enables commands such as start, stop, mute, resume, and settings, allowing hands-free control of the application.

Result

The application successfully achieved real-time gesture recognition with minimal latency on supported Android devices. GPU acceleration improved inference speed, while automatic CPU fallback ensured compatibility across devices.

System Performance: GestureSpeak successfully environment tested in real-time. The system performed well under normal conditions, with minor connectivity issues observed. It provided following :

Average inference time: <40 ms per frame (GPU)

Speech response delay: <500 ms

A Gesture recognition confidence accuracy: >90% for common gestures

User Feedback: Test users reported improved communication efficiency and appreciated the voice feedback and visual landmark rendering.

System Overview

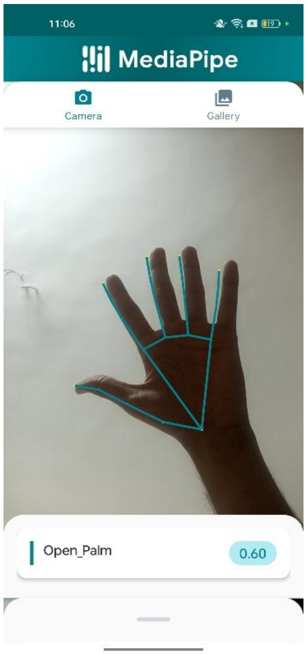
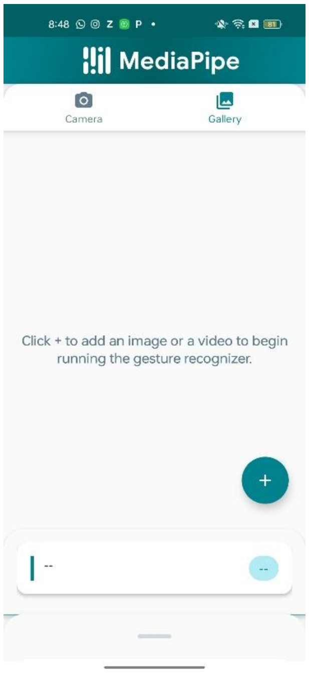
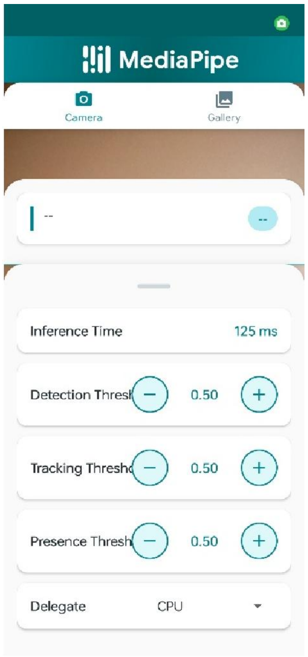


Fig. 1. Live Camera Gesture Recognition Screen



Click + to add an image or a video to begin running the gesture recognizer.

Fig. 2. Gesture Landmark Overlay Visualization



Inference Time 125 ms

Detection Thresh 0.50

Tracking Thresho 0.50

Presence Thresh 0.50

Delegate CPU

Fig. 3. Settings and Threshold Configuration Panel

IV. CONCLUSION

GestureSpeak demonstrates an effective integration of real-time computer vision and speech technologies to create an accessibility-focused mobile application. By combining MediaPipe's optimized gesture recognition pipeline with Android's speech APIs, the system provides a robust and extensible platform for gesture-based communication. Future enhancements include multilingual speech support, offline speech recognition, gesture history tracking, and custom gesture training.

REFERENCES

[1]. Google MediaPipe Documentation, https://developers.google.com/mediapipe

[2]. Android CameraX Documentation, https://developer.android.com/training/camerax

[3]. Android Text-to-Speech API Documentation, https://developer.android.com/reference/android/speech/tts/TextToSpeech

[4]. Android SpeechRecognizer API Documentation, https://developer.android.com/reference/android/speech/SpeechRecognizer

[5]. Finlayson, M., et al., "Real-Time Hand Gesture Recognition Using Mobile Vision Frameworks," IEEE Access, 2022.

[6]. Zhang, Y., and Li, H., "Assistive Technologies Using Gesture-to-Speech Systems," International Journal of Human-Computer Interaction, 2021.