WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Speech-to-Text Conversion Wearable Smart Glasses for the Hearing-Impaired using ESP 32

Dr. G. Nookaraju[1], M. Teja Sai Kumar[2], P. Naveen Kumar[3], I. Karthik Chandu[4], G. Sameer Kumar[5]

[1] Sr. Assistant Professor, Department of Electronics and Communication Engineering,GMR Institute of Technology, Rajam, Andhra Pradesh, India 532127.

[2 3 4 5]Students, Department of Electronics and Communication Engineering, GMR Institute of Technology, Rajam, Andhra Pradesh, India 532127

1nookaraju.g@gmrit.edu.in

DOI : https://doi.org/10.55248/gengpi.6.0425.14168

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.For people with hearing impairments, assistive technologies like speech-to-text apps, sign language, and hearing aids have greatly increased accessibility and improved communication. The goal of ongoing research is to create better ways to promote inclusivity and integration in society. In order to integrate Deepgram's API for real-time speech recognition and conversion tasks as a hearing assistant, this paper introduces the ESP-32 device, which will serve as the main microcontroller board. An INMP441 digital microphone built into the device records ambient sound and transforms it into a digital signal.The integrated 24-bit ADC converts analog signals to digital ones and communicates with the ESP-32 microcontroller board via I2S protocol. The ESP-32 processes digital audio data and transmits it via Wi-Fi to the Deepgram API. The API transcribes speech to text and sends it to the ESP-32 microcontroller. The text is displayed as subtitles on the glasses using a TOLED connected to the ESP-32 via I2C protocol. This proposed innovation aims to improve communication in social, educational, and business settings, fostering a shared space between hearing and non- hearing individuals through technology.

***Index Terms*—ESP32, I2S and I2C protocol,INMP441 mi- crophone, Speech-to-Text, Deepgram API, IoT Audio Applica- tions,Base64 Encoding**

I. Introduction

Wearable technology is developing quickly and presents fresh opportunities to enhance daily living, particularly for those with disabilities. Smart glasses that can convert speech to text in real time are an interesting new development in this field. For those who have hearing loss, these glasses are revolutionary because they make discussions more accessible by displaying spoken words as subtitles. In contrast to conven- tional options like external microphones or hearing aids, which can be uncomfortable and have trouble in noisy settings, smart glasses provide a hands-free, non-intrusive option.Lighting and background noise, however, can still have an impact on their accuracy.

In this work, a sophisticated speech-to-text system that uses an ESP32 microcontroller, Deepgram API, a transparent OLED (T-OLED) display, and an INMP441 microphone is integrated into smart glasses. Highly accurate speech recogni- tion is provided by the Deepgram API, while the ESP32 guarantees processing efficiency without using excessive power. The INMP441 microphone records high-quality audio, even in noisy environments, and the T-OLED display enables crisper real-time subtitles.When combined, these elements produce a more dependable and useful assistive technology for those who are hard of hearing.

The feasibility of integrating such a system into smart glasses is supported by recent research. Ali et al. [2] demon- strated the effectiveness of Google Glass in real-time scene analysis, highlighting the potential of head- mounted devices for data processing and display. Similarly, Huang et al.

[1] explored radar-based sensors for non-contact heart rate monitoring in smart glasses, emphasizing their versatility for advanced sensing and processing applications. Building on these advancements, this study introduces a comprehensive speech-to- text system designed for portability, ease of use, and real-time functionality. By combining hardware and software innovations, this solution aims to set a new benchmark for assistive wearable technology.

In conclusion, integrating real-time speech-to-text function- ality into smart glasses represents a significant advancement in assistive technology. This innovation offers a useful and approachable solution for people with hearing impairments by addressing the drawbacks of conventional systems and utilizing the most recent developments. It not only increases independence and accessibility, but it also creates opportunities for wearable assistive technology advancements in the future.

II. DESIGN

The design of the proposed smart glasses for speech-to- text conversion is mainly focused on achieving efficiency, wearability, and real-time performance while 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 flow of data with in the system. 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.

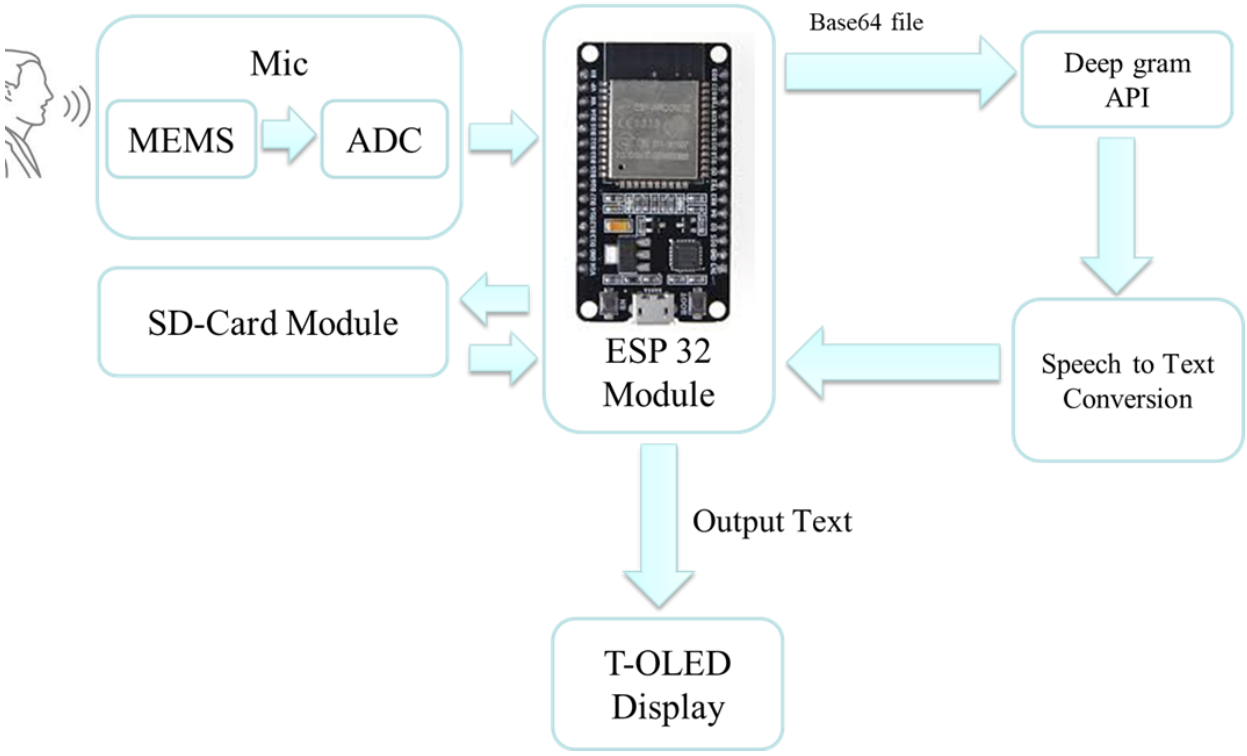

Fig. 1. The figure represents the block diagram of speech-to-text conversion system using an ESP32 module and outlines the key components and

The ESP32 microcontroller module is the core processing unit of the system. It receives the digitized audio data from the mic and processes it and stores as a audio file in the SD- card module. This audio file is processed and converted into base64 format for transmission. To perform speech recognition and transcription, the ESP32 board sends the audio file to the Deepgram API which is a cloud-based speech processing service. This API utilizes advanced machine learning models to accurately convert speech into text. Once the speech-to- text conversion is complete, the textual output is sent back to the ESP32. Finally the concerted text is displayed on the transparent OLED display.

A. Structural Design

These smart glasses are built for all-day comfort ,with a lightweight and ergonomic design that will not strain the user. The frame is made from durable yet feather-light materials such as 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 unobtru- sive.everything is designed with ease and comfort in mind,so users can go about their day without feeling weighed down.

B. Component Placement and Integration

The T-OLED display, SD card module, and INMP441 mi- crophone are all wired internally to the ESP32 microcontroller, which is embedded on one side of the frame. To keep the frame balanced, the Li-ion battery is positioned inside the other arm. The T-OLED display offers an unhindered overlay of subtitles because it is precisely positioned in the user's line of sight.

C. 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 pref- erences, or switching languages, ensuring an intuitive, hands- free experience.

III. Operation Principles

How speech-to-text conversion works with smart glasses.The following structured procedure is followed by smart glasses that convert speech to text: audio sensing, data processing, and real-time text display. By providing real- time subtitles for conversations using automatically generated subtitles for spoken words beneath the electric transparent engine OLED (T-OLED) display without human intervention, the presented system will provide hearing-impaired people with a valuable experience. We can observe that there are several operational phases, such as speech recognition, audio recording, data processing, and subtitle display.

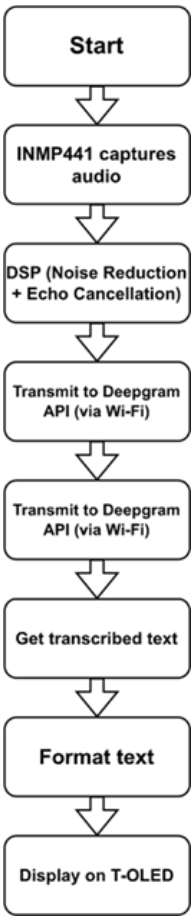

Fig. 2. Design of the speech-to-text conversion system using an ESP32 module

A. Audio Capture and Preprocessing

Using the INMP441 digital microphone, which is positioned close to the user's mouth, the system first detects speech. This microphone was selected due to its high sensitivity and ability to eliminate background noise, guaranteeing crystal- clear speech recording. The INMP441 is a high-performance, low-power, digital output, omnidirectional MEMS microphone with an integrated analog to digital converter(ADC). The complete INMP441 solution consists of a MEMS sensor, signal conditioning, analog to digital converter, anti-aliasing filter, power management, and industry-standard 24-bit I2S interface. The MEMS captures the physical audio signal and converts that signal to analog electrical signal and this signal is converted to digital signal by the integrated ADC fro digital transmission of the audio to ESP32 microcontroller board. The I2S (Inter-IC Sound) interface is used to digitally transmit the audio signal to the ESP32 microcontroller, enhancing sound quality and lowering interference. The system employs digital signal processing (DSP) methods like echo cancellation and noise reduction to further improve speech clarity. By eliminating unwanted background noise, these features aid in more precise speech recognition.

Fig. 3. The communication process between Deepgram API and ESP 32

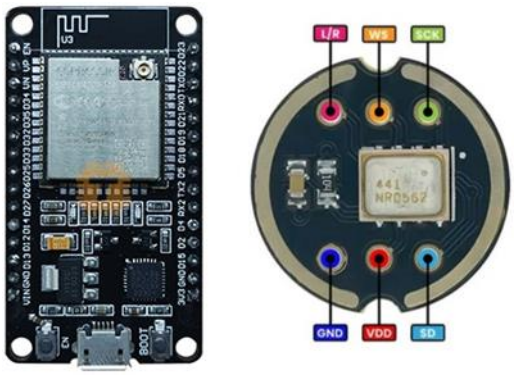

Fig. 4. Pin structure of ESP32 microcontroller and INMP441 MIC

B. Speech Recognition and Data Processing

The ESP32 microcontroller processes the raw audio that is received from the INMP441 Mic and gets it ready for real-time transmission. The microcontroller board saves the received raw audio data in the SD-card module as a file. Before the ESP32 transmits audio data to the API the stored file was converted into base64 format as the API only accepts the base64 format files. Deepgram API is a potent cloud-based speech recognition service, that offers speech recognition and speech to text and text to speech transcription with more than 95%. ESP32 transmits audio data to Deepgram API via its integrated Wi-Fi module. This API ensures a smooth and seamless experience by rapidly converting spoken words into digital text with high accuracy and little delay. The system is built to continue operating effectively even in the absence of an internet connection. In order to allow users to access subtitles after reconnecting, future updates might incorporate offline speech recognition or temporary text storage on the ESP32. This guarantees continuous communication even in places with spotty or nonexistent network coverage.

The ESP32 processes and formats the speech for display after it has been converted to text by the Deepgram API. The user can easily follow conversations thanks to the real-time subtitles displayed on the transparent OLED(T-OLED) screen that is integrated into the smart glasses. In order to provide a seamless and comfortable experience, the display is made to overlay the text without obstructing their view.

The font size is changed for clarity and the text refreshes smoothly to make the subtitles easy to read. Additionally, users can alter text position, contrast, and brightness to suit their tastes. The display uses little energy and remains bright and visible in a variety of lighting conditions because it employs low-power technology. Because of this, smart glasses are a useful and dependable aid for those who have hearing loss.

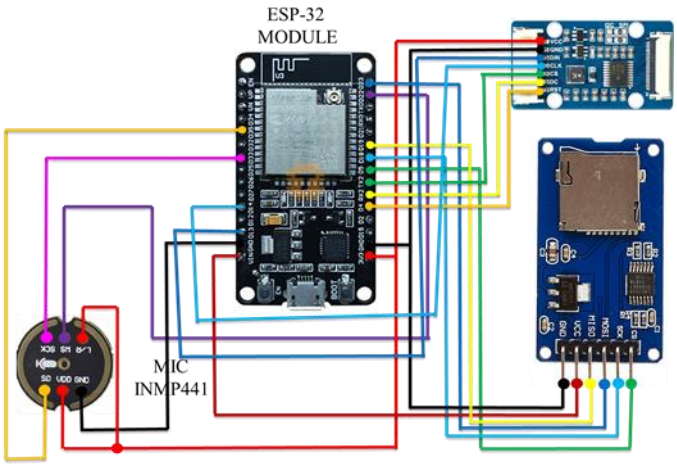

Fig. 5. Circuit Diagram of Speech-to-Text Conversion Wearable Smart Glasses for the Hearing-Impaired using ESP 32

IV. Experimental Setup

The experimental setup focuses on evaluating the perfor- mance of the proposed smart glasses in real-world conditions. The testing is conducted in various environments to assess the system's accuracy, response time, and usability for hearing- impaired individuals.

A. Hardware Setup

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

ESP32 Microcontroller: The ESP32 is a low-cost 38 pin microcontroller board with integrated Wi-Fi and Bluetooth, making it suitable for IoT applications. It is used for handling data processing, wireless communication, and interfacing with other components.

INMP441 Microphone: Captures real-time speech with digital signal processing (DSP) capabilities via the I2S in- terface with High signal to noise ratio of 61 dBA and high sensitivity of 26 dBFS. It captures the audio and converts it into digital format.

T-OLED Display: A transparent OLED screen that overlays real-time subtitles onto the user's field of vision.

SD Card Module: the SD card module is a small board that allows microcontroller to interface with standard SD or microSD cards for storing the audio files. The communication between SD Card module and ESP32 board is achieved by using I2C communication protocols

Li-ion Battery: Provides power to the entire system with efficient energy management.

B. software Setup

The software framework integrates cloud-based speech recognition and local data processing. The key elements in- clude:

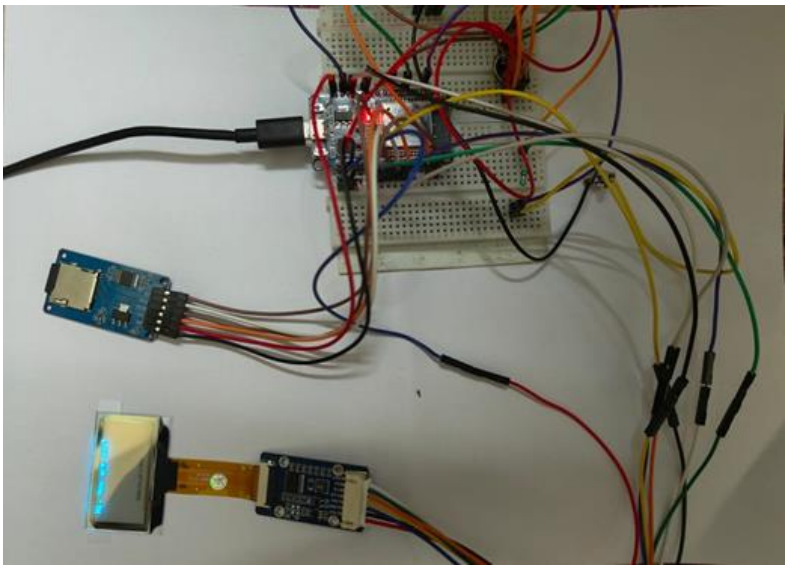

Fig. 6. Final circuit showing the converted speech as subtitles on Transparent OLED display

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 imple- mented on the ESP32.

C. Testing Environments

The prototype was tested in different real-world conditions to evaluate its accuracy and usability. Based on initial testing, the system performs well in closed-room environments but shows minor errors in classroom environments. To refine the evaluation, the following test environments are considered:
Closed Room (Ideal Conditions) – High Accuracy

A quiet indoor setting with minimal background noise.

The system functions with high accuracy, providing clear and real-time speech-to-text conversion.

Classroom or Office (Moderate Noise) – Minor Errors Observed

A structured environment with multiple people talking simultaneously.

Some minor recognition errors were observed, likely due to background conversations affecting speech isolation.

V. RESULTS

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.

Test Environment	**Accuracy (%)**	**Response Time (Seconds)**	**Observations**
Closed Room (Quiet Environment)	96%	0.8s	High accuracy, minimal errors.
Classroom (Moderate 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.

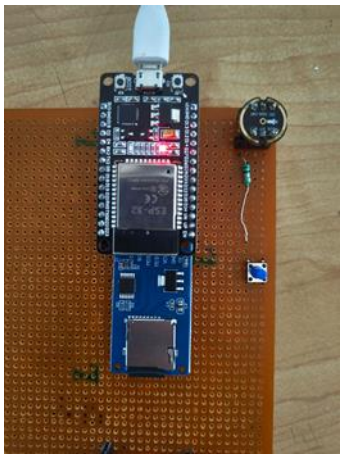

Fig. 7. Speech to text converting system designed on a on a PCB

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. 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 WiFi- ClientSecure library, which provides support for secure (SSL/TLS) Wi-Fi communication. It provides HTTPS connec- tion which is required for ESP32 fro secure communication over cloud API and to send and receive data from the Deep- gram 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 transcription 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.

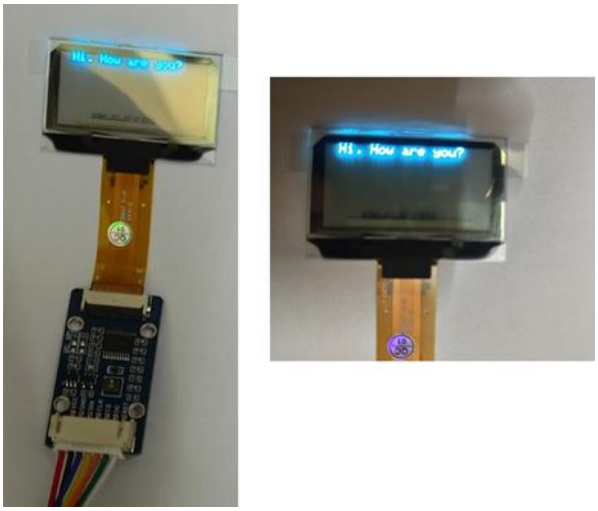

Fig. 8. T-OLED display showing the recognized speech converted into text and displayed as subtitles

VI. Conclusion

The ESP32-based real-time subtitle display system effec- tively 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 set- tings 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, includ- ing 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 set- tings 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.

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 Synchroniza- tion 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 EMBEDDINGS 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. Pe´geot 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. Gonza´lez-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.