Sensors and Materials, Vol. 3#, No. # (202#) ##–##
MYU Tokyo

S & M ****

AI-Driven AAC System for Communication Enhancement in Cerebral Palsy Using ASL Recognition

Yuh-Shihng Chang,[1] Chen-Wei Lin,[1] and Chien-Chih Chen[1*]

[1]Department of Information Management, National Chin-Yi University of Technology, Taiwan, ROC.

Keywords: Augmentative and Alternative Communication, Cerebral Palsy, Artificial Intelligence, Machine Learning, American Sign Language

Augmentative and Alternative Communication (AAC) methods are designed to facilitate effective communication for individuals with speech, language, or writing impairments. This research introduces a novel AAC system designed to enhance communication for individuals with speech, language, or writing impairments, such as those with cerebral palsy (CP). Addressing the limitations of existing AI-driven AAC solutions, the system integrates advanced AI and machine learning (ML) techniques to minimize user effort in conveying their thoughts. Utilizing a readily available webcam as a key input modality, the system employs MediaPipe to capture hand gestures corresponding to American Sign Language (ASL) alphabets. The resulting visual data is then processed and classified by a Random Forest model. By interpreting these sensor-captured gestures, the system enables users to input partial vocabulary, which subsequently prompts generative AI models to predict and complete intended text. Empirical evaluations, conducted through two distinct experiments, validate the system's viability and demonstrate its potential to significantly improve communication accessibility for individuals with CP through an accessible and intuitive gesture-based interface.

1. Introduction

Augmentative and Alternative Communication (AAC) methods aim to facilitate information exchange for individuals with speech, language, or writing disabilities, either by enhancing speech communication or providing alternative written communication.(1) Researchers are continuously developing new AAC technologies, including sign languages(2) and communication boards,(3) to improve communication effectiveness. In recent years, the integration of artificial intelligence (AI) has led to the development of AI-based AAC methods. However, personalized AAC solutions remain a challenge.

1.1 Research background

AAC systems are broadly classified into unaided and aided categories. Unaided systems utilize the user's body language, natural gestures, manual signs,(4) and facial expressions, requiring no external devices. Their effectiveness depends on the receiver's familiarity with the specific system. Sign languages, a highly structured form of unaided AAC, are considered among the most

[*]Corresponding author: e-mail: frick@ncut.edu.tw

https://doi.org/10.18494/SAM.2022.####

developed. In contrast, aided AAC systems introduce external devices to facilitate communication. Traditional examples include communication boards featuring pictures, symbols, or words,[5] letter boards, and writing tools.

Over the past decade, AI techniques have revolutionized the AAC research domain. Individuals with communication disabilities can now utilize motion recognition applications to select symbols or words[5] for speech-generating devices, enabling synthetic speech. These technologies commonly incorporate features such as text prediction, vocabulary customization, and personalized voice options.

Individuals with cerebral palsy (CP), who experience neuromuscular disorders affecting movement and posture, often face challenges with speech production and fine motor control. Motor impairments can hinder their ability to control their hands, arms, or mouth, necessitating innovative approaches to access communication tools. Fortunately, the field of AAC offers a variety of access methods that have proven particularly beneficial for the CP community.[6]

For individuals with CP using AAC, both access methods and naming systems require careful consideration. Visual-cognitive challenges associated with CP can impede symbol recognition, necessitating the selection of symbol systems based on maximum intelligence and equivalence principles. Suitable options for individuals with moderate literacy include photographs, realistic color pictures, simplified pictographic symbols, and text.[5]

1.2 Research Motivation

Direct access to technology remains a significant challenge for individuals with CP. Muscle tone fluctuations, involuntary movements, and difficulties in maintaining device positioning can hinder consistent equipment operation. These motor challenges, which vary throughout the day and with emotional states, necessitate dynamically adaptable access systems. Some users struggle with targeting and switch activation, while others experience fatigue during prolonged communication attempts, leading to communication endurance issues.

Early research once introduced Dasher[7], a data entry interface combining continuous gestures and language models to assist motion-impaired computer users. Dasher facilitates continuous text generation with minimal physical effort, reducing the precision required for effective communication.

The rapid progress in computer vision and deep learning has facilitated the emergence of sophisticated Human Pose Estimation (HPE) algorithms, with OpenPose and MediaPipe being prominent examples. For instance, the study[8] demonstrated the application of the MediaPipe Gesture Recognizer[9] in creating a hand gesture recognition system. Consequently, the development of effective communication tools for individuals with CP has become increasingly feasible.

Many developers of mobile phones, laptops, and other communication devices have integrated large language models (LLMs) based on GPT (Generative Pre-trained Transformer) architectures to significantly enhance word prediction. Devices previously relying solely on frequency and recency now leverage semantic context to provide users with more accurate predictions, including phrases and responses, thereby reducing the number of required selections[10].

To address communication challenges faced by individuals with cerebral palsy (CP), this study presents a practical AI-based AAC system. The system employs a three-stage process: handshape recognition via MediaPipe Gesture Recognizer[9] and Random Forests[11] classification, text prediction using generative AI (GAI) to generate the top five sentence completions, and speech synthesis utilizing the pyttsx3 Python library. Empirical evaluations, conducted through two distinct experiments, validate the system's effectiveness and its potential to significantly improve communication access for individuals with CP.

The subsequent sections of this paper are structured as follows: Section 2 outlines the foundational elements of the proposed method, Section 3 provides a detailed explanation of the method itself, Section 4 presents the empirical evaluation, and Section 5 concludes the paper with a summary of findings.

2. Related Work

This section outlines the foundational elements of our method: the American Sign Language (ASL) alphabet, MediaPipe's Gesture Recognizer, and an overview of Generative Artificial Intelligence (GAI).

2.1 American Sign Language: alphabets

While American Sign Language (ASL) utilizes a manual alphabet, where each letter of the English alphabet is represented by a specific handshape, it's essential to understand that this alphabet is not the core of ASL. ASL is not simply English spelled out with hand gestures. Instead, it is a fully developed, independent language with its own grammar and syntax, distinct from English, as shown in Figure 1.

The ASL manual alphabet, sometimes referred to as fingerspelling, is primarily used for specific purposes, such as spelling proper nouns, clarifying words with no established sign, and spelling technical terms or loan words. It's important to note that the sequential nature of fingerspelling contrasts sharply with the simultaneous expression of meaning characteristic of most ASL signs. ASL signs convey concepts through a combination of handshape, location, movement, palm orientation, and non-manual markers (facial expressions and body language). This multi-faceted approach to communication highlights the complexity and richness of ASL beyond simple letter-to-hand correspondence. While the manual alphabet serves as a valuable tool within ASL, it represents only a small component of the language's overall structure and usage. ASL's independence stems from its unique linguistic framework, which is rooted in visual-spatial communication and historical development.

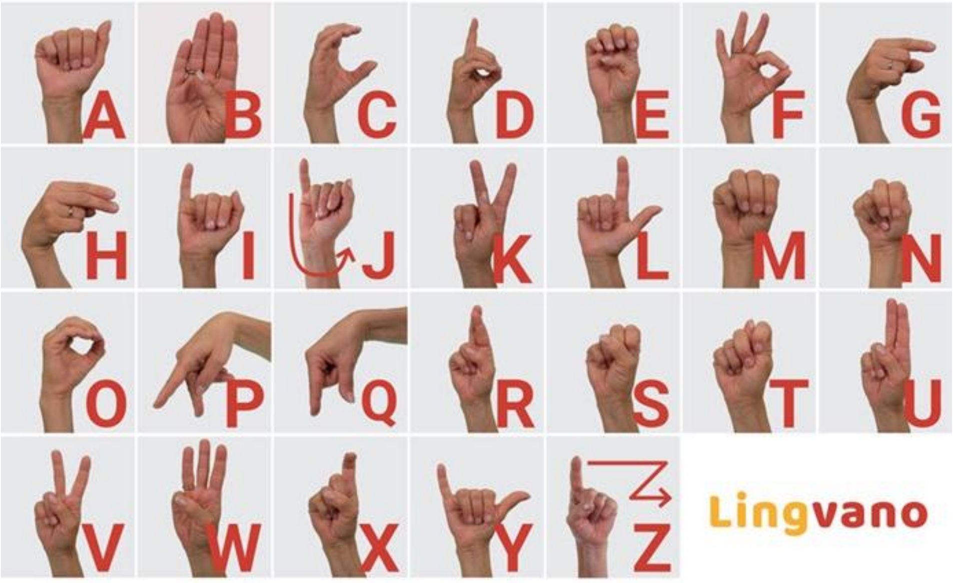


Figure 1. The American Sign Language manual alphabets(12)

2.2 MediaPipe Gesture Recognizer

Developed by Google, the MediaPipe Gesture Recognizer utilizes the MediaPipe framework for perception tasks, uniquely focusing on interpreting hand movement language through multi-step processing that identifies 21 specific key landmarks on the hand, mirroring skeletal points and tracking finger joints, knuckles, and palm movements, to classify gestures like thumbs up or open palm [8]. This technology, as illustrated by the hand landmarks in Figure 1, offers significant potential for accessibility by enabling real-time sign language translation, fostering inclusive communication between hearing and hearing-impaired individuals in the digital realm. In comparison to other HPE algorithms like OpenPose, MediaPipe Gesture Recognizer excels in real-time performance and mobile-friendly deployment due to its optimized models and lightweight architecture, providing a distinct advantage in competitions requiring fast inference speeds. However, while OpenPose often delivers higher accuracy and more detailed body pose estimations, MediaPipe Gesture Recognizer's focus on hand-specific gestures might limit its applicability in broader HPE scenarios that demand comprehensive full-body tracking.

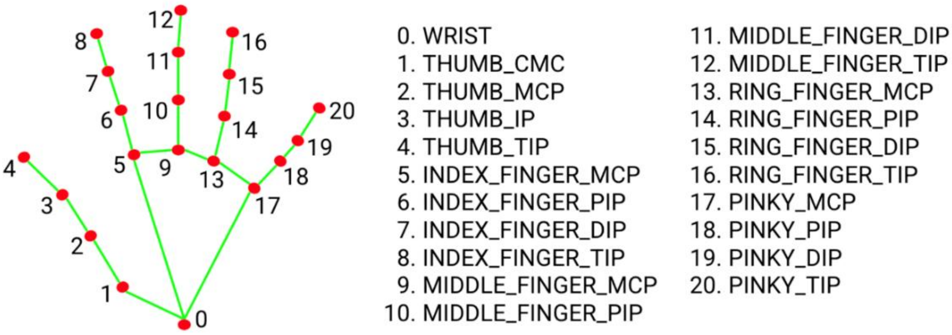


Figure 2. Hand keypoints defined in MediaPipe(9)

2.3 Generative artificial intelligence

The evolution of Generative AI (GAI), as shown in Figure 3, is a narrative of progressive innovation, commencing with foundational Early AI & ML (machine learning) Research that laid the groundwork for subsequent advancements. Initially, AI development was characterized by Statistical Methods, which employed probabilistic models and rule-based systems to analyze and interpret data. This phase transitioned into the era of Deep Learning, marked by the advent of neural networks capable of learning intricate patterns from vast datasets, thereby enabling significant improvements in tasks like image recognition and natural language processing. The emergence of Early Generative Models, such as Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs), represented a pivotal shift, allowing machines to generate novel content. For example, GAN has been utilized to solve the class-imbalanced learning issue.(13) However, the paradigm truly transformed with the introduction of Transformer & LLMs, or Transformer architectures and Large Language Models, exemplified by models like GPT, which utilized attention mechanisms to produce coherent and contextually relevant text. Concurrently, Diffusion Models, as seen in systems like DALL-E and Stable Diffusion, revolutionized image generation by iteratively refining random noise into detailed visuals. These technological breakthroughs culminated in the GAI Explosion, a period of rapid proliferation and adoption of generative AI across diverse applications. This trajectory, from the nascent stages of AI research to the current era of widespread GAI utilization, underscores the dynamic and transformative nature of artificial intelligence, highlighting its increasing capacity to create content that mirrors human creativity and understanding(14).

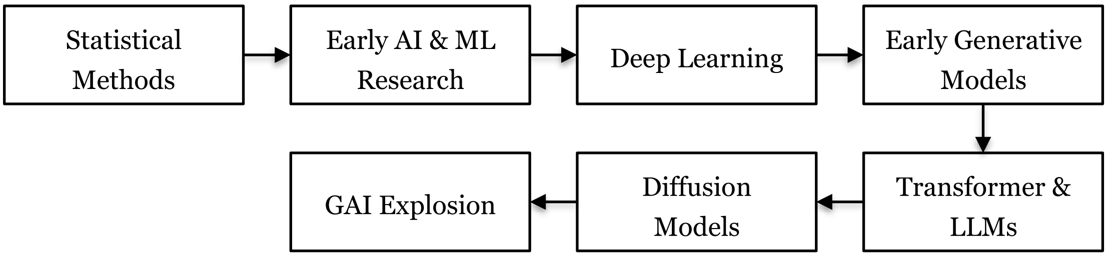


Figure 3. Evolution of Generative AI

3. Proposed Method

The proposed system, illustrated in Figure 4, comprises three distinct stages: handshape recognition, text prediction, and speech synthesis. During handshape recognition, MediaPipe is utilized to extract hand landmark coordinates, which are subsequently normalized and used to train a Random Forest classifier for letter identification. In the text prediction stage, generative AI (GAI) models are employed to generate complete sentence hypotheses based on the recognized letter sequence. The top five most probable sentence completions are then presented to the user. Finally, in the speech synthesis stage, the user-selected sentence is converted into audible speech using a Python text-to-speech library.

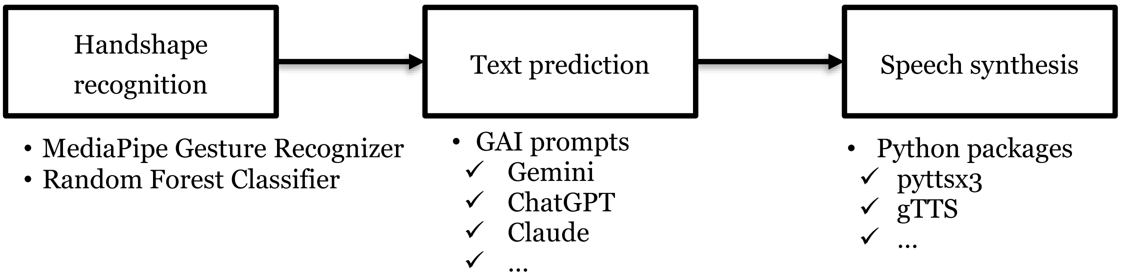


Figure 4. The main stages of the proposed method.

3.1 Handshape recognition

Handshape recognition, as depicted in Figure 5, begins with users forming ASL alphabet hand gestures. A webcam captures these gestures, and MediaPipe Gesture Recognizer extracts the hand keypoint coordinates defined in Figure 2. These coordinates serve as input for our pretrained Random Forest classifiers. Specifically, the process involves:

1. Gesture Capture: The webcam records the user’s ASL hand gesture.
2. Keypoint Extraction: MediaPipe Gesture Recognizer outputs the coordinates of hand keypoints as a dictionary object.
3. Normalization: A Min-Max normalization transforms these coordinates into the range [0, 1], as:

   $$x' = \frac{x-min}{max-min}, \tag{1}$$

   where x' is the normalized value, x is the original coordinate, and min and max are the minimum and maximum values of the original coordinates output by MediaPipe Gesture Recognizer, respectively.
4. Classification: The normalized keypoint coordinates are then fed into the Random Forest classifiers.
5. Alphabet Output: The classifiers output the corresponding ASL alphabet.

For example, when users form the gestures for 'i', 'l', and 'u', the respective keypoint coordinates are extracted, normalized, and used to accurately identify those letters.

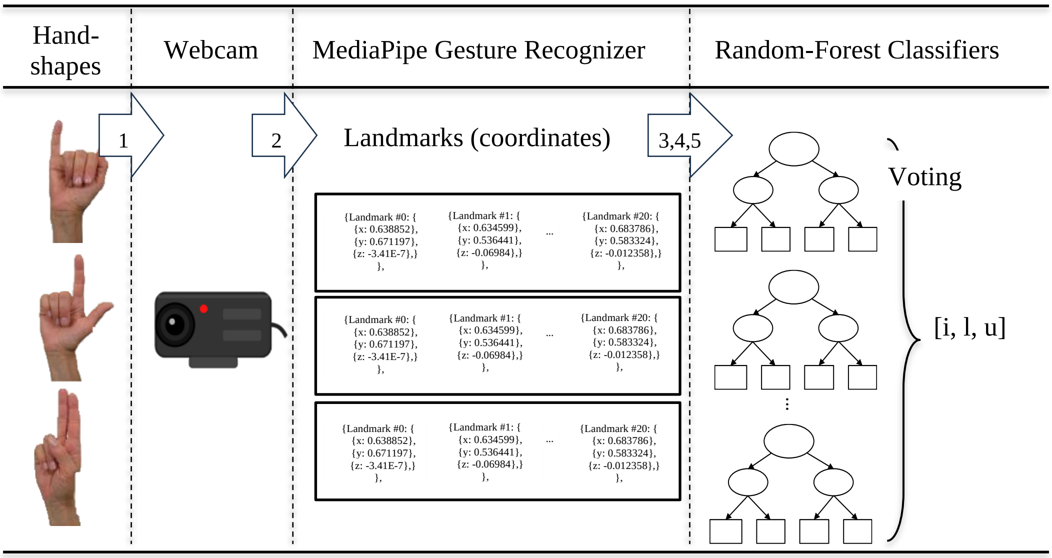

Hand-shapes
Webcam
MediaPipe Gesture Recognizer
Random-Forest Classifiers
1
2
Landmarks (coordinates)
3,4,5
{Landmark #0: { {x: 0.638852}, {y: 0.671197}, {z: -3.41E-7},} },
{Landmark #1: { {x: 0.634599}, {y: 0.536441}, {z: -0.06984},} },
...
{Landmark #20: { {x: 0.683786}, {y: 0.583324}, {z: -0.012358},} },
Voting
[i, l, u]

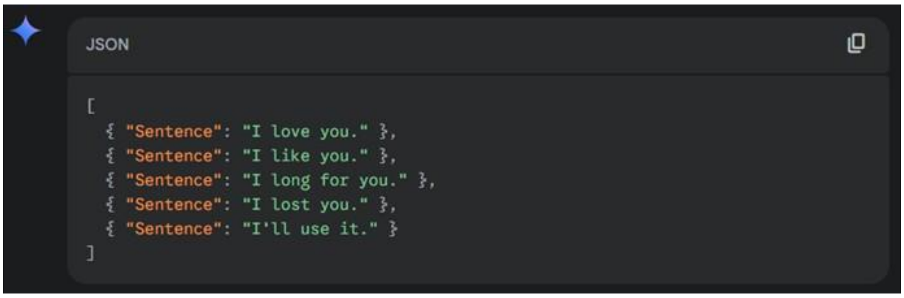
Figure 5. The processes in handshape recognition.

3.2 Text prediction

A key aspect of large language model (LLM) training involves masked language models (MLMs), which facilitate the development of LLM's understanding of human cognition. In this approach, specific words or tokens within an input sequence are randomly masked. The LLM is then trained to predict these masked elements by utilizing the contextual information provided by the surrounding words. Consequently, LLMs like ChatGPT, Gemini, and Claude can effectively infer user intent even when presented with incomplete or partial input. This study utilizes this capability, assuming robust MLM training, to explore LLMs' ability to complete text from partial vocabulary or abbreviations. For instance, prompting Gemini with 'Complete the top 5 possible sentences with this schema: {"Sentence": str} but without any explanation if the sentence "i l u" is incomplete' yields the response shown in Figure 6. Additionally, language translation is available as the response shown in Figure 7.

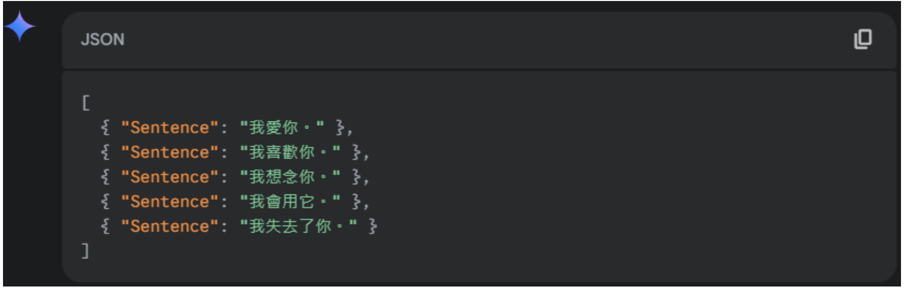
3. 3 Speech synthesis

Depending on user needs, text-to-speech (TTS) conversion can be incorporated. Given the established maturity of TTS technology, numerous Python packages are available, such as pyttsx3 and gTTS. pyttsx3 was selected for this study due to its robust offline capabilities, cross-platform support, and straightforward usage.



```json
[
  { "Sentence": "I love you." },
  { "Sentence": "I like you." },
  { "Sentence": "I long for you." },
  { "Sentence": "I lost you." },
  { "Sentence": "I'll use it." }
]
```

Figure 6. Gemini's output for 'i l u' sentence completion.



```json
[
  { "Sentence": "我愛你。" },
  { "Sentence": "我喜歡你。" },
  { "Sentence": "我想念你。" },
  { "Sentence": "我會用它。" },
  { "Sentence": "我失去了你。" }
]
```

Figure 7. Gemini's output for 'i l u' sentence completion in traditional Chinese.

4. Empirical Evaluation

In this section, we will first implement the handshape recognition and then the text prediction.

4.1 Handshape recognition

Handshape recognition involves two stages. First, MediaPipe, a Python package, extracts hand landmarks (coordinates) from webcam input. Second, a pre-trained classifier identifies letters based on these coordinates. This study utilizes a Random Forest classifier(11), trained on data from Kaggle(15). Random Forests were chosen for their ability to integrate multiple weak classifiers trained on diverse datasets, mitigating issues like overfitting and bias.

Our experimental evaluation involved individuals with CP and a control group without CP. To assess our classifier's performance, we enlisted one author with CP and ten student volunteers. Participants were instructed to hold each hand gesture for one second after ceasing movement, a condition under which all gestures could theoretically be correctly identified. The results, summarized in Table 1, indicate a recognition accuracy of 54.615% for individuals with CP and 94.808% for the control group. Lower accuracy for certain static gestures like 'A', 'M', 'N', 'S', and 'T' may be attributed to their visual similarity. Furthermore, dynamic gestures such as 'J' and 'Z' presented additional recognition challenges due to their inherent motion.

This initial outcome, however, did not meet our objective of enhancing communication effectiveness for individuals with CP. To address this, we collected additional training data by recording more videos from our author with CP, capturing 100 hand coordinate sets per letter using MediaPipe. After retraining our classifier with this augmented dataset, the results, presented in Table 2, showed an improved recognition accuracy of 75.769% for individuals with CP and a slightly decreased accuracy of 91.115% for the control group. Comparing Tables 1 and 2 highlights the necessity of customizing a recognition model specifically for individuals with CP, especially given the observed trade-off where improved accuracy for the CP group corresponded with a reduction for the control group.

Letter	Successes (CP)	Successes (No CP)	Letter	Successes (CP)	Successes (No CP)
A	9	91	N	9	84
B	3	95	O	9	98
C	1	98	P	7	98
D	5	99	Q	8	99
E	8	98	R	5	98
F	7	99	S	8	96
G	6	96	T	5	90
H	2	96	U	6	99
I	3	96	V	7	96
J	1	88	W	6	97
K	5	90	X	4	98
L	8	96	Y	2	99
M	6	85	Z	2	86
Total	64	1227	Total	78	1238

Table 1. The experimental results of our Random-Forest classifier.

Letter	Successes (CP)	Successes (No CP)	Letter	Successes (CP)	Successes (No CP)
A	10	82	N	9	86
B	6	86	O	9	97
C	6	90	P	7	98
D	8	89	Q	8	97
E	8	98	R	8	94
F	9	99	S	8	92
G	10	90	T	8	90
H	8	88	U	8	92
I	8	83	V	7	95
J	7	88	W	6	98
K	8	88	X	4	98
L	10	96	Y	5	89
M	6	83	Z	6	83
Total	104	1160	Total	93	1209

Table 2. The experimental results of our re-trained classifier.

4.2 Text prediction

To evaluate the text prediction capability of GAI models, we propose an approach called Partial Word Masking (PWM). Table 3 illustrates the PWM rate calculation with four examples. Consider the first example: for the sentence "I agree," we tested various letter combinations. We observed that "I ag" was the shortest input that still resulted in "I agree" being among the top five predictions by Gemini 2.0 Flash. The PWM rate for each word is calculated as the proportion of masked letters. For "I," the rate is (1-1)/1 = 0, as no letters were masked. For "agree," with five letters and two given ("ag"), the masked length is (5-2) = 3, resulting in a PWM rate of 3/5 = 0.6. The PWM rate for the entire sentence "I agree" is then defined as the maximum of the individual word PWM rates: Max(0, 0.6) = 0.6.

Original sentences	Successes conditions	Word masked rate	Sentence masked rates
I agree	I ag	I: (1-1)/1=0 agree: (5-2)/5=0.6	Max(0, 0.6) = 0.6
Not yet	Not yet	Not: (1-1)/1=0 yet: (5-2)/5=0	Max(0, 0)=0
See you	See y	See: (1-1)/1=0 you: (5-2)/5=0.67	Max(0, 0.67)=0.67
I didn't mean it	I din m i	I: (1-1)/1=0 didn't: (6-3) / 6=0.5 mean: (4-1)/4 = 0.75 it: (2-1)/2=0.5	Max(0, 0.5, 0.75, 0.5)=0.75

Table 3. Examples for computing partial word masking rates.

We evaluated Gemini 2.0 Flash using the first one hundred sentences from (http://www.eng.fju.edu.tw/etc/quiz/lifeeng.htm). The results, presented in Table 4, show the frequency of "Hit events," where Gemini 2.0 Flash correctly predicted the original sentence when presented with partially masked words. Notably, approximately half of the sentences were accurately inferred when the PWM rate fell within the [60%, 80%] range. Conversely, only about 30% of sentences were correctly predicted at PWM rates below 50%. These findings suggest that GAI models like Gemini 2.0 Flash have the potential to significantly reduce user input demands in the proposed AAC system, particularly when a moderate level of word information is provided.

PWM rate ranges	Hit events
[0%, 20%)	5
[20%, 40%)	5
[40%, 60%)	21
[60%, 80%)	49
[80%, 100%]	20
Total	100

Table 4. The PWM rates and their accuracy events.

5. Conclusion

The novel AAC system presented in this research offers a promising avenue for enhancing communication accessibility for individuals with severe disabilities, particularly those with cerebral palsy. By integrating hand gesture recognition powered by MediaPipe and a Random Forest classifier with the predictive capabilities of GAI models, this system aims to significantly reduce the physical and cognitive demands typically associated with AAC use. Our initial experimental evaluation of hand gesture recognition revealed a notable difference in accuracy between individuals with CP (54.615%) and a control group (94.808%), highlighting the challenges posed by motor impairments. Subsequent efforts to improve recognition for individuals with CP through additional training data yielded a substantial increase in accuracy to

75.769%, albeit with a slight decrease for the control group (91.115%), underscoring the need for tailored models. Furthermore, the partial word masking experiment with Gemini 2.0 Flash demonstrated the potential of GAI models to infer intended sentences even with significant portions of words masked, with approximately 50% accuracy within a 60-80% masking range. This suggests that by providing even fragmented sign input, the proposed AAC system can utilize the predictive power of advanced language models to complete intended messages, thereby lowering the user burden. The combined results of these experiments validate the viability of our integrated approach and its potential to significantly improve communication effectiveness and independence for individuals who rely on AAC. Future work will focus on refining the hand gesture recognition model for improved robustness and exploring user-centered design principles to optimize the overall usability and real-world impact of this novel AAC system.

References

1 T. Griffiths, R. Slaughter, and A. Waller: Journal of Enabling Technologies **18** (2024) 232. doi: http://dx.doi.org/10.1108/JET-01-2024-0007

2 E. Efeoğlu and A. Tuna: Kirklareli University Journal of Engineering and Science **10** (2024) 219. doi: http://dx.doi.org/10.34186/klujes.1546178

3 R. Scherer, M. Billinger, J. Wagner, A. Schwarz, D. T. Hettich, E. Bolinger, M. Lloria Garcia, J. Navarro, and G. Müller-Putz: Annals of Physical and Rehabilitation Medicine **58** (2015) 14. doi: http://dx.doi.org/https://doi.org/10.1016/j.rehab.2014.11.005

4 D. Kouremenos and K. Ntalianis, Glam-Sign: Greek Language Multimodal Lip Reading with Integrated Sign Language Accessibility 2025) doi: http://dx.doi.org/10.48550/arXiv.2501.05213

5 L. Pope and J. Light: Augmentative and Alternative Communication (2025) 1. doi: http://dx.doi.org/10.1080/07434618.2025.2458868

6 Y. Guerrier, S. Bosquet, O. Valadier, N. Cauchois, V. Delcroix, K. M. de Oliveira, and C. Kolski: HCI International 2024 – Late Breaking Papers (2025) 14. doi: http://dx.doi.org/

7 D. Ward, A. Blackwell, and D. Mackay, Dasher---a Data Entry Interface Using Continuous Gestures and Language Models 2000) p 129 doi: http://dx.doi.org/10.1145/354401.354427

8 C. N. Rang, P. Jerónimo, C. Mora, and S. Jardim: Procedia Computer Science **256** (2025) 198. doi: http://dx.doi.org/https://doi.org/10.1016/j.procs.2025.02.112

9 MediaPipe Gesture Recognizer: Mediapipe Gesture Recognizer, https://ai.google.dev/edge/mediapipe/solutions/vision/gesture_recognizer?hl=zh-tw (2025/4/15,

10 D. J. Bailey, F. Herget, D. Hansen, F. Burton, G. Pitt, T. Harmon, and D. Wingate: Aphasiology (2025) 1. doi: http://dx.doi.org/10.1080/02687038.2024.2445663

11 L. Breiman: Mach. Learn. **45** (2001) 5. doi: http://dx.doi.org/

12 Lingvano: The Alphabet in Asl Sign Language, https://www.lingvano.com/asl/blog/sign-language-alphabet/ (2025/4/15,

13 C.-C. Chen, Y.-S. Lin, and H.-Y. Chen: Sensors and Materials **36** (2024) 4835. doi: http://dx.doi.org/10.18494/SAM5224

14 L. Banh and G. Strobel: Electronic Markets **33** (2023) 63. doi: http://dx.doi.org/10.1007/s12525-023-00680-1

15 Kaggle: American Sign Language Fingerspelling Recognition, https://www.kaggle.com/competitions/asl-fingerspelling/data (2025/4/17,

About the Authors

First Author received her B.S. degree from ABC University, Japan, in 2000 and her M.S. and Ph.D. degrees from the XY Institute of Technology, Japan, in 2002 and 2005, respectively. From 2005 to 2009, she was an assistant professor at ABC University, Japan. Since 2010, she has been a professor at DEF University. Her research interests are in MEMS, bioengineering, and sensors. (xxxxxxxx@xxx.edu.jp)

Chen-Wei Lin received his B.S. degree from National Chin-Yi University of Technology in Taiwan in 2024. He is currently a graduate student in the Department of Information Management at the same university. His research interest focuses on applying artificial intelligence to Augmentative and Alternative Communication. (eexx3939@gmail.com)

Chien-Chih Chen is an assistant professor in the Department of Information Management of National Chin-Yi University of Technology, Taiwan. His current interests are focused on machine learning with small data sets. His articles have appeared in *Decision Support Systems*, *Omega*, *Automation in Construction*, *Computers and Industrial Engineering*, *International Journal of Production Research*, *Neurocomputing*, and other publications. (frick@ms14.hinet.net)