

TalkLens – Integrated Vision and Sign Language Recognition for Inclusive Human Interaction

1st Dr. Sanjay Malode

Artificial Intelligence and Data Science
K.D.K. College of Engineering
Nagpur, India

2nd Ganesh Dhere

Artificial Intelligence and Data Science
K.D.K. College of Engineering
Nagpur, India

3rd Gouri Kashettiwar

Artificial Intelligence and Data Science
K.D.K. College of Engineering
Nagpur, India

4th Rhushikesh Ugemuge

Artificial Intelligence and Data Science
K.D.K. College of Engineering
Nagpur, India

5th Vaishnavi Bele

Artificial Intelligence and Data Science
K.D.K. College of Engineering
Nagpur, India

6th Yarmika Narad

Artificial Intelligence and Data Science
K.D.K. College of Engineering
Nagpur, India

Abstract—Communication barriers greatly impact the independence, social engagement, and overall quality of life for individuals with visual, speech, or hearing challenges. This review looks at the TalkLens framework, a new solution aimed at overcoming these barriers. It integrates real-time vision-based object detection, sign language recognition, and natural language processing with empathetic artificial intelligence. The framework uses technologies like TensorFlow SSD MobileNet for object detection, MiniGPT-4 for understanding vision and language, Sentence-BERT for processing text, and natural text-to-speech synthesis to enable smooth communication. By combining these elements, TalkLens can interpret visual and textual inputs accurately and turn them into meaningful outputs while responding empathetically to users' needs. The system not only improves accessibility but also encourages social inclusion by allowing users to interact confidently in various real-world situations. Additionally, TalkLens tackles the issues faced by traditional assistive technologies by offering real-time, adaptable, and smart support that fosters active participation in educational, professional, and social settings. This review critically evaluates the design, methods, and possible applications of TalkLens, emphasizing its role in promoting independence, enhancing communication effectiveness, and contributing to a more inclusive society. By bringing together recent advancements in vision-language AI, sign language processing, and empathetic interaction systems, this study highlights the potential of TalkLens to create meaningful and inclusive human-computer interactions.

Keywords—Vision-based object detection, Sign language recognition, Empathetic AI, Natural language processing, Text-to-speech synthesis, Inclusive communication.

I. INTRODUCTION

Communication is a fundamental part of human life. It allows people to share emotions, exchange ideas, and form relationships. However, for millions worldwide who are visually, hearing, or speech-impaired, communication can be challenging. These sensory impairments create barriers that limit participation in education, jobs, and social activities. The World Health Organization (WHO) states that over 1 billion people have some form of disability, with nearly 430 million living with disabling hearing loss. Even with the growth of assistive technologies, a gap still exists between available solutions and the complex communication needs of differently-abled users.

Traditional assistive systems, such as text-to-speech devices, Braille tools, or sign language translators, usually focus on just one aspect of disability. Some systems only help with visual impairment by turning text or images into speech. Others are designed only to interpret sign language gestures. While these tools can be effective on their own, they often do not support inclusive, two-way communication involving multiple disabilities. This divide forces users to rely on others or limits them to specific situations, which can lower their independence and confidence.

The growth of artificial intelligence (AI) has created new possibilities for closing these gaps. Deep learning models now allow for real-time visual understanding, natural

language comprehension, and human-like speech generation, capabilities that once seemed like science fiction. By combining these technologies in one framework, we can develop systems that understand and respond to users in context, similar to human interactions. The TalkLens framework is envisioned as an inclusive AI-powered ecosystem. It combines vision-based object detection, sign language recognition, and empathetic natural language interaction into a single platform, helping users communicate and navigate their surroundings more naturally and confidently.

TalkLens aims to go beyond just providing assistive functions. It seeks to enable empathetic interaction. Unlike traditional AI systems that only handle input and output, empathetic AI focuses on grasping intent, tone, and emotional context, resulting in a more human-like experience for users. By using models like TensorFlow SSD MobileNet for object detection, MiniGPT-4 for visual-linguistic reasoning, Sentence-BERT for contextual understanding, and modern text-to-speech synthesis for expressive audio feedback, TalkLens brings together different types of AI into one effective communication platform.

The future of human-centered AI depends on its ability to meet the needs of diverse users, consider emotional contexts, and ensure fairness in communication. TalkLens represents a significant step toward that future. It envisions a world where technology not only assists but genuinely connects people, regardless of their abilities.

II. LITERATURE SURVEY

The field of assistive technology has seen rapid growth in recent years, thanks to the use of artificial intelligence (AI), deep learning, and computer vision. Many studies have looked into ways to improve accessibility for people with visual, hearing, and speech impairments. One of the early projects in this area is Diksha R. Pawar's "Digital Eyes" Android app, which uses real-time object detection to help visually impaired people identify objects around them. By providing audio feedback about the environment, it significantly improves navigation safety and awareness, showing the real value of vision-based assistive tools [1]. However, its focus is mostly on visual impairment, which points to the need for systems that integrate multiple senses for better communication.

Xianwei Jiang and colleagues (2024) conducted a thorough review of deep learning methods in sign language recognition. They explained how convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have greatly improved gesture classification accuracy [2]. They stressed the importance of having diverse datasets and being able to adapt in real time across different sign languages. They noted that scaling these solutions and understanding context are still major

challenges. This review lays the groundwork for combining gesture-based communication systems with advanced language models to create more natural translations between sign and spoken language.

Similarly, Dechrit Maneetham and his team (2024) developed a compact object recognition system designed for devices like the Raspberry Pi 4, which have limited resources [3]. Their work showed that by optimizing neural network architectures, reliable real-time recognition can be achieved without heavy computing power. This is especially useful for portable assistive systems like TalkLens, which aim to provide ongoing support through wearables and smartphones.

Another significant development in combining vision and language is MiniGPT-4, introduced by Zhu et al. (2023) [4]. This model successfully merges pretrained visual encoders with large language models, enabling better visual reasoning and generating descriptive captions for complex images. The MiniGPT-4 framework has opened doors for multimodal AI systems that can understand and communicate using both visual and textual cues. For TalkLens, this merging is essential for connecting visual context with natural human language to provide meaningful, context-aware feedback.

For understanding semantics, Sentence-BERT (SBERT), created by Reimers and Gurevych (2019) [5], marks a breakthrough in natural language processing. Unlike traditional BERT models, SBERT applies a Siamese network structure that generates rich embeddings, allowing for effective comparison and contextual matching of sentences. This feature makes it especially useful for interpreting the intent behind gestures or text in sign language translation, ensuring that emotional tone and contextual meaning are not lost.

Recent advances in speech synthesis have also significantly contributed to assistive communication. State-of-the-art models that rely on mel spectrogram predictions can now produce human-like, expressive speech that closely resembles natural tone and rhythm. Such systems are crucial for real-time feedback in inclusive platforms, enabling users to communicate in a more natural and emotionally resonant way instead of sounding robotic.

Together, these studies highlight the growing trend towards multi-modal AI, where computer vision, natural language processing, and speech synthesis come together to improve human-AI interaction. However, most of these efforts focus on separate components rather than integrated frameworks. This gap highlights the unique position of TalkLens, which combines object detection, sign language recognition, and natural language generation in one system. By utilizing the strengths of existing technologies, TalkLens aims to

create an empathetic and inclusive platform that empowers differently-abled users through intelligent, context-aware communication.

III. PROBLEM STATEMENT

Millions of individuals with disabilities face ongoing communication barriers that limit their engagement in daily life, education, work, and social interactions. Despite significant advancements in artificial intelligence and assistive technologies, most available systems focus on just one disability, such as visual, hearing, or speech impairment. This lack of integration leads to fragmented communication experiences, as users often have to rely on several tools to express or understand information. Consequently, true two-way interaction becomes challenging, which decreases opportunities for meaningful engagement and independence.

Additionally, traditional assistive devices mainly serve as supportive aids rather than as caring communication partners. They tend to be rigid and lack an understanding of context and emotional cues. These systems struggle to interpret real-world complexities, like tone, facial expressions, environmental signals, or intent—factors that are crucial for natural human communication. This lack of adaptive intelligence can make users feel dependent and socially isolated, reinforcing barriers instead of eliminating them.

In many public, professional, and academic environments, the lack of inclusive and intelligent communication systems continues to limit accessibility and equality. There is an urgent need for a unified framework that can combine vision-based object detection, sign language recognition, and natural language processing to create smooth, two-way communication that is both accurate and compassionate.

TalkLens aims to address this important gap by merging these technologies into a single, accessible system that not only interprets but also understands and responds with care. By empowering differently-abled individuals to interact confidently in everyday situations, TalkLens seeks to change communication from a challenge into a shared human experience.

IV. METHODOLOGY AND TECHNIQUES

The method used by TalkLens focuses on creating an empathetic AI-based communication system. This system allows smooth interactions among people with visual, speech, or hearing impairments. It combines computer vision, speech processing, and natural language understanding to create real-time, two-way communication. The development process follows a flexible design strategy, which supports easy integration.

1. Development Approach

The system is built in modular stages to ensure it can scale and function accurately. Each main module, including object detection, gesture recognition, and speech processing, is developed and tested separately before the complete system is combined. This method supports continuous improvements based on user feedback and testing outcomes.

2. Tools and Technologies

- Programming Language: Python 3.x
- Frameworks and Libraries: TensorFlow, Keras/PyTorch, OpenCV, NumPy
- Text-to-Speech Engine: Google TTS API / pyttsx3
- Speech Recognition: Google Speech-to-Text / Vosk
- Datasets:
 - COCO Dataset (Object Detection)
 - Custom Gesture and Environmental Datasets (for Sign and Scene Recognition)
- IDE: Jupyter Notebook / Visual Studio Code

3. System Workflow

1. Start Application: The system sets up camera and microphone inputs.
2. Input Capture: It captures live video and audio from the surroundings.
3. Processing:
 - The Object Detection Module identifies and classifies nearby objects or obstacles.
 - The Gesture Recognition Module interprets non-verbal signs for communication.
 - The Speech Recognition Module converts spoken words into text commands.
4. Output Generation: This step converts detected objects, gestures, or speech into accessible formats using Text-to-Speech (TTS) or on-screen text.
5. Interactive Query Response: Users can interact with the system through questions like "Where am I?" or "What's in front of me?"
6. Priority Handling: Critical alerts, such as "Step down!" or "Obstacle ahead!" take precedence over non-urgent narration.

4. Core Modules

- Camera and Audio Capture Module: Gathers real-time visual and audio input.
- Detection and Mapping Module: Uses deep learning to find, locate, and describe objects.
- Gesture Recognition Module: Turns sign language gestures into text or speech.
- Speech Processing Module: Manages speech-to-text and text-to-speech conversions.
- Accessibility Output Module: Provides output as audio for the visually impaired or as text for the hearing impaired.
- Feedback Logger: Records user interactions and system responses for improvement.

5. Algorithm and Model Design

Object Detection & Spatial Mapping:

- Input: Live video frames
- Processing:
 - Frame preparation and feature extraction using CNN
 - Object classification and bounding box prediction
 - Distance and direction estimation
- Output: Context-aware narration, for example, "Door ahead, 3 meters to the right."

Text-to-Speech Workflow: This converts processed text output into natural-sounding speech with adjustable tone, speed, and urgency settings.

6. Testing and Evaluation

- Unit Testing: Each module is tested on its own to ensure it functions correctly.
- Integration Testing: This ensures that modules work together smoothly for real-time performance.
- User Testing: This is done with differently-abled users to evaluate comfort, response time, and usability.

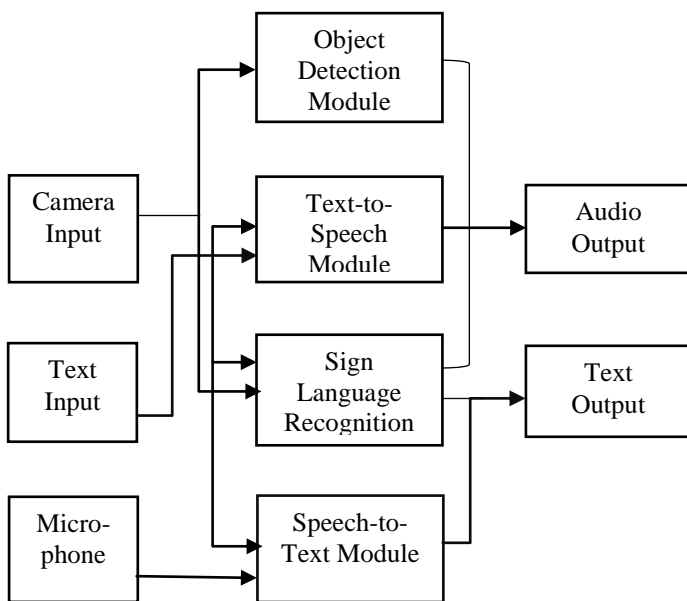
Evaluation Metrics:

- Accuracy in detection and recognition
- Delay between input and output response
- User satisfaction and accessibility
- System stability during extended use

7. Maintenance and Feedback

The system continually learns from user interactions. The administrator updates datasets, retrains models, and adjusts performance based on logged feedback and error repor

V.ARCHITECTURE



VI. PROPOSED APPROACH FOR TalkLens

With the help of modern artificial intelligence technologies, TalkLens offers a unique and caring framework designed to improve communication for individuals with visual, speech, or hearing challenges. The system uses deep learning-based sign language recognition algorithms that can convert gestures into both text and speech. This ensures that non-verbal expressions are made into meaningful communication in real time. It also includes real-time object detection, powered by TensorFlow SSD MobileNet, which helps users perceive and respond to their surroundings more effectively. This feature brings together visual cues and verbal interaction.

To provide a better understanding of environmental inputs and improve comprehension beyond isolated gestures, TalkLens uses vision-language models like MiniGPT-4. This model can interpret complex visual and contextual scenarios. Additionally, Sentence-BERT embeddings improve the system’s natural language processing by capturing subtle relationships and meanings. This ensures more accurate recognition of user intent and personalized communication. The integration of mel spectrogram-based text-to-speech synthesis completes the communication loop, producing realistic, expressive, and easily understandable spoken responses in a natural conversational tone.

Beyond its technical setup, TalkLens strongly emphasizes human-centered design principles focused on empathy, inclusivity, and flexibility. Unlike traditional assistive tools that often follow strict, rule-based patterns, TalkLens learns from user interactions. It adjusts to personal preferences, tone, and environmental context. This flexibility allows users with different abilities to communicate naturally without feeling limited by rigid system behavior. The emotional sensitivity in its AI design ensures that responses remain polite, aware of context, and socially appropriate, promoting dignity and independence among users.

Moreover, the system’s design allows for future growth and integration with new technologies. It can be expanded to include regional sign languages, multilingual translation, and even emotion recognition to interpret facial expressions and tone variations. Integrating wearable devices like smart glasses or haptic feedback systems can further enhance user experience by providing instant awareness of the environment and private communication cues. These additions would make TalkLens not just a communication aid but also a personal assistant that learns, evolves, and interacts with empathy and intelligence.

By combining advanced AI with emotional understanding, TalkLens goes beyond just technical

innovation; it becomes a bridge between technology and humanity. Its seamless, adaptable, and inclusive design ensures that communication is a universal right, where every gesture, sound, and expression is recognized and valued. This vision positions TalkLens as a significant step toward a future where technology not only assists but truly connects people, promoting equality, confidence, and compassion across diverse human experiences.

VII. CONCLUSION

The development of TalkLens marks an important step toward inclusive communication powered by artificial intelligence. It is designed not just as a helpful tool, but as an empathetic medium that connects people with visual, hearing, or speech impairments to the larger world through smooth, intelligent interaction. TalkLens uses modern AI components, including TensorFlow SSD MobileNet for real-time vision, MiniGPT-4 for contextual interpretation, Sentence-BERT for understanding meaning, and mel spectrogram-based text-to-speech synthesis to create a unified system for natural, two-way communication. Unlike traditional assistive systems that focus on single disabilities separately, TalkLens takes a broader approach by combining visual, language, and auditory elements into one seamless experience. This combination improves accessibility and also encourages independence, social confidence, and emotional well-being for users with different abilities. When used in educational, professional, or public settings, this technology can change the concept of equality by allowing people to interact freely without relying on others. As the system continues to develop, it can grow to support regional sign languages, multilingual translation, emotion recognition, and wearable device integration to offer even more adaptable and personalized help. Ultimately, TalkLens shows how empathetic AI can turn inclusion from a goal into a reality. By merging innovation with compassion, it connects technology and humanity, ensuring that communication becomes a universal right, where no voice, gesture, or vision goes unheard or unseen.

VIII. REFERENCES

- [1] D. R. Pawar, "Digital Eyes: Real-time object detection for visually impaired," 2025.
- [2] [2] X. Jiang, et al., "Recent advances in deep learning for sign language recognition," 2024.
- [3] D. Maneetham, et al., "Compact real-time object recognition on Raspberry Pi 4," 2024.
- [4] D. Zhu, et al., "MiniGPT-4: A vision-language model for complex visual-linguistic tasks," 2023.
- [5] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT networks," 2019.
- [6] Neha, Surya Pratap Singh, Ashutosh Pratap Singh, Abhishek Kumar Singh, "Assistive Device for Blind, Deaf and Dumb" 2025.
- [7] Sven Topp, Shuangshuang Xiao, Basil Duvernoy, Jeraldine Milroy, Zhanat Kappassov, Nurlan Kabdyshev, Roope Raisamo, Vincent Hayward & Mounia Ziat, "Mediated and non-mediated tactile fingerspelling: a comparative study" 2024.
- [8] Zifan Jiang, Amit Moryossef, Mathias Müller, Sarah Ebling, "Machine Translation between Spoken Languages and Signed Languages Represented in SignWriting" 2023.
- [9] Aishwarya, Akshata Chougule, Shruti J. Ingaleswar, Vaishnavi G. Pattar, Prof. S. C. Hiremath, "Bridging Communication Gaps: A Literature Survey on Assistive Technologies for Individuals with Disabilities in India" 2024.
- [10] Bader Alsharif, Ali Salem Altaher, Ahmed Altaher, Mohammad Ilyas and Easa Alalwany, "Deep Learning Technology to Recognize American Sign Language Alphabet" 2023.
- [11] Kepeng Wu, Zecheng Li, Hezhen Hu, Wengang Zhou, Houqiang Li, "Cross-Modal Consistency Learning for Sign Language Recognition" 2025.
- [12] Naoki Kimura, Michinari Kono, Jun Rekimoto*, "SottoVoce: An Ultrasound Imaging-Based Silent Speech Interaction Using Deep Neural Networks" 2023.
- [13] Arnav Gangal, Anusha Kuppahally, Malavi Ravindran, "Sign Language Recognition with Convolutional Neural Networks"
- [14] Rajat Singhal, Jatin Gupta, Akhil Sharma, Anushka Gupta, Navya Sharma, "INDIAN SIGN LANGUAGE DETECTION FOR REAL-TIME TRANSLATION USING MACHINE LEARNING" 2025.
- [15] Soukeina Elhassen, Lama Al Khuzayem, Areej Althohali, Ohoud Alzanzami, Nahed Alowaidi, "CONTINUOUS SAUDI SIGN LANGUAGE RECOGNITION: A VISION TRANSFORMER APPROACH" 2025.
- [16] Rachel Beeson, Korin Richmond, "Silent Speech Recognition with Articulator Positions Estimated from Tongue Ultrasound and Lip Video" 2023.

- [17] Savita B Patil , Jeevan Gowda B , Manjula A , Nagalakshmi B G, "SMART COMMUNICATION FRAMEWORK FOR BLIND , DEAF AND DUMB" 2023.
- [18] Nicole Agaronnik, BS, Eric G. Campbell, PhD, Julie Ressalam, MPH, CHES, and Lisa I. Iezzoni, MD, MSc, "Communicating with Patients with Disability: Perspectives of Practicing Physicians"
- [19] Jestin Joy, Kannan Balakrishnan, "A prototype Malayalam to Sign Language Automatic Translator"
- [20] Savindu H.P., Iroshan K.A., Panangala C.D., Perera W.L.D.W.P., De Silva A.C., "BrailleBand: Blind Support Haptic Wearable Band for Communication using Braille Language" 2019.
- [21] J Brabyn, KD Seelman, and S Panchang. Aids for people who are blind or visually impaired. An Introduction to Rehabilitation Engineering, ed. RA Cooper, H. Ohnabe, and DA Hobson, pages 287–313, 2007.
- [22] Tanay Choudhary, Saurabh Kulkarni, and Pradyumna Reddy. A braillebased mobile communication and translation glove for deaf-blind people. In Pervasive Computing (ICPC), 2015 International Conference on, pages 1–4. IEEE, 2015.
- [23] Ben AG Elsendoorn. Assistive technology for the hearing-impaired, deaf and deafblind, by marion e. hersch and michael a. johnson, eds. Technology and Disability, 16(2):111–113, 2004.
- [24] Prof. Sonali Karthik, Shraddha Deshmukh, Archis Save, Rani Shah, "Effective Communication Between Blind, Mute And Deaf People Using A Multi-Model Approach" 2024.
- [25] Divyansh Bisht, Manthan Kojage, Manu Shukla, Yash Patil, Priyanka Bagade Smart Communication System Using Sign Language Interpretation.
- [26] Sartha Tambe; Yugchhaya Galphat; Nilesh Rijhwani; Aishwarya Goythale; Janhvi Patil Analyzing and Enhancing Communication Platforms available for a Deaf-Blind user.
- [27] Naga Thulasi Vundelli; Dharahas venkat G; Hima Bindu Valiveti; Vinay Kumar H; Pratyusha Chowdari Two Way Communicator between Blind-Deaf-Dumb Person and Normal People.