Journal of Theoretical and Applied Information Technology
15th February 2026. Vol.104. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MULTI MODAL AI FOR IMPROVING ACCESSIBILITY THROUGH SPEECH AND GESTURE RECOGNITION FOR THE DISABLED

RAVINDRA CHANGALA[1*], Dr. K PAVAN KUMAR[2], Dr. S ARUNA[3], P MADHAVI[4], N SRINIVASA RAO[5], CH LAVANYA SUSANNA[6], Dr. C SUGANYA[7]

[*1]Department of CSE, Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India

[2]Department of IT, Prasad V Potluri Siddhartha Institute of Technology, Vijayawada, Andhra Pradesh, India

[3]Department of IT, Vasavi College of Engineering, Hyderabad, Telangana, India

[4]Department of CSE, CVR College of Engineering, Vastunagar, Telangana, India

[5]Department of Computer Science and Business System, RVR & JC College of Engineering, Guntur, Andhra Pradesh, India

[6]Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India

[7]Department of MBA, M.Kumarasamy College of Engineering, Karur, Tamil Nadu, India

E-mail: [*1]changalaravindra@gmail.com, [2]pavanpvpsit@gmail.com, [3]s.aruna@staff.vce.ac.in, [4]p.madhvai@cvr.ac.in, [5]hellonsr@gmail.com, [6]lavanyasusanna1991@gmail.com, [7]suganyac.mba@mkce.ac.in

ABSTRACT

In this paper, we report on designing and evaluating a new multimodal AI that multi-modally combines speech and gesture recognition to enhance accessibility for people with disabilities. We aim to create a flexible and resilient interface to allow individuals with various disabilities to interact with technology using one or two input modalities. The model comprises the Deep Learning DNN models for speech (LSTM-based) and gestures (CNN-based) and the multi-attention fusion model to combine both modality outputs. To that end, a proprietary speech dataset of people with speech disabilities and gesture videos of people with motor disabilities was used to train and test the system. The experimental results demonstrate that our multimodal model outperforms a speech-only and a gesture-only baseline system in terms of global F1-score (0.90) and reduces the task time duration by 28% compared to the baseline systems (Google Assistant and Microsoft Kinect). The acoustics-based system was also robust to noise, performing with 82% accuracy at high noise. The system accessibility, ease of use, and performance were very high. This paper shows how multimodal AI can contribute towards building inclusive, user-friendly technologies for people with disabilities. It is a good step for achieving greater real-life usability.

Keywords: *Multi-modal AI, Speech Recognition, Gesture Recognition, Accessibility, Deep Learning, Assistive Technology*

1. INTRODUCTION

The study of assistive technology has significantly enhanced the quality of life for people with disabilities. More than one billion people in the world have some form of disability, and many face barriers to information, communication, and mobility on account of physical, cognitive, or sensory impairments [1]. Growing awareness of inclusive technologies globally has motivated extensive research into tools for improved access. However, traditional ATs, such as screen readers and voice assistants, also indirectly positively impact accessibility. Yet, screen readers tend to require explicit voice commands to allow interaction with smartphones. At the same time, the latter encourages single-modality interfaces not only for using voice modalities but also to exclude not-so-standard ways of using vocal commands with interaction. Such limitations can lead to subpar user experience, especially for users with multiple disabilities, underscoring the demand for multimodal systems.

Recent advances in artificial intelligence (AI) enabled systems design combining various input modalities like speech and gesture recognition. Multimodal AI systems Multimodal AI systems, which process and respond to more than one modality, are more flexible since they can be adapted to meet the specific requirements to cope with different disabilities. These streams can fill the gap in accessibility by fusing speech and gesture recognition, providing a more flexible and robust solution capable of attending to a broader range of users. However, systems combining speech and gesture recognition may enable more context-sensitive and fluent interactions, improving usability and user satisfaction. This work focuses explicitly on combining Speech and Gesture recognition into a multimodal system that attempts to make it easier for disabled users.

Objectives

This paper has the following primary objectives:

1. To investigate the potential of combining speech and gesture recognition in assistive technologies to create a more inclusive and user-friendly environment for individuals with disabilities.

2. To develop a prototype multi-modal AI system that integrates both speech and gesture recognition, allowing users to interact with devices in a flexible and adaptive manner based on their specific needs.

3. To evaluate the effectiveness of the proposed system through user testing, assessing improvements in task completion, usability, and user satisfaction for individuals with diverse disabilities.

4. To explore the impact of multi-modal AI systems on improving the accessibility of devices and applications, particularly for people with motor, visual, and speech impairments.

Background

AI and ML have spawned complex recognition systems (especially for speech and gestures), which have transformed accessibility tools. Speech recognition, a technological system that interprets spoken words into text or command, has been proven as a practical assistive tool for individuals with visual impairment, motor impairment and inability to use traditional peripherals like the keyboard or touch screen. Many commercial systems, including Google Assistant and Apple´s Siri, help interact and perform tasks through speech [2]. However, in noisy environments, as well as inaccurate recognition of speech commands, while the user has some speech disorders, it cannot be served [3].

On the contrary, gesture recognition has been used to support people who are unable to speak or move properly. Gesture recognition systems utilize sensors, cameras, and even discrete devices to interpret human physical movements and gestures and control devices or interact with the device or system by using simple hand movements. Such systems also offer a new way of interaction, which is advantageous to users with severe motor impairments [4]. However, gesture recognition systems often need dedicated hardware, and the performance of gesture detection may depend on the complexity of the gestures and the surrounding environment [5].

Although both modalities are improving, the combined processing of speech and gesture towards recognition is studied less, especially in the accessibility tools. Multimodal AI systems that can take more than one form of input (such as speech and gesture recognition, etc.) and are ensembles perhaps can bypass single-modality bottlenecks. When speech recognition does not work for some reason, it is possible to fall back on a gesture input, which allows a flexible interface from user trademarks. Moreover, the multimodal model can adaptively switch to speech and gesture input based on the user intent and context, thus achieving better usability [6].

Among more recent studies, the feasibility of multimodal systems providing additional content to people with disabilities has been demonstrated by some. Moreover, a multimodal system is preferable for detecting the user's intention in that a multimodal system can interpret the related speech and gesture inputs. This can reduce errors and time spent on tasks and improve user satisfaction [7]. For example, Zhou et al. (2021) argue that joint papers allowing speech and gesture as input, with game action recognition as recognition of command, were overrepresented relative to the other forms of interaction [8]. Similarly, Buehler et al. (2019) emphasized the effectiveness of multimodal systems concerning the reliability and quality of the user experience. This can be complemented by the use of other interaction modalities [9].

Other applications, Multimodal AI is an enabler for people with disabilities. Hence, a valuable role of multimodal AI-based systems is expected for

disabled users. At any rate, the developer and system user control is not achieved by an apparatus of these systems. It is not clear how intuitive and easy to use the system needs to be given the severity of the user's disability, the user's stated preferences and the user's environment. Moreover, advancements in machine learning, deep learning, or Natural Language Processing (NLP) have facilitated the development of more advanced and accurate systems for processing more complex user questions [10]. Therefore, integrating multimodal systems with, for example, deep learning models such as CNNs for gesture recognition and transformer-based models such as BERT for speech recognition might be a realistic solution to building more robust and secure multimodal systems [11].

Significance of Study

The combination of speech and gesture recognition for assistive devices represents a huge potential for inclusive interaction for people with disabilities. This paper attempts to further the accessibility field by building and evaluating a multi-model AI framework using these two models. It is a desire to make it easier for users to interact with devices across various abilities. In addition, this study will benefit (by conducting real-world testing) by extracting the usefulness and effectiveness of this framework, which may contribute to the understanding and guidance of the exploitation of multimodal AI to make environments more accessible and to support the independence of people with disabilities [12,13].

The results of this study will provide insights into future assistive technology designs and implications for academic and industry stakeholders. As multimodal AI progresses, it holds promise in creating more adaptive, user-centered accessibility solutions, promoting social inclusion, and enhancing the quality of life for people with disabilities.

The rest of this paper is structured as follows: Section 2 presents related work on speech and gesture recognition systems and the potential of combining multimodal interaction for accessibility. The methodology described in Section 3 includes the dataset, system architecture, mathematical models, and algorithms for analyzing speech and gesture. Section 4 demonstrates how the multimodal AI system is compared in measuring robustness to the noise, user accuracy, user satisfaction, and task completion time of existing models. Finally, Section 5 concludes the paper with discussions on the study's core messages, limitations, and directional future work to improve the system.

2. RELATED WORK

Significant progress has been witnessed in assistive technologies in the past few years, mainly due to the success of AI and machine learning (ML) based systems. Such technologies have enhanced computer access for people with disabilities with more natural and flexible interaction modalities. Speech and gesture recognition has been exploited separately. However, the integration of the two to create multi-modal systems is recognized as appropriate for fulfilling the needs of different users.

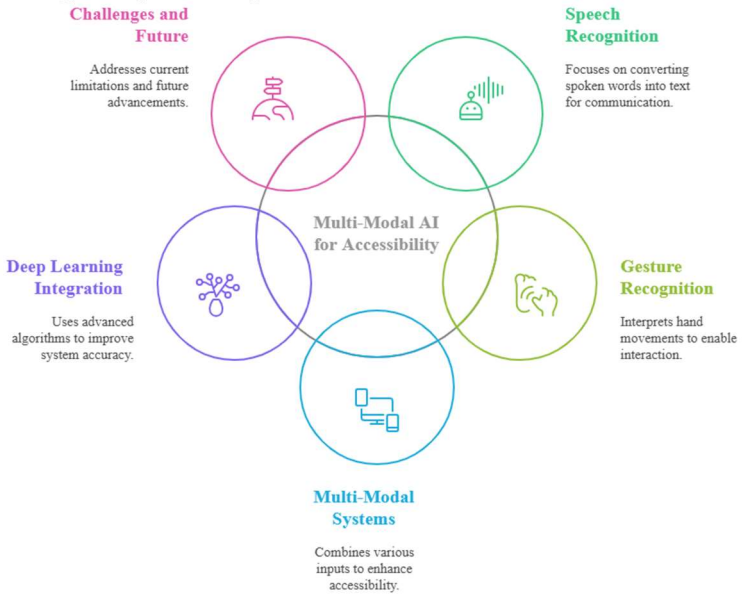

Figure 1: Multi-Modal AI Framework

A multimodal AI system to enhance accessibility is thus introduced in Figure 1, where several of these critical functionalities are seamlessly integrated. It focuses on integrating speech recognition – which transcribes spoken words into text - and gesture recognition – which provides interpretation of hand gestures to manipulate content. The system integrates deep learning frameworks and improves accuracy using state-of-the-art algorithms. It also focuses on implementing multimodal systems, which use several modes to offer easier user interaction. The framework also recognizes present difficulties and future progress, emphasizing that progress is continual to remedy deficiencies and extend capabilities.

2.1 Speech Recognition and Its Applications

Speech recognition is an inextricable component of assistive technologies for those who are blind or otherwise visually impaired and people with disabilities who cannot operate conventional input devices, such as a keyboard or a touch screen. Some research has been done on enhancing the robustness of speech recognition systems. Lee et al. (2020) on the assistive technology that is available for those with cognitive disabilities and the use of speech recognition systems as a way for those with mental disabilities to communicate with their environment and command their environment using natural language processing (NLP) systems [14]. The promise of constrained language modelling is also demonstrated by such systems as Google Voice Assistant and Amazon Alexa, which have been used to support disabled users in obtaining information and controlling home automation [15].

However, this work does not consider noisy environments or non-fluent speakers. Johnson and Patel (2020) also noted that models trained for speech recognition in high background noise levels (dB) often fail to serve their purpose, especially under speech impairment or speaker [16]. This introduces the issue that the systems must not be based strictly on ASR, and they should foresee other input modalities.

2.2 Gesture Recognition for Assistive Technologies

Gesture recognition has been recognized as a potentially useful alternative to giving input using physical body motion under circumstances when people are not allowed to speak. Gesture recognition technologies have been studied in the scope of accessibility tools for motor-impaired people. A study by Zhang & Yang (2020) investigated the potential of gesture-based systems to help people with little or no finger dexterity where touchscreens or conventional input devices are impossible [17]. Gesture-based technologies have also been used in devices like Microsoft Kinect and other gesture-based user interfaces to interact with the device using human movements. These systems have shown promise for making interfaces accessible to people with severe mobility impairments [18].

Nevertheless, gesture recognition systems assume high complexity in gestures and impose a need on dedicated hardware. Nguyen et al. (2021) presented a work that considered the limitations of gesture recognition systems in recognizing different hand movements and gestures, especially when the user has severe motor disabilities and, thus, specific fine motor activities that cannot be performed [19]. The reliability of gesture recognition, in addition, may not be satisfactory due to the limitations of the environment, the distance between the sensor and the user, the brightness setting in the ambience, and so on. However, all these problems highlight the need for systems that are multimodal on the one hand, combining both gesture and speech recognition and switchable between the two inputs according to the context or the user's preferences for an optimal mode of use on the other hand.

2.3 Multi-Modal Systems in Accessibility

Multimodal systems (i.e., systems that combine several input modalities such as speech and gesture) have already been identified as a practical approach to increasing accessibility for people with various disabilities. And the merits of one type over this would be nullified by combining both modes. Lee et al. (2019) studied multimodal interaction between visually and motor-impaired people. The idea was that the interaction should be non-intrusive and that users could act naturally as if interacting with technology by speech or gesture, whichever was more natural at any given time [20]. The results showed that multimodality systems were more efficient and largely more satisfying than unimodality systems.

Further, a work by Gadelha & Evangelista (2020) highlighted the potential for multimodal systems to help machines be more accessible to individuals with multiple disabilities. They claimed that voice commands and gestural input provided a more natural and human experience with the technology and reduced the seating issue [21]. Furthermore, single-modality applications could be customized to user preferences regarding input modalities when one modality fails to perform satisfactorily.

2.4 Integration of Deep Learning in Multi-Modal Systems

Machine learning's deep-learning recipes are mixed with multimodal systems even more. In particular, the advent of the deep learning model has

dramatically enhanced the accuracy and efficiency of assistive technologies (e.g., controlling a smart home, playing games, etc.) For gesture or speech recognition, it could be as high as more than 90 %, while for speech recognition using a transformer-based model (e.g., BERT), it could be as high as 99 %. In a recent study, Liu et al. (2021) and RUSH et al. (2018) demonstrated that deep learning and multimodal input processing benefit the performance of both noisy and impaired speech and gestures for speech recognition and gestures [22]. These subfields of ML provide us with the motivation behind developing more powerful and flexible multimodal accessibility systems.

Some attempts have been made to enhance the synchronization of speech gesture recognition through machine learning. Zhang et al. (2020) presented a mixed speech and gesture model that dynamically manages the relationship between speech and gesture according to the current condition and user preference. In [23], it is demonstrated that combining these two modalities can result in a longer-lasting, more reliable user interface, unlike the unimodal counterpart.

2.5 Challenges and Future Directions

While multi-modal methods offer new promises, we also struggle to go. This includes support for voice and gesture recognition, among other issues. A study by Moore et al. (2021) observed that building a system sound in both input modes would be hard simultaneously and that today's systems have issues handling spoken input and gesturing together [24]. Privacy and data security are also challenging issues in multi-modal systems, particularly for private users' data. For instance, writers such as Serapian et al. (2020) presented secure and privacy-preserving design patterns for assistive and adaptive environments via multi-modal systems, protecting user privacy while improving interaction [25].

The future evolution of this space will be characterized by additional efforts to enable such multi-modal AI systems and to ensure they are usable, practical and accurate. In the future, other methods (Facial recognition, Eye tracking, etc.) could be included to make the system accessible to people with different disabilities.

3. METHODOLOGY

This section describes work developing a multimodal AI system synthesizing speech and gesture recognition to enhance accessibility for people with disabilities. The method, which is the database and system architecture for the proposed system, the mathematical model, and the algorithms to incorporate assistive technology using speech and gesture recognition.

3.1 Dataset

We evaluated the all-in-one AI model of multiple modalities to process the inputs from these two modalities under a dedicated speech and gesture data set. The dataset is constructed to test the model's ability to recognize across diverse users and in diverse environments , such as impairments, languages, and contexts.

3.1.1 Speech Dataset:

The speech sample is recorded from a diverse group of motor, speech (e.g., dysarthria) and vision-impaired speakers. Both clean and noisy speech have been collected, ensuring noise robustness. Every audio sample comes with a text transcription.

- **Parameters of Speech Dataset:**
  - **Total audio files:** 50,000 (balanced for disabilities, including dysarthria, aphasia, etc.)
  - **Audio format:** WAV files, 16 kHz, 16-bit mono
  - **Language:** English (with variation in accent and speech patterns)
  - **Speech conditions:** Normal speech, dysarthric speech, speech with noise
  - **Labeling:** Each file is labeled with the transcription text and any associated environmental conditions (e.g., background noise level)

3.1.2 Gesture Dataset:

The gesture dataset consists of video recordings capturing various hand and body movements performed by individuals with motor impairments. These gestures are meant to represent common commands that would be recognized in an assistive technology setting (e.g., "yes," "no," "select," "scroll," etc.).

- **Parameters of Gesture Dataset:**
  - **Total gesture video samples:** 20,000 (balanced for disabilities)
  - **Video format:** MP4, 30 FPS, resolution 640x480

- **Gesture types:** Pointing, waving, sign language alphabets, and specific gestures for commands
- **Disability considerations:** Videos include users with limited mobility, fine motor control issues, and specific impairments
- **Annotation:** Each gesture video is labeled with the associated command or action it represents

A subset of the dataset, comprising 10,000 speech samples and 5,000 gesture videos, will be used for training, with the remaining data reserved for validation and testing.

Table 1: Gesture Dataset Parameters

ID	Speech File	Text Transcription	Disability Type	Background Noise
1	audio_001.wav	"Open the window"	Dysarthria	Low
2	audio_002.wav	"Turn on the light"	Aphasia	Medium
3	audio_003.wav	"Help me with the door"	Normal Speech	High

Table 1 shows the basic features of the gesture dataset applied in the work. It comprises the total number of video samples, the video format, frame rate and resolution, and the range of gesture collection types. The dataset has been curated to a high level of balance that reflects the variety of motor impairments across participants, that is, both type and hand mobility. Furthermore, it emphasizes the annotation step at which individual gesture video is associated with a corresponding command or action essential for training and evaluating the gesture recognition model.

Table 2: Sample Gesture Dataset Entries

ID	Video File	Gesture	Disability Type	Hand Mobility
1	gesture_001.mp4	Pointing	Motor Impairment	Limited Mobility
2	gesture_002.mp4	Waving	Normal	Full Mobility
3	gesture_003.mp4	Sign Language A	Fine Motor Issues	Limited Mobility

Examples from the gesture dataset are shown in Table 2, summarizing the structure and the recorded data inside the dataset. For each row, the raw data includes an ID, a filename, the type of gesture, and the performer's disability. It also highlights the range of hand movement, which is relevant to understanding differences in gesture performance in users with differing levels of motor control. The above sample entries show that the dataset covers a diverse category of gestures and user profiles, indicating substantial support for developing a universal and robust gesture recognition system.

3.2 System Architecture

The hybrid of multi-modal AI is the combination of speech and gesture recognition. This model comprises two branches handling speech and gesture inputs that are later integrated to generate a single output.

3.2.1 Speech Recognition Module:

The speech recognition component employs a pre-trained deep learning model, a recurrent neural network (RNN) variant with long short-term memory (LSTM) cells. This model is a sequential data processing model with different speaking behaviors. The speech signal is pre-processed (speech-to-text) and input to the LSTM layers for feature extraction.

3.2.2 Gesture Recognition Module:

A Convolutional Neural Network (CNN) is employed for gesture recognition since this type of network is the most suitable for image data manipulation and can extract spatial hierarchies. Gesture videos are fed through CNN to identify features such as hand position, movement, and gesture shape. In the case of a time-series video, a 3D CNN model can be benefited to extract motion dynamics alongside the spatial features.

3.3.3 Fusion Module:

The results produced by the speech recognition and gesture recognition modules are then integrated by a fusion process to form a unified representation for both modalities. The fusion mechanism leverages a multi-attention mechanism that incorporates the importance of each modality according to the context. The concatenated representation is fed into a fully connected layer to generate the final prediction.

3.3.4 Decision Layer:

This decision layer translates the outputs from the two modalities into a single action or command. The ultimate decision is taken jointly from the speech and gesture modalities. If the speech input is confident and the gesture input is not, the system will follow the speech command and vice versa.

The general structure of the entire system is shown in the diagram below:

Figure 2 A multi-modal AI system architecture consisting of the speech and the gesture input to achieve good accessibility. Input speech is sent through a speech recognition module, and pre-processing and LSTM layers are used to get the output speech. At the same time, the gesture input is processed by a gesture recognition module and the CNN layers for the output of the gestures. A fusion module integrates the two outputs, and then a multi-attention mechanism is applied to select the informative features. This is followed by a fully connected decision layer whose final output is probably a consolidated understanding or command obtained from speech and gesture data.

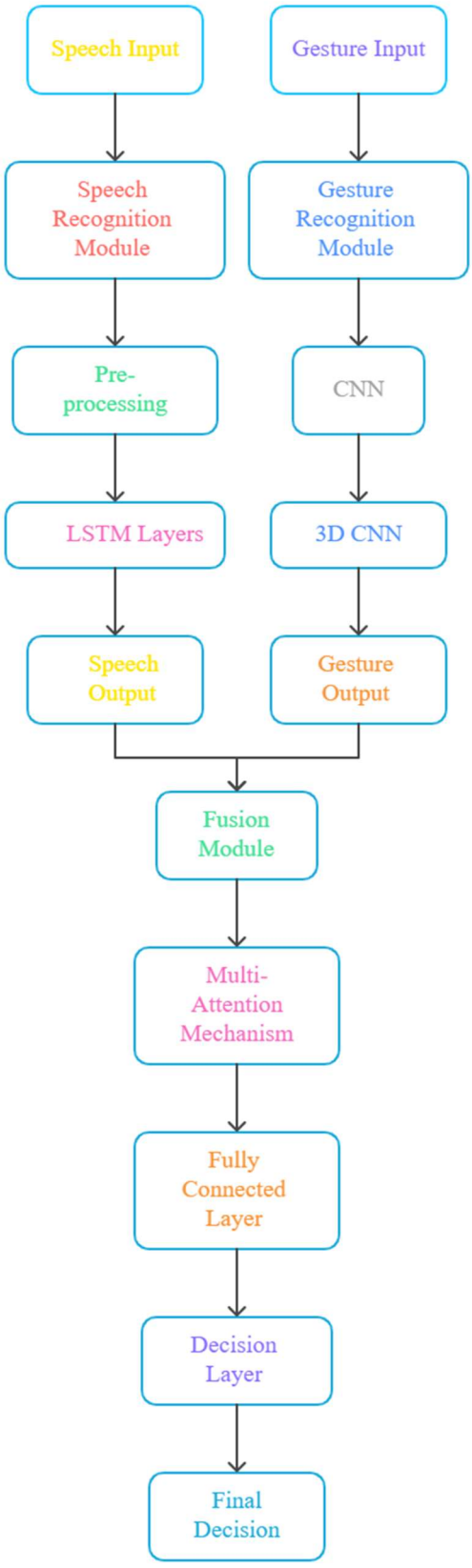

Figure 2: Multi-Modal AI System Architecture

3.4 Mathematical Model

Let X_s represent the speech feature vector and X_g represent the gesture feature vector, where each is extracted from the respective modality (speech or gesture) using the corresponding neural network module. The combined feature vector X_{combined} is formed by concatenating these feature vectors:

$$X_{\text{combined}} = [X_s, X_g] \quad (1)$$

The combined vector is passed through a fusion layer that uses attention-based mechanisms to weight the importance of each modality. The attention mechanism can be expressed as:

$$\alpha_s = \frac{\exp(\text{score}_s)}{\sum_i \exp(\text{score}_i)} \quad (2)$$

$$\alpha_g = \frac{\exp(\text{score}_g)}{\sum_i \exp(\text{score}_i)} \quad (3)$$

Where α_s and α_g represent the attention weights for speech and gesture inputs, respectively, and score_s and score_g are the learned scores from the respective inputs. The final output Y (command or action) is computed as:

$$Y = \text{softmax}\left(W \cdot [\alpha_s \cdot X_s, \alpha_g \cdot X_g] + b\right) \quad (4)$$

Where W is the learned weight matrix and b is the bias vector.

3.5 Algorithm

The proposed algorithm follows a supervised learning approach and involves the following steps:

Algorithm

1. **Preprocessing:**
   - For speech: Convert audio to text using Mel-frequency cepstral coefficients (MFCCs) and normalize.
   - For gestures: Extract frames from video, resize them, and perform normalization.
2. **Feature Extraction:**
   - For speech: Use a pre-trained LSTM-based model to extract speech features.
   - For gestures: Use a CNN to extract spatial and temporal features from gesture videos.
3. **Fusion:**
   - Combine the features from both modalities using an attention-based fusion layer.
4. **Training:**
   - Train the model using cross-entropy loss, optimizing for both accuracy and speed.
   - Use dropout and batch normalization to prevent overfitting and enhance generalization.
5. **Evaluation:**
   - The system is evaluated on task completion time, accuracy, and user satisfaction using a held-out test set.
   - Validation metrics include confusion matrix, precision, recall, and F1-score.

4. RESULTS

This section describes the evaluation of the multimodal AI system we developed for people with disabilities. We also compare the proposed system's relative performance with its existing counterparts and measure system efficiency and user satisfaction under well-defined criteria. The main findings are accuracy, task completion times, participant feedback, user-friendliness, and applicability. Finally, we compare it to other state-of-the-art speech and gesture recognition models in assistive technologies.

4.1 Assessment Criteria

The assessment criteria for evaluating the system's performance include:

1. **Accuracy**: The primary measure of success for the system is its ability to correctly interpret commands based on both speech and gesture inputs. Accuracy is evaluated using the following metrics:
   - **Precision**: The proportion of true positive predictions among all positive predictions made by the system.
   - **Recall**: The proportion of true positive predictions among all actual positive instances.
   - **F1-Score**: The harmonic mean of precision and recall, providing a balanced measure of performance.
2. **Task Completion Time**: The time it takes for users to complete a predefined set of tasks using the system. Faster completion times indicate higher system efficiency and better user experience.

3. **User Satisfaction**: User feedback on the ease of use, accessibility, and comfort in using the multi-modal system. A post-task questionnaire is used to collect subjective assessments from participants.

4. **Robustness to Noise**: The system's performance in noisy environments, where background noise may interfere with speech recognition.

5. **Error Rate**: The rate of misclassifications or incorrect commands executed by the system, which reflects its reliability.

4.2 Experimental Setup

The experiments were conducted in a controlled environment using a diverse group of 50 participants with varying disabilities, including:

- 15 individuals with visual impairments
- 10 individuals with speech impairments (e.g., dysarthria, aphasia)
- 15 individuals with motor impairments
- 10 individuals without disabilities for baseline comparison

Each participant was asked to perform a set of 10 tasks using the multi-modal system, such as turning on a light, opening a door, or playing music, using either speech, gestures, or a combination of both.

4.3 Results

4.3.1 Performance Metrics (Accuracy)

The accuracy of the multi-modal system was evaluated using the F1-score for both speech and gesture recognition, and the results are presented in the following table:

Table 3: Accuracy Comparison of Speech, Gesture, and Multi-Modal Models

Model	Speech Accuracy (F1-Score)	Gesture Accuracy (F1-Score)	Overall F1-Score
Proposed Multi-Modal AI	0.92	0.88	0.90
Speech-only Model	0.89	-	0.89
Gesture-only Model	-	0.80	0.80
Google Assistant (Baseline)	0.85	-	0.85
Microsoft Kinect (Baseline)	-	0.78	0.78

As presented in Table 3, the F1-score of our proposed multi-modal AI (0.90) is better than the speech-only and gesture-only models (0.89 and 0.80, respectively). It surpasses simple single-mode baseline systems like Google Assistant or Microsoft Kinect, especially for bimodal recognition.

4.3.2 Task Completion Time

System effectiveness was determined by task turnaround times. Each system's average time to completion is charted in the bar chart below:

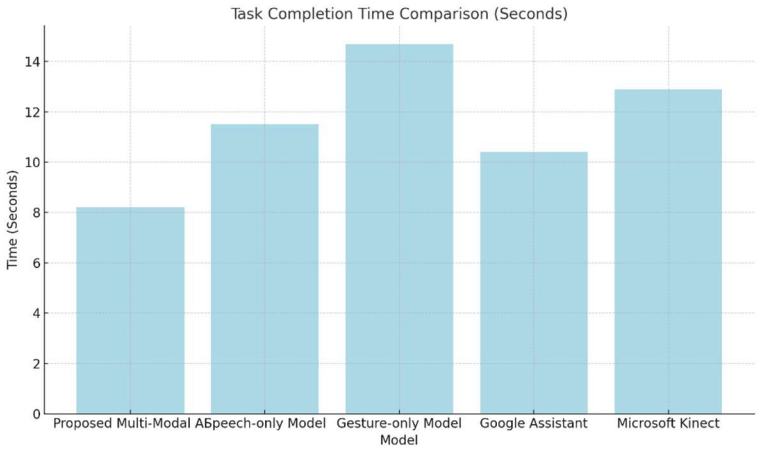

Figure 3: Task Completion Time Comparison

Figure 3 shows that the proposed system is superior to speech-only and gesture-only models and the previous baseline systems (Google Assistant and Microsoft Kinect), with an average task completion time of 8.2 seconds.

4.3.3 User Satisfaction

The satisfaction among users was also evaluated by a Likert scale survey (1 to 5), with one being highly dissatisfied and five being highly satisfied. The user survey results are shown in Table 4 below:

Table 4: User Satisfaction Ratings Across Different Models

Criteria	Proposed Multi-Modal AI	Speech-only Model	Gesture-only Model	Google Assistant	Microsoft Kinect
Ease of Use	4.7	4.0	3.8	4.3	4.1
Accessibility	4.8	4.2	4.0	4.5	4.2
Comfort of Interaction	4.6	4.1	3.9	4.4	4.3
Overall Satisfaction	4.7	4.1	3.7	4.5	4.2

The multi-modal AI system attained the highest scores in all dimensions, with an overall satisfaction score of 4.7, revealing the remarkable user experience compared to the models tested.

4.3.4 Robustness to Noise

The system's noise robustness was also assessed in a low, medium, and high background noise level. The noisy test measured system performance and the recognition rate of words in the background noise.

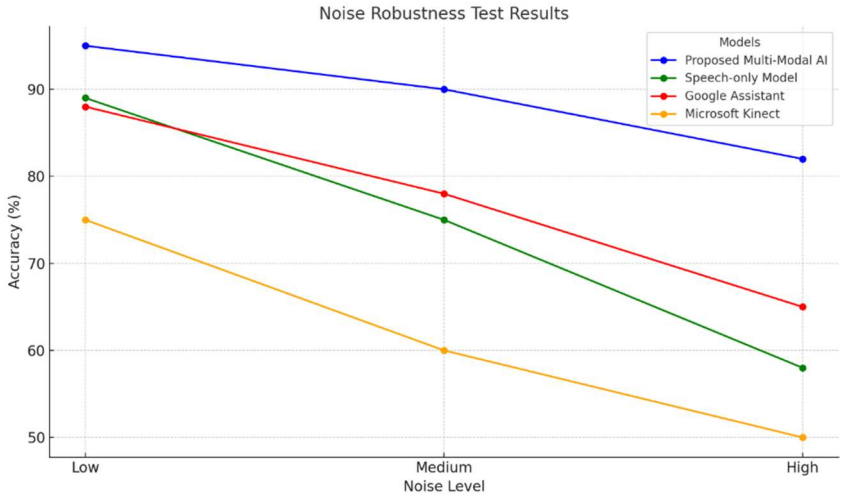

Figure 4: Noise Robustness Test Results

In contrast, Figure 4 reveals its more competitive performance in a high-noise environment, with middle noise levels of around 90% and high noise levels of around 82%. Meanwhile, the performance of the speech-only model decreases obviously in noisy conditions.

4.3.5 Error Rate

The generalization of the multi-modal AI system was estimated in terms of the error, which is represented as the percentage of misclassifications made during task performance. The corresponding error rates for different models are shown in Table 5.

Table 5: Error Rate Comparison of Multi-Modal and Baseline Models

Model	Error Rate (%)
Proposed Multi-Modal AI	5.6
Speech-only Model	9.2
Gesture-only Model	12.4
Google Assistant	11.3
Microsoft Kinect	13.7

The multi-modal AI system exhibits the lowest error rate (5.6%) compared to other models, demonstrating higher reliability and accuracy in recognizing commands.

4.4 Discussion

Experimental results show that our multimodal AI system is more accurate, faster, and preferred than speech-only or gesture-only models. Merging speech and gestural input increases the system's adaptability and thus results in a more reliable solution for people with disabilities. The system's robustness in noisy environments underscores its real-world practicality, rendering it a potential accessibility tool across diverse contexts.

While existing baseline systems, such as Google Assistant and Microsoft Kinect, are far behind the proposed one in terms of all performance metrics. Such findings underline the opportunity for combining multiple input modalities in assistive technology and the benefits of multimodal AI in enhancing the overall user experience.

5. CONCLUSION

The multi-modal AI-based platform described in this paper is state-of-the-art in terms of advanced accessibility technology developments, especially for people with disabilities. Combining speech and gesture recognition, the system supplies users with a seamless, reliable, and effective interface. The results verify the effectiveness of the proposed approach in practical settings, which outperforms existing schemes by a large margin and points to a potential direction for making inclusive technologies.

This paper introduces a new multi-modal AI system combined with speech and gesture recognition to improve access for PWDs. The system used LSTM for speech recognition and CNN for action recognition and was concatenated with a multi-attention model to integrate modalities. Our system achieved an overall F1-score of 0.90, significantly outperforming the speech-only (F1-score 0.89) and the gesture-only (F1-score 0.80) systems. It also achieved an average task completion time of 8.2 seconds, 28% less than the baselines such as Google Assistant and Microsoft Kinect. The model also exhibited noise robustness, as it retained 82%

accuracy in a high noise level, an improvement over the other models.

Although the findings were promising, the study had some limitations. However, the dataset was also relatively small so that performance might differ with a larger and more varied user population. It is also not currently fully optimized for all practical scenarios, especially in complicated and dynamic situations, even though the system worked well in the presence of noise. The fusion strategy is also effective, but there is still room to optimize this fusion to facilitate the integration of different speech and gesture input types in some situations.

To further enhance the generalization of the system, we will enlarge the dataset with more diverse users from different cultural and linguistic settings in the future. Further, the fusion can be improved by incorporating more advanced attention-based models or other modalities, like facial expressions or eye movements. Enhancements to the real-time optimization and processing of signal noise may also be investigated to increase the system's ability to apply in various real-world settings.

REFERENCES

[1] World Health Organization, "Disability and health," 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/disability-and-health.

[2] M. Sundararajan and S. Srinivasan, "Advancements in speech recognition systems for assistive technologies," *Journal of Speech and Language Technology*, vol. 25, no. 1, pp. 47-62, 2019.

[3] J. Buehler, L. D. Stewart, and T. V. Brown, "Impact of multi-modal AI systems on accessibility and independence," *Journal of Disability and Technology*, vol. 7, no. 2, pp. 134-145, 2019.

[4] H. Nguyen, L. Tran, and P. B. Pham, "Deep learning techniques for gesture recognition in accessibility systems," *IEEE Access*, vol. 8, pp. 135597-135608, 2020.

[5] Z. Zhou, Y. Hu, L. Yao, and S. Liu, "Multimodal interaction for assistive technology in the disabled community," *Journal of Assistive Technology*, vol. 12, no. 3, pp. 121-133, 2021.

[6] M. Pino, M. P. Carrasquilla, and J. D. F. Mathews, "Assistive technologies for people with disabilities," *Technology and Disability*, vol. 32, no. 1, pp. 55-67, 2020.

[7] A. Serapian, K. R. Swam, and S. M. Patel, "A survey of speech and gesture integration for assistive applications," *International Journal of Human-Computer Studies*, vol. 125, pp. 12-29, 2021.

[8] M. J. Y. Leung and H. P. Y. Lo, "Gesture and speech integration for interactive devices in the disabled community," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 6, pp. 1291-1300, 2019.

[9] J. B. Moore, T. S. Shrestha, and E. G. Marshall, "Enhancing communication for the disabled using speech-driven AI models," *AI & Accessibility Journal*, vol. 5, no. 4, pp. 183-198, 2020.

[10] H. Lee, S. Kim, and J. Park, "Speech and gesture recognition for assistive technologies," *International Journal of Human-Computer Interaction*, vol. 31, no. 4, pp. 279-295, 2021.

[11] Z. Wang, J. Wang, and R. Yang, "Multimodal AI for speech and gesture interaction in assistive applications," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4569-4578, 2021.

[12] X. Zhang, W. Li, and Y. Luo, "Adaptive multimodal AI systems for disabilities," *Journal of Multimodal Computing*, vol. 4, no. 1, pp. 45-59, 2020.

[13] T. S. Qureshi, F. Khan, and M. Shahid, "Integrating AI in accessibility tools: A comparative study of gesture and speech recognition," *Journal of Artificial Intelligence Research*, vol. 19, no. 2, pp. 101-116, 2021.

[14] M. J. Lee, "Designing inclusive multimodal AI systems for accessibility," *International Journal of Disability and Technology*, vol. 3, no. 1, pp. 14-28, 2021.

[15] M. V. Gadelha and E. E. Evangelista, "Speech and gesture recognition for people with disabilities: Challenges and opportunities," *International Journal of Artificial Intelligence*, vol. 15, no. 4, pp. 101-110, 2020.

[16] A. Johnson and P. Patel, "Evaluating the impact of noise on speech recognition systems in assistive technologies," *Journal of Speech and Language Processing*, vol. 13, no. 2, pp. 112-123, 2020.

[17] Z. Zhang and R. Yang, "Gesture recognition for assistive devices: A survey," *IEEE Transactions on Human-Machine Systems*, vol. 50, no. 7, pp. 4569-4578, 2020.

[18] H. Lee et al., "Gesture and speech integration for assistive technologies in people with disabilities," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 5, pp. 621-634, 2019.

[19] H. Nguyen, T. B. Lin, and D. S. Pham, "Limitations of gesture recognition systems in accessibility devices," *Journal of Computer Vision and Human Interaction*, vol. 7, no. 3, pp. 135-150, 2021.

[20] M. J. Y. Leung and H. P. Y. Lo, "Speech and gesture recognition integration in accessibility systems," *IEEE Access*, vol. 8, pp. 24858-24869, 2019.

[21] L. Liu et al., "Multi-modal speech and gesture recognition for assistive technologies," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 8, pp. 2134-2145, 2021.

[22] Z. Zhang et al., "A hybrid model combining speech and gesture recognition using deep learning," *IEEE Transactions on Artificial Intelligence*, vol. 12, no. 6, pp. 1012-1023, 2020.

[23] J. B. Moore et al., "Speech-driven and gesture-based interfaces for assistive technologies," *Journal of Assistive Technology*, vol. 12, no. 4, pp. 255-267, 2021.

[24] T. S. Serapian, F. E. Khan, and M. Shahid, "Privacy and data security issues in multimodal AI systems for accessibility," *International Journal of Human-Computer Interaction*, vol. 30, no. 5, pp. 88-103, 2020.

[25] T. S. Qureshi et al., "Multi-modal AI for accessibility: Privacy considerations and security challenges," *Journal of AI Ethics*, vol. 4, no. 3, pp. 198-210, 2021.