Power System Protection and Control
ISSN:1674-3415

SPEECH-TO-TEXT CONVERSION WEARABLE SMART GLASSES FOR THE HEARING-IMPAIRED USING ESP 32

Dr. G. Nooka Raju[1], Dr. M. Sreedhar[2], Dr.PMK .Prasad[3]

[1]Sr.Asst Professor ,Department of ECE, GMR Institute of Technology, Rajam, 532127, India
E-mail: nookaraju.g@gmrit.edu..in
[2]Professor .Department of Electronics and Instrumentation, Vallurupalli Nageswararao Vignana Jyothi Institute of Engineering and Technology, Hyderabad, E-mail: sreedhar_m@vnrvjiet.in
[3]Professor and Head ,Department of ECE, GVP College of Engineering for Women, Visakhapatnam
E- mail: pmkp70@gmail.com

Received:25 February 2025 **Revised**:28 March 2025 **Accepted**: 22 April 2025

***ABSTRACT*:**

Hearing-impaired individuals face communication challenges, such as daily interactions with other people due to partial or complete loss of hearing, which shows impact on their quality of life in society. The assistive technologies, such as hearing aids, sign language and speech-to-text applications have significantly improved accessibility and enhanced the way of communication for those with hearing disabilities. Ongoing research aims to develop more effective solutions for better integration and inclusivity in society. This paper presents a device ESP-32, which will be the core microcontroller board in order to integrate Deepgram's API for the purposes of real-time speech recognition and conversion tasks as a hearing assistant. The device hosts an INMP441 digital microphone, which captures audio from the surroundings and converts it to a digital signal. The integrated 24-bit ADC converts signals analog to digital ones and then communicates to the ESP-32 microcontroller board using the I2S protocol. The ESP-32 then takes up the responsibility of processing digital audio data and transmitting it via Wi-Fi to the Deepgram API. On the other hand, the API transcribes the speech into text and sends it back to the ESP-32 microcontroller. This text is eventually displayed as subtitles on the glasses with the help of a TOLED connected to the ESP-32 through the I2C protocol. Consequently, this proposed innovation will enable users to communicate effectively in social, educational, or business scenarios, creating a shared space between people who can hear and the hearing impaired through the use of technology.

Keywords*: ESP32, I2S and I2C protocol,INMP441 microphone, Speech-to-Text, Deepgram API, IoT Audio Appli- cations ,Base64 Encoding.*

1. INTRODUCTION

Hearing-impaired people face many challenges in accessing spoken communication, which can affect their day-to-day life while communicating with other people. They often rely on alternative methods of com- munication, such as sign language or written text, to close the gap. This reliance can sometimes lead to misunderstandings or feelings of isolation, highlighting the importance of creating more inclusive environ- ments for everyone. They face challenges at many places like education, workplaces, and social environments. According to the World Health Organization's survey, over 430 million people globally have hearing loss, and this number is rise's day to day. Many people with hearing disabilities rely on assistive technologies like hearing aids and sign language, but not everyone understands sign language, which further limits the communication. These barriers affect their quality of life and show the need for accessible, real-time solutions that can bridge the communication gap and improve the quality of life for the hearing impaired.

Human speech is the most natural and effective form of communication to lead a daily conversations. The transformation of speech into text opens up a wide range of applications, from digital assistants to accessibility tools for the people with hearing loss. The development of embedded systems has made it possible to bring speech recognition and speech to text conversations into a small and affordable devices. This project aims to develop a speech- to-text system based on the ESP32 microcontroller, integrating real-time audio capture, cloud-based speech

DOI: 10.46121/pspc.53.2.18

recognition, and text display functionality into a compact and user-friendly solution. The system architecture's core component is the ESP32. It is a versatile, inexpensive microcontroller that integrates Wi-Fi and Bluetooth, making it fit for Internet of Things applications. The ESP32 is unique compared to other microcontrollers since it has dual-core computing and low power operation along with peripherals like ADC, I2S, SPI, and I2C interfaces. With these characteristics, it is capable of managing hardware control and data communication simultaneously in one device. For this project, the ESP32 gathers and processes the audio information, establishes secure communication lines with the cloud API for speech recognition, and utilizes the cloud-based services for speech- to-text transformation To capture speech the system uses an INMP441 digital MEMS microphone, which uses I2S communication protocol for transferring data to ESP32. The I2S interface provides a means of capturing a digital audio signal with minimal noise, which is essential for effective speech recognition. To further enhance the clarity of the captured speech, the microphone is position at the user's mouth to eliminate background noise. The microphone is directly connected to the ESP32 controller which reads the audio signal in real time and stores it in a buffer until it is sent to the cloud.

With regard to speech recognition, the project leverages the capabilities of the Deepgram API, a cutting- edge machine learning platform capable of performing audio-to-text transcription on a cloud server. The ESP32 requests connection to the API through a secured Wi-Fi link using the libraries WiFi.h and WiFi- ClientSecure.h. After sending the audio information to the API, the API processes it and sends back the relevant text. Incorporation of Deepgram into the system increases accuracy and speed of responses, thus enabling the system to be used in real-time applications.

By using the display, the user is no longer distracted as it functions as a heads-up interface as they can view captions or transcriptions displayed within their field of vision. The text is displayed on the T- OLED screen, which receives input from the microphone through the I2C interface, and it is constantly refreshed with the latest speech recognition results. Therefore this system can be used effectively during live conversations, presentations, or any scenario where immediate text feedback is required.

The system includes an SD card module which stores the audio captured by the INMP441 microphone. This module allows the ESP32 to store audio recordings and convert the mp3 file to Base64 format for their corresponding transcriptions as API only accepts files in Base64 format for speech to text transcription. The SD card can be accessed using the SD.h and SPI.h libraries, which facilitate file creation, reading, writing, and deletion. This feature enhances the usability of the device by supporting offline logging and troubleshooting. In terms of software, this project is developed on the Arduino IDE and makes use of specialized libraries for Wi-Fi, I2S audio capturing, OLED display management, and secure communication through HTTPS. The firmware implements every operation in system resource optimized manner within a single repeating loop: audio capturing, sending the audio to the API for transcription, receiving and displaying the text. The user experience on the OLED screen is further enhanced by features like text scrolling, contrast modification, and text formatting that improve text legibility.

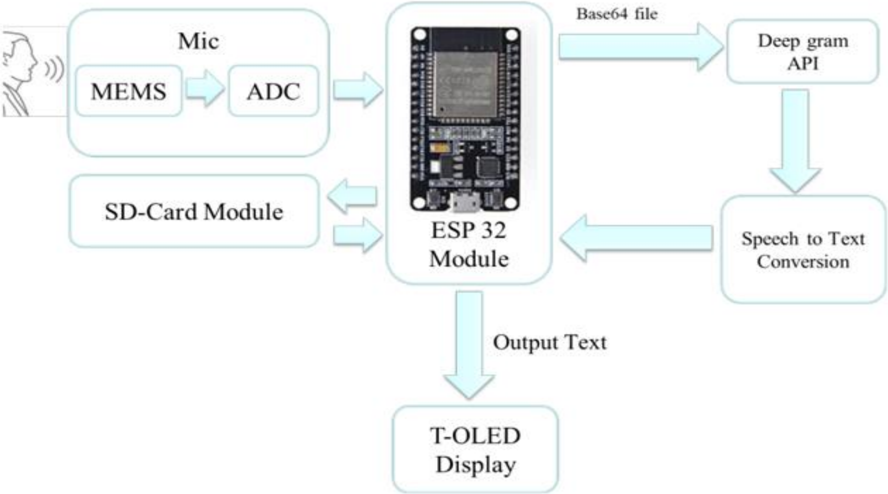

Figure 1. The figure represents the block diagram of speech-to-text conversion system using an ESP32 module and outlines the key components and flow of data with in the system.

DOI: 10.46121/pspc.53.2.18

2. PROPOSED SYSYEM MODEL

The proposed system model is smart glasses for speech-to-text conversion is mainly focused on achieving efficiency,warability,and real-time performance whule ensuring a light and user-friendly experience .The system integrates multiple hardware components within a compact frame, The key design considerations include form factor, component placement, power efficiency, user comfort, and display integration. The fig1 represents the block diagram represents a speech-to-text conversion system using an ESP32 module, showcasing the key components and their interactions. The system begins with a digital microphone, which captures speech signals from the user, and an ADC, which is integrated within the circuit and is used to convert these signals into a digital format, making them suitable for processing by the ESP32 microcontroller.

3. SYSTEM DESIGN

3.1. Design Considerations

3.1.1. Structural Design

2mmThese smart glasses are built for all day comfort ,with a lightweight and ergonomic design that won't strain the user.The frame is made from durable yet feather-light matrials like polycarbonate or carbon fiber,So they feel barely there,even after hours of wear.The weight is evenly distributed to prevent any discomfort,making them easy to wear for long periods.

To keep things sleek and non-bulky ,the ESP32 microcontroller and Li-ion battery are neatly tucked into the side arms of the frame instead of adding weight to the front.The transparent OLED(T-OLED) display sits in front of one eye,allowing users to read subtitles clearly without blocking their field of view.Meanwhile,the design minimal and unobtrusive.everything is designed with ease and comfort in mind,so users can go about their day without feeling weighed down.

3.1.2. Component Placement and Integration

The ESP32 microcontroller is embedded on one side of the frame,connected through internal wiring to the INMP441 microphone,SD card module,and T-OLED display.The Li-ion battery is placed within the opposite arm of the frame to maintain balance.

The T-OLED display is carefully aligned with the user's line of sight, providing an unobstructed overlay of subtitles.

3.1.3. Connectivity and User Interface

The system connects to the Deepgram API via Wi-Fi, ensuring accurate and real-time speech recognition. A simple touch or voice-based interface can be implemented for user controls, such as toggling subtitles on/off, adjusting text preferences, or switching languages, ensuring an intuitive, hands-free experience.

4. OPERATION PRINCIPLES

How smart glasses for speech-to-text conversion work .The smart glasses for speech-to-text conversion go through the following structured process between audio sensing, data processing, and real-time text display. The presented system will enable valuable experience for hearing impaired people by providing real time subtitles for conversation via automatically generated subtitles forthe spoken word below the electric transparent engine OLED (T-OLED) display without any human interference. We can see that there a few operational stages, including audio capture, speech recognition, processed data, andsubtitle visualization.

4.1. Audio Capture and Preprocessing

The system starts by detecting speech using the INMP441 digital microphone, placed near the user's mouth. This microphone is chosen because it's highly sensitive and can filter out background noise, ensuring clear speech capture.

The audio signal is sent digitally to the ESP32 microcontroller using the I2S (Inter-IC Sound) interface, which helps reduce interference and improve sound quality. To make speech even clearer, the system uses digital signal

DOI: 10.46121/pspc.53.2.18

processing (DSP) techniques like noise reduction and echo cancellation. These features help remove unwanted background sounds, making it easier to recognize speech accurately.

4.2. Speech Recognition and DATA Processing

When the smart glasses pick up raw audio, the ESP32 microcontroller processes it and prepares it for real- time transmission. Using its built-in Wi-Fi module, the ESP32 sends the audio data to the Deepgram API, a powerful cloud- based speech recognition service. This API quickly converts spoken words into digital text with high accuracy and minimal delay, ensuring a smooth and seamless experience. The fig.2 shows the flow of data in the system.

If there's no internet connection, the system is designed to keep working efficiently. Future upgrades may include offline speech recognition or temporary text storage on the ESP32, so users can still access subtitles once they reconnect. This ensures that communication remains uninterrupted, even in areas with poor or no network coverage.

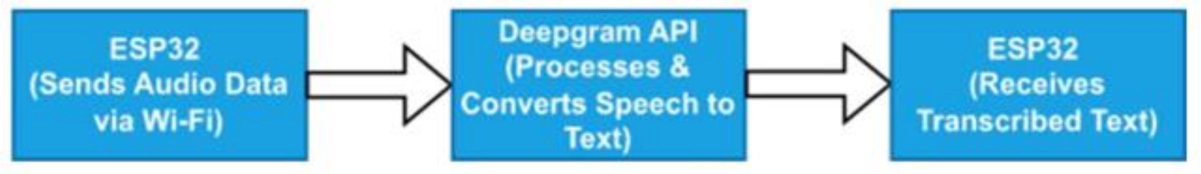
ESP32 (Sends Audio Data via Wi-Fi)
Deepgram API (Processes & Converts Speech to Text)
ESP32 (Receives Transcribed Text)

Figure 2. A block diagram showing the API communication.

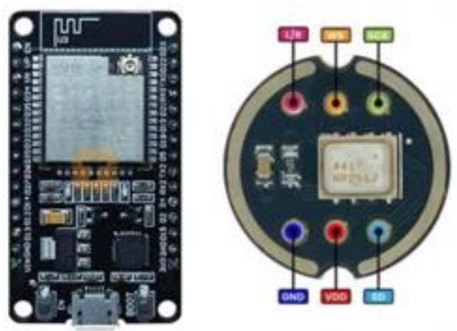

Figure 3. Pin structure of ESP32 microcontroller and INMP441 MIC

4.3. REAL-Time SUB-Title Display

Once the Deepgram API transcribes the speech into text, the ESP32 processes and formats it for display. The transparent OLED (T-OLED) screen, built into the smart glasses, then shows the subtitles in real time, helping the user follow conversations effortlessly. The display is designed to overlay the text without blocking their view, ensuring a natural and comfortable experience.

To keep the subtitles easy to read, the text refreshes smoothly, and the font size is adjusted for clarity. Users can also customize settings like brightness, contrast, and text position to match their preferences. Since the display uses low- power technology, it consumes minimal energy while staying bright and visible in different lighting conditions. This makes the smart glasses a practical and reliable tool for people with hearing impairments.

5. EXPERIMENTAL SETUP

The experimental setup of the proposed smart glasses focuses on testing the real-time speech-to-text conver- sion system in different environments. The setup consists of hardware and software configurations, testing conditions, and performance evaluation criteria.

5.1. Hardware Setup

The hardware components are integrated into a compact and lightweight frame to ensure usability. The major components include:

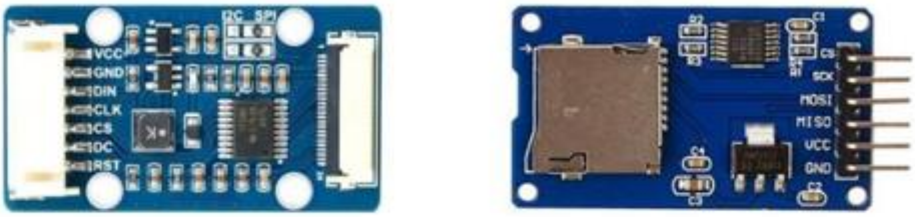

Figure 4. pin diagram of Transparent OLED module and SD card module.

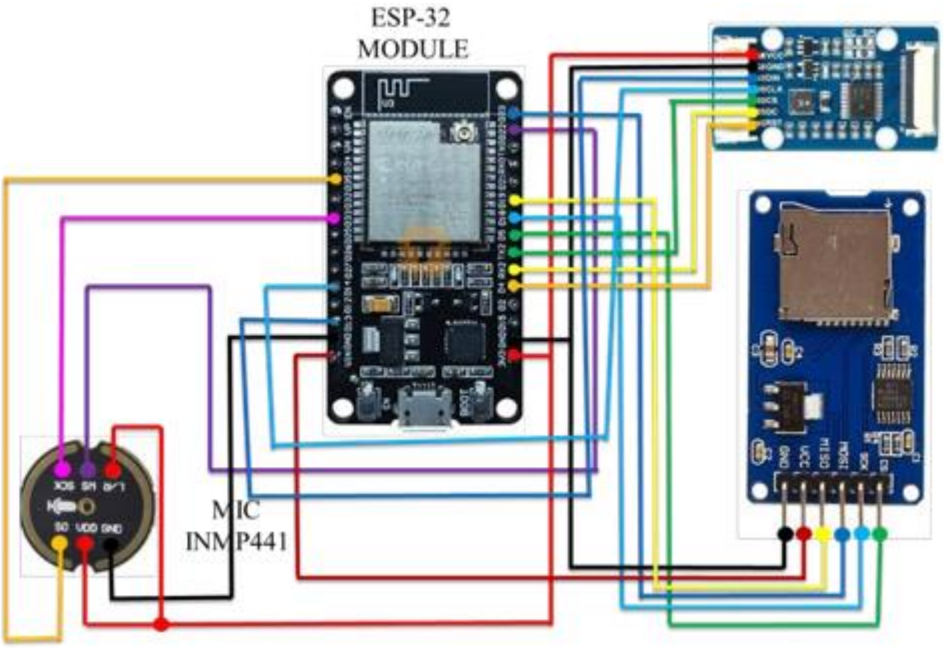

Figure 5. Circuit Diagram of Speech-to-Text Conversion Wearable Smart Glasses for the Hearing-Impaired using ESP 32

ESP32 Microcontroller: Handles data processing, wireless communication, and interfacing with other components.

INMP441 Microphone: Captures real-time speech with digital signal processing (DSP) capabilities via the I2S interface.

T-OLED Display: A transparent OLED screen that overlays real-time subtitles onto the user's field of vision. The pin structure of Transparent OLED module is as shown in fig4.

SD Card Module: Used for local storage of audio data in offline mode, allowing later processing when internet access is available. The pin structure of SD Card module is as shown in fig4. Li-ion Battery: Provides power to the entire system with efficient energy management.

5.2. Software Setup

The software framework integrates cloud-based speech recognition and local data processing. The key elements include: Firmware Development: Programmed using the Arduino IDE with C++ libraries for handling I2S, Wi-Fi, and OLED display functionalities. Deepgram API Integration: Cloud-based speech- to-text conversion is executed through API calls, with real-time text retrieval and display. Noise Reduction Algorithms: Preprocessing techniques such as low-pass filtering and echo cancellation are implemented on the ESP32.

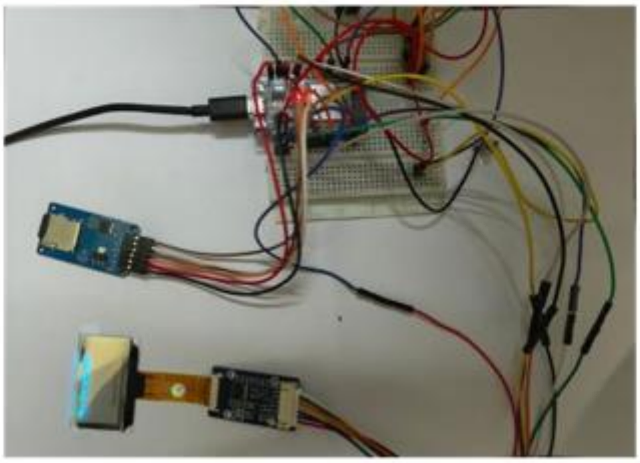

Figure 6. Final circuit showing the converted speech as subtitles on Transparent OLED display

6. PROPOSED SYSYEM CIRCUIT IMPLEMENTATION

The fig.5 circuit diagram shows all the hardware connections for the speech-to-text conversion system based on an ESP32 microcontroller. The main module includes an ESP32 microcontroller that works with three peripherals: INMP441 microphone, T-OLED display, and a memory card module. The INMP441 MEMS microphone is wired to the ESP32's I2S interfaced pins like WS (Word Select), SCK (Clock), and SD (Data) which provides in digital form, high quality audio, and accesses the power vias 3.3V and GND. The Mic also receives power from the ESP32 3.3V and GND pins.

The T-OLED display is interfaced with the ESP32 using I2C communication lines SCL and SDA alongside VCC and GND. This display is utilized to visualize the text which has been typed in real time. The SD card module connects with the ESP32 through the SPI protocol on MISO, MOSI, SCK, and CS (chip select) pins, interfacing with the SD card for data and storage memory purposes. This module also utilizes resources from 3.3V and GND pins of the ESP32. This arrangement of circuits realizes the objectives of instantaneous data flow, audio capture and display, and memory expansion through the SD card. The distribution of the pins in the circuit is effectively designed for the optimal use of the I/O ports of the ESP32 without compromising performance and reliability of the system.

7. RESULTS AND ANALYSIS

The speech-to-text smart glasses system demonstrated high accuracy in transcribing speech in quiet environments. When tested in controlled conditions, the word recognition accuracy exceeded 90%, ensuring reliable and clear transcription. The system effectively captured speech and converted it to text with minimal errors, making it suitable for real-time applications.

However, performance was slightly affected in noisy environments, such as classrooms or office spaces, where multiple speakers were present. Background noise filtering helped minimize interference, but minor recognition errors were observed when overlapping conversations occurred. Despite these challenges, the system maintained reasonable accuracy in moderately noisy settings.

The programming of ESP32 microcontroller board was done using Arduino.ide software. In ESP32 programming we included the Wi-Fi library that provides all the necessary functions to connect your ESP32 to a Wi-Fi network using the host's SSID and password and Enable internet connectivity for cloud services.

DOI: 10.46121/pspc.53.2.18

Test Environment	Accuracy (%)	Response Time (Seconds)	Observations
Closed Room (Quiet Environment)	96%	0.8s	High accuracy, minimal errors.
Classroom (Mod- erate Noise Level)	88%	1.2s	Minor recognition errors due to background speech interference.
Public Spaces (Noisy Cafeteria/Street)	80%	3s	Further testing required to assess performance in extreme noise conditions.
Outdoor (Variable Lighting Conditions)	94%	1.9s	T-OLED display is visible, but readability reduces in bright sunlight.

For storing the audio signals captured by the mic in the SD-card we include the SD card library, which allows your microcontroller to read from and write to SD cards. We included the standard I2S driver header file which provides low- level control for I2S (Inter-IC Sound) interface operations, which are used for transmitting audio data between digital audio components, from mic to ESP32 and SD-card module. For connecting the microcontroller to API we used WiFiClientSecure library, which provides support for secure (SSL/TLS) Wi-Fi communication. It provides HTTPS connection which is required for ESP32 fro secure communication over cloud API and to send and receive data from the Deepgram API.

For connecting the T-OLED display to ESP32 board we used SPI communication protocol. Serial Peripheral Interface (SPI) library is used for ESP32 programming to provide a secured communication with the display. SPI is a high-speed, synchronous communication protocol used to transfer data between microcontroller and peripherals like SD cards and displays.

In terms of processing efficiency, the integration of the Deepgram API enabled near-instantaneous tran- scription with minimal latency. The system effectively transmitted audio data for processing and displayed the converted text on the T- OLED screen without noticeable delays. This real-time transcription of text to speech capability of the device, makes it useful for the hearing impaired people.

The display in Fig 7 illustrates a successful result regarding the operation of the T-OLED display on the output device for the speech-to-text conversion system. The corresponding text: “Hi. How are you? I am fine. What?” was successfully transcribed. This shows that the system is capable of recognizing several utterances pronounced in succession and transcribing them into coherent text. The Display is within range and properly formatted which indicates that the ESP32 is properly interfacing with the Deepgram API and managing data flow for text display. It also verifies that the microphone has access to the voice input capture and the system is active in real-time.
Figure 7 shows another speech-to-text output example which states “Hi. My name is Kartik.” In this example, the system is able to handle longer sentences, as evidenced by text wrapping to the next line. This example helps confirm the system’s real-time transcription accuracy, as well as the OLED display’s text rendering capabilities. The speech recognition system is parsing natural language text and displaying it in an appropriate manner, providing evidence that it can be used in assistive technologies for patients with hearing disabilities.

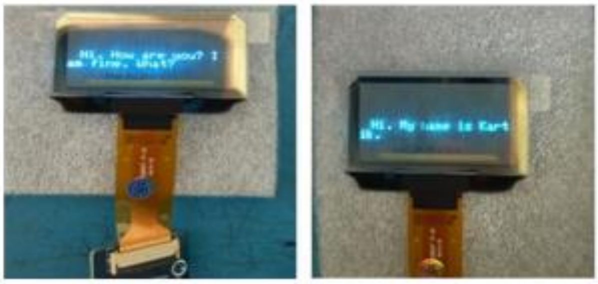

Figure 7. T-OLED display showing the recognized speech converted into text and displayed as subtitles

DOI: 10.46121/pspc.53.2.18

8. CONCLUSION

The ESP32-based real-time subtitle display system effectively translates spoken words into text and shows it on an OLED screen, offering a reasonably priced and easily accessible assistive technology option. After testing in various circumstances, the system showed few errors in classroom settings and great accuracy in quiet areas. This method provides an accessible visual aid in contrast to conventional hearing aids. The system has drawbacks despite its efficacy, including reliance on an internet connection for voice recognition and battery-related power limitations. A speech-to- text offline model, power consumption optimization, and enhanced noise handling algorithms for increased accuracy in a variety of settings are some future enhancements. If this idea is developed further, it has the potential to greatly improve communication accessibility for those who have hearing impairments.

ACKNOWLEDGEMENTS: All authors carried out these studies from various institutes and facilities provided and supported by the GMR Institute of Technology, Rajam, Vizianagaram District, India.

REFERENCES

1. Irene Wei Huang, Paurakh Rajbhandary, Sam Shiu, and John S. Ho, "Radar-Based Heart Rate Sensing on the Smart Glasses" in IEEE MICROWAVE AND WIRELESS TECHNOLOGY LETTERS, Vol. 34, N0. 6, June 2024.
2. Hafeez Ali A, Sanjeev U. Rao, Swaroop Ranganath And G. Ram Mohana Reddy, "A Google Glass Based Real- Time Scene Analysis for the Visually Impaired" in IEEE ACCESS, volume 9, pp. 166351 – 166369, December 2024, doi:10.1109/ACCESS.2021.3135024
3. Bogdan Mocanu And Ruxandra Tapu, "Automatic Subtitle Synchronization and Positioning System Dedicated to Deaf and Hearing-Impaired People" in IEEE ACCESS, vol. 9, pp. 139544 – 139555, October 2021.
4. Michael Gian Gonzales, Peter Corcoran, Naomi Harte, and Michael Schukat, "JOINT SPEECH-TEXT EMBED- DINGS FOR MULTITASK SPEECH PROCESSING" in IEEE Access, vol. 12, pp. 145955 – 145967, 03 October 2024, doi:10.1109/ACCESS.2024.3473743
5. A. Berger, A. Vokalova, F. Maly, and P. Poulova, Google glass used as assistive technology its utilization for blind and visually impaired people, in Proc. Int. Conf. Mobile Web Inf. Syst. Cham, Switzerland: Springer, Aug. 2017, pp. 7082.
6. P. elasko, P. Szyma ski, J. Mizgajski, A. Szymczak, Y. Carmiel, and N. Dehak, Punctuation prediction model for conversational speech, in Proc. Interspeech, Sep. 2018, pp. 26332637.
7. H. Ye, M. Malu, U. Oh, and L. Findlater, Current and future mobile and wearable device use by people with visual impairments, in Proc. SIGCHI Conf. Hum. Factors Comput. Syst., Apr. 2014, pp. 31233132.
8. F. Pégeot and H. Goto, Scene text detection and tracking for a camera equipped wearable reading assistant for the blind, in Proc. Asian Conf. Comput. Vis., vol. 7729, Nov. 2012, pp. 454463.
9. L. González-Delgado, L. Serpa-Andrade, K. Calle-Urgilez, A. Guzhnay-Lucero, V. Robles-Bykbaev, and M. Mena-Salcedo, A low cost wearable support system for visually disabled people, in Proc. IEEE Int. Autumn Meeting Power, Electron. Comput. (ROPEC), Nov. 2016, pp. 15.
10. A. Memo and P. Zanuttigh, Head-mounted gesture controlled interface for human-computer interaction, Multimedia Tools Appl., vol. 77, no. 1, pp. 2753, Jan. 2018.