Gesture-Aware Zero-Shot Speech Recognition for Patients with Language Disorders

Seungbae Kim[1*], Daeun Lee[2*], Brielle Stark[3], Jinyoung Han[4†]

[1] Department of Computer Science and Engineering, University of South Florida
[2] Yale School of Medicine, Yale University
[3] Department of Speech, Language and Hearing Sciences, Indiana University Bloomington
[4] Department of Applied Artificial Intelligence, Sungkyunkwan University
seungbae@usf.edu, da-eun.lee@yale.edu, bcstark@iu.edu, jinyounghan@skku.edu

Abstract

Individuals with language disorders often face significant communication challenges due to their limited language processing and comprehension abilities, which also affect their interactions with voice-assisted systems that mostly rely on Automatic Speech Recognition (ASR). Despite advancements in ASR that address disfluencies, there has been little attention on integrating non-verbal communication methods, such as gestures, which individuals with language disorders substantially rely on to supplement their communication. Recognizing the need to interpret the latent meanings of visual information not captured by speech alone, we propose a gesture-aware ASR system utilizing a multimodal large language model with zero-shot learning for individuals with speech impairments. Our experiment results and analyses show that including gesture information significantly enhances semantic understanding. This study can help develop effective communication technologies, specifically designed to meet the unique needs of individuals with language impairments.

Introduction

Language disorders, such as aphasia, arise from damage to brain regions responsible for language production and comprehension. Aphasia is most often caused by acquired brain injuries, like stroke, and persists chronically in at least 30% of cases (Broca et al. 1861; Wasay, Khatri, and Kaul 2014). This means that many are living with aphasia – indeed, there are nearly two million with aphasia in the USA alone (Simmons-Mackie and Cherney 2018). Individuals with language disorders face significant communication challenges due to difficulties in processing and understanding language, resulting in impaired social interactions and reduced quality of life (El Hachioui et al. 2017). As voice-assisted technologies like Siri and Alexa become integral to daily activities, the inability of individuals with language disorders to interact effectively with these systems exacerbates their communication barriers, leading to frustration and further marginalization (Rohlfing et al. 2021).

Automatic Speech Recognition (ASR) systems are designed to transcribe spoken language into text, serving as a cornerstone for many voice-driven technologies. While recent advancements have made ASR systems more adept at handling disfluencies such as stuttering, their performance heavily depends on clear and consistent audio input (Radford et al. 2023). For individuals with language disorders, speech distortions, substitutions, and disjointed delivery pose significant challenges to accurate ASR transcription (Sanguedolce, Naylor, and Geranmayeh 2023). As a result, current ASR solutions often fall short in addressing the needs of this population.

To enhance the robustness of ASR, researchers have explored Audio-Visual Speech Recognition (AVSR) systems that combine auditory and visual information (Cheng et al. 2023b). These systems leverage visual features such as lip movement (Hu et al. 2023; Cheng et al. 2023a) and facial expressions (Zadeh et al. 2016; Busso et al. 2008) to improve speech recognition accuracy. However, these approaches are not ideal for individuals with language disorders, who often experience concomitant motor speech disorders and facial hemiplegia, which can result in 'masked' facial expressions (Multani et al. 2017; Duffy et al. 2012).

In contrast, individuals with language disorders frequently rely on non-verbal communication, such as gestures, to compensate for their verbal limitations (Stark and Oeding 2023). Iconic gestures, in particular, serve as powerful tools for conveying meaning when spoken language is insufficient (de Kleine et al. 2023; van Nispen, van de Sandt-Koenderman, and Krahmer 2017; Stark and Cofoid 2022; Stark and Oeding 2023). Unlike speech or facial expressions, iconic gestures offer a visual representation of concepts, enabling individuals with language impairments to express ideas more effectively (Lee et al. 2023). However, current ASR and AVSR systems fail to consider the latent semantic information encoded in these gestures, leaving a critical gap in understanding the full context of communication for individuals with language disorders.

To address the limitations of current ASR and AVSR systems, we introduce a gesture-aware zero-shot speech recognition framework specifically designed for individuals with language disorders. Our approach harnesses the capabilities of multimodal large language models (LLMs) to incorpo-

[*]Equal contribution.
[†]Corresponding author.

Copyright © 2025, GenAI4Health Workshop @ Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

rate linguistic, acoustic, and gestural information, enabling a deeper understanding of the speaker's intended meaning. Unlike traditional systems that rely solely on spoken or visual cues like lip movements, our method emphasizes the integration of iconic hand gestures—gestures that visually represent concepts—to enrich the transcription process. This allows the system to produce transcripts that not only reflect spoken content but also capture the latent meanings conveyed through gestures, which are crucial for individuals with impaired speech.

The core of our method lies in leveraging a zero-shot framework to align and synthesize multimodal inputs. Our system processes disfluent or incomplete speech signals while simultaneously analyzing visual data to identify and interpret gestures. By fusing these streams of information, the system generates semantically enriched transcripts that bridge the gaps left by speech alone. This zero-shot design eliminates the need for task-specific training, making the approach adaptable to a wide range of scenarios and linguistic contexts. This adaptability is particularly advantageous for language-disordered populations, where speech patterns and gestures vary widely between individuals.

Our experiments show that our proposed model successfully generates transcripts by incorporating significant gestural information to ascertain the latent intent of individuals with language disorders, which is not conveyed through speech alone. This findings indicate that incorporating nonverbal information can effectively aid in advancing the creation of more inclusive and effective communication technologies designed specifically for the distinctive requirements of individuals with language impairments.

Related Work

ASR as Assistive Technologies

Assistive technologies play a critical role in supporting clinicians to deliver assessments and therapies, as well as in facilitating communication for individuals with aphasia. In speech-language pathology, such tools are often referred to as augmentative and alternative communication (AAC) systems (Beukelman and Mirenda 1998). ASR technology has shown promise in enhancing communication by enabling real-time, accurate feedback to individuals with aphasia (Ballard et al. 2019; Barbera et al. 2021). This is particularly valuable as individuals with aphasia frequently face challenges with self-monitoring (Oomen, Postma, and Kolk 2001; Sampson and Faroqi-Shah 2011) and often benefit from external cues to improve their performance (Tompkins, Scharp, and Marshall 2006; Conroy, Sage, and Lambon Ralph 2009; Schwartz et al. 2016). Additionally, ASR can improve communication efficiency and quality by compensating for writing or grammatical impairments through speech-to-text conversion and delivering feedback during therapy (Ballard et al. 2019; Barbera et al. 2021). Consequently, ASR-based applications have gained popularity in delivering speech-language services for individuals with aphasia and other neurogenic conditions, such as Parkinson's disease (Hoover and Carney 2014; McCrocklin 2016; Strik et al. 2009).

However, despite advancements in ASR, reliance on voice-based systems alone often results in transcription errors when processing disfluent or incomplete speech, a hallmark of language impairments (Jefferson 2019; Le et al. 2016). These inaccuracies can hinder communication and reduce the overall effectiveness of these tools, further exacerbating barriers faced by individuals with language disorders. To address these challenges, AVSR systems have emerged, combining visual inputs—such as lip movements and facial expressions—with audio to enhance recognition accuracy (Gabeur et al. 2022; Afouras et al. 2018; Dupont and Luettin 2000; Ma, Petridis, and Pantic 2021; Noda et al. 2015; Mroueh, Marcheret, and Goel 2015; Feng et al. 2017). However, AVSR systems often fall short for individuals with language and speech disorders. Speech disorders such as dysarthria and apraxia frequently disrupt articulation, rendering lip movements unreliable. Moreover, neutral or ambiguous facial expressions common in these populations provide limited contextual insight (Tong, Sharifzadeh, and McLoughlin 2020; Salama, El-Khoribi, and Shoman 2014).

Gesture into Assistive Technology

Gestures play a vital role in communication, often conveying critical information that complements, enhances, or even substitutes spoken language (Kita 2009; Kendon 1994; McNeill, Pedelty, and Levy 1990). This holds true for individuals with and without language impairments, as gestures are co-produced with speech across languages and cultures, aiding both speakers and listeners (Kita 2009). Remarkably, even congenitally blind speakers, who have never observed gestures, use gestures as frequently as sighted individuals, emphasizing their fundamental role in communication (Iverson and Goldin-Meadow 1998). For speakers, gestures illustrate abstract concepts, emphasize key points, and provide additional information that speech alone may struggle to convey (McNeill, Pedelty, and Levy 1990; Kita 2000; Goldin-Meadow and Alibali 2013). For listeners, gestures offer visual cues that complement verbal messages, enhancing understanding in noisy settings, during non-native language processing, and when speech is ambiguous (Cook and Tanenhaus 2009; Goldin-Meadow 1999). Notably, listeners often recall information received via gestures as if it were spoken, underscoring their semantic importance (Cassell 2000; Kelly et al. 1999).

Among gestures, representational gestures—those that visually represent objects or actions—are particularly valuable for enhancing communication (McNeill, Pedelty, and Levy 1990). Iconic gestures, a subtype of representational gestures, depict concrete referents (e.g., a kicking motion to signify "kick") (Novack and Goldin-Meadow 2017). Other representational gestures include pantomimes, which occur without speech, and deictic gestures, such as pointing, which may emphasize or add meaning to speech(Kendon 1994). Additionally, emblems like a "thumbs up" convey culturally specific meanings independent of speech (Kendon 1994). However, not all gestures are communicative. Non-communicative movements, such as tucking hair behind the ear, do not contribute to the conveyed message and can be irrelevant to the communicative process.

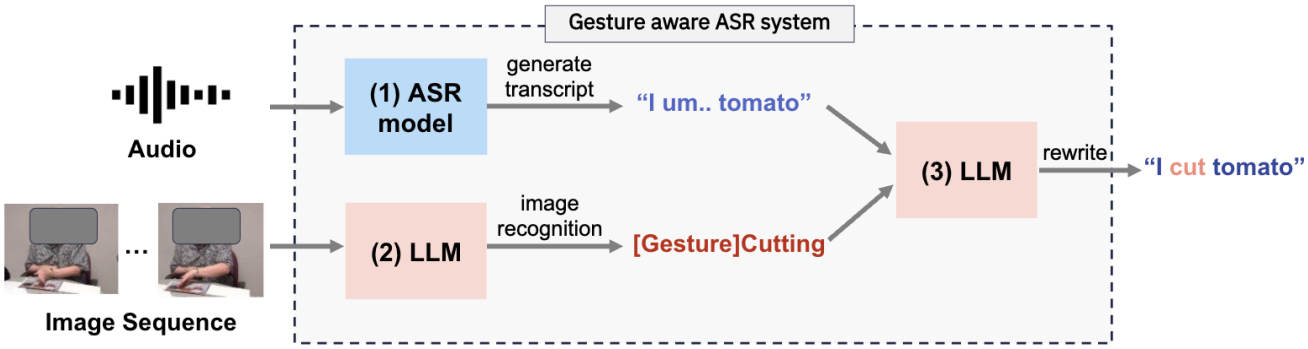

Figure 1: The overall process of the proposed system. Our model integrates incomplete speech and visual data (i.e., iconic gestures) and generates semantically enriched transcripts.

For assistive technologies, particularly AI-based systems, distinguishing communicative gestures from non-communicative movements is essential. Systems must accurately interpret gestures to capture the full intent of a speaker's message, especially when assisting individuals with language impairments. Thus, effectively integrating linguistic and co-speech gesture information is critical for improving the utility and effectiveness of AI-driven assistive tools.

Challenges in Gesture Interpretation for Assistive Technology

Accurately interpreting the wide variety of gestures individuals use is a significant challenge, as gestures vary in form and meaning, with some being non-communicative. Even gestures conveying the same intent, such as cutting, can differ—for example, an "enacting" gesture mimics the action of cutting with a flat hand, while a "handling" gesture simulates gripping an invisible knife (Poggi 2008; Hassemer and Winter 2018). This variability complicates consistent interpretation, particularly in zero-shot settings. While recent computer vision methods focus on generating gestures from text inputs for avatars or video sequences (Ginosar et al. 2019; Ahuja, Lee, and Morency 2022; Liu et al. 2022), they do not address gesture understanding or intent recognition, relying instead on predefined datasets. Existing ASR and AVSR systems also fall short, as they lack the ability to incorporate conversational context or personalized knowledge, both critical for interpreting gestures tied to specific topics or individual habits. For instance, one person may favor "handling" gestures while another prefers "enacting." Additionally, task-specific training required by many systems is impractical for real-time applications, as it necessitates frequent retraining to accommodate new users and scenarios. These limitations underscore the need for adaptable systems capable of real-time, personalized gesture interpretation.

Method

Given that the dataset provides video recordings of persons with language disorders, the proposed model aims to improve the transcription of audio from individuals with language disorders, focusing on enhancing contextual understanding, incorporating gesture, and resolving ambiguities in speech. It consists of three interconnected components: Speech Recognition, Gesture Recognition, and Contextual Rewriting. Figure 1 provides a schematic overview of the system.

Speech Recognition The first step in the pipeline is audio processing using an ASR system. This component takes an audio signal $\mathbf{A}$ as input and outputs a preliminary transcript T_{ASR}, which represents the system's initial interpretation of spoken words. Let the ASR model be represented by f_{ASR} as follows,

$$T_{\text{ASR}} = f_{\text{ASR}}(\mathbf{A}), \quad T_{\text{ASR}} = \{w_1, w_2, \ldots, w_n\} \tag{1}$$

where w_i is the i-th word in the transcript. Consider an audio input from a patient saying, "I um... tomato." The ASR model may generate a raw transcript as incomplete speech. This transcript lacks clarity and completeness, as it fails to convey the full meaning intended by the speaker. The limitations arise because ASR systems rely solely on the audio signal and are unable to incorporate accompanying non-verbal cues, such as gestures, that could provide additional context.

Gesture Recognition The second component focuses on identifying gestures from the video frames corresponding to the speech. The input is a sequence of video frames $\mathbf{V} = \{v_1, v_2, \ldots, v_m\}$, where each frame v_i captures a snapshot of the speaker's hand or body movements. A gesture recognition model, f_{Gesture}, processes these frames to detect meaningful gestures as follows,

$$\mathbf{G} = f_{\text{Gesture}}(\mathbf{V}), \quad \mathbf{G} = \{g_1, g_2, \ldots, g_k\} \tag{2}$$

where g_i represents a detected gesture and its associated meaning. Due to the scarcity of labeled datasets for iconic gestures, the framework employs a multimodal large language model to perform zero-shot gesture recognition. This approach allows the model to infer the meanings of gestures based on generalizable patterns learned from other modalities. Iconic gestures, which represent concrete referents, are

of particular interest because they often clarify or complement the spoken language.

If the video captures the speaker making a back-and-forth motion with a flat hand, the gesture recognition model identifies this as a cutting motion and assigns it the label g="cut". Iconic gestures, however, can vary significantly between individuals. For instance, one person might perform a "handling" gesture by mimicking the act of holding and moving an invisible knife, while another might use an "enacting" gesture, simulating the motion of cutting without mimicking the act of holding. The proposed model is designed to handle such variability by recognizing these different representations and mapping them to a unified semantic meaning, ensuring that the gesture's intent is accurately interpreted regardless of individual differences in gesture style.

Contextual Rewriting The third component integrates the outputs from the ASR and gesture recognition models to generate a contextually enriched and semantically accurate transcript. This process is handled by a LLM which takes the initial transcript T_{ASR}, the detected gestures $\mathbf{G}$ as inputs. The final transcript, T_{Final}, is generated as follows:

$$T_{\text{Final}} = f_{\text{LLM}}(T_{\text{ASR}}, \mathbf{G}), \quad T_{\text{Final}} = \{w'_1, w'_2, \ldots, w'_p\} \quad (3)$$

where w'_i represents a word in the final transcript. The model directly uses the recognized gestures as the sole non-verbal input, streamlining the process while focusing on the integration of multimodal signals.

For example, if the ASR transcript reads, "I um... tomato," and the gesture recognition model identifies the label g="cut" from a cutting motion, the LLM rewrites the transcript as "I cut tomato." This correction ensures that the final output reflects both the spoken and gestural communication, improving the system's ability to understand and represent the speaker's intent accurately. This approach enables the generation of accurate and enriched transcripts, even in cases where speech alone might be ambiguous or incomplete. By leveraging the co-speech gesture input, the system enhances its interpretive power without requiring additional external context.

Experiments

We collected the dataset from AphasiaBank (MacWhinney et al. 2011; Forbes, Fromm, and MacWhinney 2012), a shared database created by clinical experts for aphasia research; the corpus information is summarized in Table 1. The dataset includes video recordings of the language evaluation test process between a pathologist and a subject, which also contains human-annotated transcriptions and subjects' demographic information.

For our study, we selected the "Peanut Butter Sandwich Task" as the focal activity where people are asked to explain the procedure for making a sandwich. This task has been demonstrated to associate with high rates of iconic gesturing (Stark and Cofoid 2022; Stark and Oeding 2023; Pritchard et al. 2015; Illes 1989).

We performed a detailed analysis of the dataset to examine the distribution and characteristics of gestures used by participants during this task. Table 2 provides a summary of the data, showing six distinct gesture types: cutting, spreading, folding, eating, layering, and opening. Each gesture corresponds to an action performed to describe or demonstrate a step in the sandwich-making process. For example, the cutting gesture involves a back-and-forth hand motion mimicking the act of slicing, while the spreading gesture typically represents the act of spreading condiments with a sweeping motion of the hand. The data highlights the natural integration of gestures with speech, making it a valuable resource for gesture-aware ASR systems.

Among the gestures, spreading was the most frequently used, appearing in 269 utterances across 169 users, with an average duration of 4.53 seconds. This gesture reflects a critical step in the sandwich-making process, as participants commonly describe spreading peanut butter or other ingredients. Following this, folding emerged as the second most frequently observed gesture, with 159 instances across 145 users, averaging 2.95 seconds in duration. The cutting gesture ranked third, appearing 109 times across 92 users, with an average duration of 3.76 seconds. These three gestures collectively account for a significant portion of the dataset,

Table 1: The corpora from AphasiaBank

Corpus	Site
ACWT (Bynek 2013)	Aphasia Center of West Texas
Adler (Szabo 2013)	Adler Aphasia Center
APROCSA (Wilson 2021)	Vanderbilt University Medical Center
BU (Hoover 2013)	Boston University
Capilouto (Capilouto 2008)	University of Kentucky
CC-Stark (Stark 2022)	file for CC
CMU (MacWhinney 2013)	Carnegie Mellon University
Elman (Elman 2011, 2016)	Aphasia Center of California
Fridriksson (Fridriksson 2013)	University of South Carolina
Garrett (Garrett 2013)	Pittsburgh, PA
Kansas (Jackson 2013)	University of Kansas
Kempler (Kempler 2013)	Emerson College
Kurland (Kurland 2013)	University of Massachusetts, Amherst
MSU (Boyle 2013)	Montclair State University
NEURAL (Stark 2023)	NEURAL Research Lab, Indiana University
Richardson (Richardson 2008)	University of New Mexico
SCALE (McCall 2013)	Snyder Center for Aphasia Life Enhancement
STAR (Corwin 2013)	Stroke Aphasia Recovery Program
TAP (Silverman 2013)	Triangle Aphasia Project
TCU (Muñoz 2013)	Texas Christian University
Thompson (Thompson 2013)	Northwestern University
Tucson (Hirsch Kruse 2013)	University of Arizona
UCL (Dean 2021)	University College London
UMD (Faroqi-Shah and Milman 2018)	University of Maryland
UNH (Ramage 2013)	University of New Hampshire
Whiteside (Whiteside 2013)	University of Central Florida
Williamson (Williamson 2013)	Stroke Comeback Center
Wozniak (Wozniak 2013)	InteRACT
Wright (Wright 2013)	Arizona State University

Table 2: Summary of data statistics of 'Peanut Butter Task' based on gesture annotation.

label	# utt	# user	duration (sec)		
			mean	min	max
[gesture:cutting]	109	92	3.76	0.21	18.17
[gesture:eating]	54	48	2.01	0.12	20.79
[gesture:folding]	159	145	2.95	0.09	21.31
[gesture:layering]	45	31	3.92	0.39	16.99
[gesture::opening]	34	24	4.56	0.31	16.29
[gesture:spreading]	269	169	4.53	0.14	26.42
total	670	288	3.99	0.26	19.77

Table 3: The Average Word Error Rate (WER) between original transcript and ASR results.

Model	WER
Whisper (Radford et al. 2023)	0.557
Seamless (Barrault et al. 2023)	0.891
Wave2Vec (Baevski et al. 2020)	0.630
Whisper (conf $>$ 0.2)	0.519

indicating their central role in describing sandwich preparation. These findings align with linguistic analysis of the same story, demonstrating cutting, spreading, and folding as core parts of the story (Dalton et al. 2020).

The dataset also revealed interesting variability in gesture execution. For instance, while spreading gestures were consistently observed, their duration ranged from brief motions lasting just 0.14 seconds to extended actions of up to 26.42 seconds. Similarly, cutting gestures varied in duration, from 0.21 seconds to 18.17 seconds, highlighting differences in participants' expressive styles and cognitive processing abilities. Gestures such as layering and opening were less frequently observed, occurring 45 and 34 times respectively, but exhibited average durations comparable to the more common gestures, suggesting that even less frequent gestures are articulated with similar complexity.

This analysis highlights the diversity in both the frequency and execution of gestures among participants, emphasizing the importance of robust recognition systems capable of capturing these variations. The prevalence of gestures like spreading, folding, and cutting also demonstrates their importance as key features for understanding communication in tasks involving procedural explanations. These findings provide a strong foundation for training and evaluating our gesture-aware ASR framework, which seeks to seamlessly integrate gesture information into speech transcription.

Results

We conducted experiments to evaluate the performance of ASR models in generating initial transcripts. These experiments are essential for assessing the quality of the ASR output, which serves as the input for our gesture-aware contextual rewriting model. Additionally, we performed a case study on selected samples to highlight the improvements made by our proposed approach compared to the initial ASR outputs.

Speech Recognition Models

The foundation of a successful gesture-aware ASR system lies in obtaining high-quality initial transcripts from the speech recognition stage. To this end, we compared state-of-the-art ASR models on the AphasiaBank dataset to determine their performance in generating accurate transcripts. The results, summarized in Table 3, highlight the average Word Error Rate (WER) for each model, calculated by comparing their output with the ground truth human-annotated transcripts. WER serves as a critical metric for evaluating ASR systems, representing the proportion of errors (insertions, deletions, and substitutions) in the generated transcript relative to the total number of words in the ground truth.

Among the models tested, Whisper (Radford et al. 2023) demonstrated the best performance, achieving the lowest WER of 0.557. This indicates that Whisper is particularly effective in handling speech from participants with aphasia, despite challenges such as disfluencies, atypical speech patterns, and background noise. Wave2Vec (Baevski et al. 2020) performed moderately well with a WER of 0.630, while Seamless (Barrault et al. 2023), designed for multilingual ASR, showed a higher WER of 0.891, indicating potential limitations in its robustness to the specific speech characteristics in this dataset.

To further improve the accuracy of the initial transcripts, we applied a confidence-based filtering approach to Whisper's outputs. By utilizing Whisper's token-level confidence scores, we excluded tokens with a confidence score below 0.2, thereby retaining only the most reliable portions of the transcript. This refinement reduced Whisper's WER from 0.557 to 0.519, further solidifying its position as the most accurate ASR model in our evaluations.

These results highlight the importance of selecting a robust ASR model for datasets featuring non-standard speech patterns. Whisper, particularly with confidence-based filtering, provides a strong baseline for generating initial transcripts. The reduced WER achieved by Whisper demonstrates its ability to capture key elements of speech even under challenging conditions, thereby enhancing the reliability of the downstream contextual rewriting process.

Table 4: The result of a case study for comparison between Whisper and the proposed model. *[gesture:]* indicates the type of iconic gestures used by speakers while speaking.

Index	Origianl	Whisper	Ours
1	w [gesture:layering] is right.	There's right there.	You put things right there, doing a **layering** action.
2	[gesture:folding] uz@u uh right yes.	right.	**folding** it right.
3	um [gesture:cutting] [gesture:cutting] banana.	Um... Banana.	**cutting** banana.
4	and [gesture:eating].	and	and **eating**
5	but a w knife [gesture:spreading] do the stuff like um	night do this stuff like oh	do this night stuff like oh **spreading**

A Case Study

To evaluate the effectiveness of the proposed multimodal model in capturing and incorporating the latent meaning conveyed by iconic gestures, we conducted a case study comparing its outputs with those of the Whisper ASR model. Table 4 presents the results, showcasing the original transcripts in the dataset, the transcript generated by Whisper, and the transcript generated by our multimodal approach. The goal of this comparison was to determine whether our model can enhance transcript quality by inferring and integrating the semantic meaning of gestures observed during speech.

As shown in the table, Whisper provides a literal transcription of the spoken words but fails to capture the contex-

tual or semantic information conveyed by the speaker's gestures. For example, in the first example, the original input includes the layering gesture, indicating an action of stacking or placing items on top of each other. Whisper transcribes this as "*There's right there*," which misses the gesture's implicit meaning. In contrast, our proposed model generates the enriched transcript, "*You put things right there, doing a layering action*," effectively incorporating the gesture's semantic context into the spoken narrative.

In the second example, the speaker performs a folding motion, which is entirely absent from Whisper's transcription, resulting in the simple word "*right*." However, our model identifies and integrates this gesture into the transcript, producing "*folding it right*," which provides a more comprehensive description of the speaker's intention.

Similarly, in the third and fourth examples, iconic gestures like cutting and eating are completely omitted from Whisper's transcription, leading to outputs that lack essential contextual details. The proposed model enriches these outputs by explicitly including the actions, resulting in transcripts like "*cutting banana*" and "*and eating*," which more accurately reflect the speaker's message.

The fifth example highlights a limitation in handling more complex sentences where gestures provide significant contextual cues. While our model attempts to incorporate the spreading gesture into the generated transcript, the result, "*do this night stuff like oh spreading*," is not fully accurate. This inaccuracy stems from the initial transcription provided by Whisper, which included the word "*night*," leading our model to generate a sentence misaligned with the speaker's actual intent. This example underscores the importance of having accurate initial transcripts for improving the final output quality, especially in scenarios where subtle contextual nuances are critical.

It is important to note that the proposed model is capable of inferring and incorporating the speaker's underlying intention more effectively. By leveraging gestures alongside incomplete or disfluent speech, our approach captures the full semantic meaning of the speaker's communication. For example, when a speaker's verbal output is fragmented or ambiguous, the model uses the accompanying gestures to infer the intended message, producing transcripts that are both richer and more contextually accurate. Unlike Whisper, which relies solely on the spoken words and often misses crucial gestural context, our method bridges the gap between speech and gesture, offering a more comprehensive representation of the speaker's intent.

Concluding Remarks

In this paper, we proposed a novel approach utilizing multimodal LLMs to generate gesture-aware speech recognition transcripts for patients with language disorders. Our framework integrates verbal speech and iconic gestures, enabling the generation of enriched transcripts that capture the latent meaning conveyed through both modalities. Through extensive experimentation, we demonstrated that the proposed method effectively contextualizes incomplete or disfluent speech by incorporating gesture information, leading to more accurate and meaningful representations of the speaker's intent. These findings highlight the potential of our approach to significantly contribute to the field of speech and language therapy, offering innovative tools that can enhance the quality of life for individuals with language disorders by facilitating better communication and assessment methods.

Ethical Statement

Our dataset was obtained from AphasiaBank with the approval of the Institutional Review Board (IRB) and adheres to the data sharing guidelines set by TalkBank[1]. This includes complying with the Ground Rules for all TalkBank databases, which are based on the American Psychological Association Code of Ethics (Association et al. 2002).

Limitation & Future Work

While the results are promising, we recognize several limitations and outline our plans to extend this work further.

One primary limitation is the absence of a definitive ground truth for quantitative evaluation. Since our model generates transcripts by synthesizing speech and gesture data from scratch, traditional benchmarks, such as comparisons with standard speech recognition outputs, are insufficient. Moreover, existing original transcripts lack gesture annotations, making direct comparisons challenging. In future work, we aim to address this gap by collaborating with certified pathologists to conduct qualitative assessments, such as A-B preference tests, to evaluate the effectiveness of gesture-enriched transcripts in accurately conveying the speaker's intentions.

To support quantitative evaluations, we plan to develop novel metrics that assess transcript quality, including grammar accuracy, semantic consistency, and the integration of multimodal information. Such metrics will provide a more objective basis for assessing our model's performance and facilitate comparisons with other multimodal and unimodal approaches.

Another limitation of this study is its focus on structured gestures from a specific task, the Peanut Butter Sandwich Task. While this task offers a controlled context for testing our approach, it does not encompass the diversity of gestures and communication patterns seen in everyday scenarios. As part of our future work, we plan to expand the scope of our model to include tasks such as the Cinderella Story Recall Task (Bird and Franklin 1996), which involves unstructured and complex narrative gestures. This expansion will allow us to evaluate the adaptability and robustness of our model in handling varied linguistic and gestural contexts.

In summary, while this study establishes a strong foundation for gesture-aware speech recognition, we aim to refine and extend our methods through collaborative qualitative evaluations, the development of robust quantitative metrics, and broader task applications. These efforts will ensure that our approach continues to evolve, ultimately contributing to more effective communication tools and interventions for individuals with language disorders.

[1] https://talkbank.org/share/ethics.html

Acknowledgements

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ICAN (ICT Challenge and Advanced Network of HRD) support program (IITP-2024-RS-2023-00259497) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation) and under the Global Research Support Program in the Digital Field program (RS-2024-00425354) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

References

Afouras, T.; Chung, J. S.; Senior, A.; Vinyals, O.; and Zisserman, A. 2018. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(12): 8717–8727.

Ahuja, C.; Lee, D. W.; and Morency, L.-P. 2022. Low-resource adaptation for personalized co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20566–20576.

Association, A. P.; et al. 2002. Ethical principles of psychologists and code of conduct. *American psychologist*, 57(12): 1060–1073.

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460.

Ballard, K. J.; Etter, N. M.; Shen, S.; Monroe, P.; and Tien Tan, C. 2019. Feasibility of automatic speech recognition for providing feedback during tablet-based treatment for apraxia of speech plus aphasia. *American journal of speech-language pathology*, 28(2S): 818–834.

Barbera, D. S.; Huckvale, M.; Fleming, V.; Upton, E.; Coley-Fisher, H.; Doogan, C.; Shaw, I.; Latham, W.; Leff, A. P.; and Crinion, J. 2021. NUVA: a naming utterance verifier for aphasia treatment. *Computer Speech & Language*, 69: 101221.

Barrault, L.; Chung, Y.-A.; Meglioli, M. C.; Dale, D.; Dong, N.; Duquenne, P.-A.; Elsahar, H.; Gong, H.; Heffernan, K.; Hoffman, J.; et al. 2023. SeamlessM4T-Massively Multilingual & Multimodal Machine Translation. *arXiv preprint arXiv:2308.11596*.

Beukelman, D. R.; and Mirenda, P. 1998. *Augmentative and alternative communication*. Paul H. Brookes Baltimore.

Bird, H.; and Franklin, S. 1996. Cinderella revisited: A comparison of fluent and non-fluent aphasic speech. *Journal of neurolinguistics*, 9(3): 187–206.

Boyle, M. 2013. AphasiaBank English Protocol PWA MSU Corpus.

Broca, P.; et al. 1861. Remarks on the seat of the faculty of articulated language, following an observation of aphemia (loss of speech). *Bulletin de la Société Anatomique*, 6: 330–57.

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42: 335–359.

Bynek, K. 2013. AphasiaBank English Protocol PWA ACWT Corpus.

Capilouto, G. 2008. AphasiaBank English Controls Capilouto Corpus.

Cassell, J. 2000. Embodied conversational interface agents. *Communications of the ACM*, 43(4): 70–78.

Cheng, X.; Jin, T.; Huang, R.; Li, L.; Lin, W.; Wang, Z.; Wang, Y.; Liu, H.; Yin, A.; and Zhao, Z. 2023a. Mixspeech: Cross-modality self-learning with audio-visual stream mixup for visual speech translation and recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15735–15745.

Cheng, X.; Jin, T.; Li, L.; Lin, W.; Duan, X.; and Zhao, Z. 2023b. OpenSR: Open-Modality Speech Recognition via Maintaining Multi-Modality Alignment. *arXiv preprint arXiv:2306.06410*.

Conroy, P.; Sage, K.; and Lambon Ralph, M. A. 2009. The effects of decreasing and increasing cue therapy on improving naming speed and accuracy for verbs and nouns in aphasia. *Aphasiology*, 23(6): 707–730.

Cook, S. W.; and Tanenhaus, M. K. 2009. Embodied communication: Speakers' gestures affect listeners' actions. *Cognition*, 113(1): 98–104.

Corwin, M. 2013. AphasiaBank English Protocol PWA STAR Corpus.

Dalton, S. G. H.; Kim, H.; Richardson, J. D.; and Wright, H. H. 2020. A compendium of core lexicon checklists. In *Seminars in speech and language*, volume 41, 045–060. Thieme Medical Publishers.

de Kleine, N.; Rose, M. L.; Weinborn, M.; Knox, R.; and Fay, N. 2023. Does Gesture Improve the Communication Success of People with Aphasia?: A Systematic Review. *Aphasiology*, 1–25.

Dean, M. 2021. AphasiaBank English Protocol PWA UCL Corpus.

Duffy, J. R.; et al. 2012. *Motor speech disorders: Substrates, differential diagnosis, and management*. Elsevier Health Sciences.

Dupont, S.; and Luettin, J. 2000. Audio-visual speech modeling for continuous speech recognition. *IEEE transactions on multimedia*, 2(3): 141–151.

El Hachioui, H.; Visch-Brink, E. G.; de Lau, L. M.; van de Sandt-Koenderman, M. W.; Nouwens, F.; Koudstaal, P. J.; and Dippel, D. W. 2017. Screening tests for aphasia in patients with stroke: a systematic review. *Journal of neurology*, 264: 211–220.

Elman, R. J. 2011. Starting an aphasia center? In *Seminars in speech and language*, volume 32, 268–272. © Thieme Medical Publishers.

Elman, R. J. 2016. Aphasia centers and the life participation approach to aphasia. *Topics in Language Disorders*, 36(2): 154–167.

Faroqi-Shah, Y.; and Milman, L. 2018. Comparison of animal, action and phonemic fluency in aphasia. *International journal of language & communication disorders*, 53(2): 370–384.

Feng, W.; Guan, N.; Li, Y.; Zhang, X.; and Luo, Z. 2017. Audio visual speech recognition with multimodal recurrent neural networks. In *2017 International Joint Conference on neural networks (IJCNN)*, 681–688. IEEE.

Forbes, M. M.; Fromm, D.; and MacWhinney, B. 2012. AphasiaBank: A resource for clinicians. In *Seminars in speech and language*, volume 33, 217–222. Thieme Medical Publishers.

Fridriksson, J. 2013. AphasiaBank English Protocol PWA Fridriksson Corpus.

Gabeur, V.; Seo, P. H.; Nagrani, A.; Sun, C.; Alahari, K.; and Schmid, C. 2022. Avatar: Unconstrained audiovisual speech recognition. *arXiv preprint arXiv:2206.07684*.

Garrett, K. 2013. AphasiaBank English Protocol PWA Garrett Corpus.

Ginosar, S.; Bar, A.; Kohavi, G.; Chan, C.; Owens, A.; and Malik, J. 2019. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3497–3506.

Goldin-Meadow, S. 1999. The role of gesture in communication and thinking. *Trends in cognitive sciences*, 3(11): 419–429.

Goldin-Meadow, S.; and Alibali, M. W. 2013. Gesture's role in speaking, learning, and creating language. *Annual review of psychology*, 64(1): 257–283.

Hassemer, J.; and Winter, B. 2018. Decoding gestural iconicity. *Cognitive science*, 42(8): 3034–3049.

Hirsch Kruse, F. 2013. AphasiaBank English Protocol PWA Tucson Corpus.

Hoover, E. 2013. AphasiaBank English Protocol PWA BU Corpus.

Hoover, E. L.; and Carney, A. 2014. Integrating the iPad into an intensive, comprehensive aphasia program. In *Seminars in speech and language*, volume 35, 025–037. Thieme Medical Publishers.

Hu, Y.; Li, R.; Chen, C.; Qin, C.; Zhu, Q.-S.; and Chng, E. S. 2023. Hearing Lips in Noise: Universal Viseme-Phoneme Mapping and Transfer for Robust Audio-Visual Speech Recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15213–15232.

Illes, J. 1989. Neurolinguistic features of spontaneous language production dissociate three forms of neurodegenerative disease: Alzheimer's, Huntington's, and Parkinson's. *Brain and language*, 37(4): 628–642.

Iverson, J. M.; and Goldin-Meadow, S. 1998. Why people gesture when they speak. *Nature*, 396(6708): 228–228.

Jackson, S. 2013. AphasiaBank English Protocol PWA Kansas Corpus.

Jefferson, M. 2019. Usability of automatic speech recognition systems for individuals with speech disorders: Past, present, future, and a proposed model.

Kelly, S. D.; Barr, D. J.; Church, R. B.; and Lynch, K. 1999. Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of memory and Language*, 40(4): 577–592.

Kempler, D. 2013. AphasiaBank English Protocol PWA Kempler Corpus.

Kendon, A. 1994. Do gestures communicate? A review. *Research on language and social interaction*, 27(3): 175–200.

Kita, S. 2000. How representational gestures help speaking. *Language and gesture*, 1: 162–185.

Kita, S. 2009. Cross-cultural variation of speech-accompanying gesture: A review. *Language and cognitive processes*, 24(2): 145–167.

Kurland, J. 2013. AphasiaBank English Protocol PWA Kurland Corpus.

Le, D.; Licata, K.; Persad, C.; and Provost, E. M. 2016. Automatic assessment of speech intelligibility for individuals with aphasia. *IEEE/ACM transactions on audio, speech, and language processing*, 24(11): 2187–2199.

Lee, D.; Son, S.; Jeon, H.; Kim, S.; and Han, J. 2023. Learning Co-Speech Gesture for Multimodal Aphasia Type Detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 9287–9303.

Liu, H.; Zhu, Z.; Iwamoto, N.; Peng, Y.; Li, Z.; Zhou, Y.; Bozkurt, E.; and Zheng, B. 2022. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European conference on computer vision*, 612–630. Springer.

Ma, P.; Petridis, S.; and Pantic, M. 2021. End-to-end audio-visual speech recognition with conformers. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7613–7617. IEEE.

MacWhinney, B. 2013. AphasiaBank English Protocol PWA CMU Corpus.

MacWhinney, B.; Fromm, D.; Forbes, M.; and Holland, A. 2011. AphasiaBank: Methods for studying discourse. *Aphasiology*, 25(11): 1286–1307.

McCall, D. 2013. AphasiaBank English Protocol PWA SCALE Corpus.

McCrocklin, S. M. 2016. Pronunciation learner autonomy: The potential of automatic speech recognition. *System*, 57: 25–42.

McNeill, D.; Pedelty, L. L.; and Levy, E. T. 1990. Speech and gesture. In *Advances in psychology*, volume 70, 203–256. Elsevier.

Mroueh, Y.; Marcheret, E.; and Goel, V. 2015. Deep multimodal learning for audio-visual speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2130–2134. IEEE.

Multani, N.; Galantucci, S.; Wilson, S. M.; Shany-Ur, T.; Poorzand, P.; Growdon, M. E.; Jang, J. Y.; Kramer, J. H.; Miller, B. L.; Rankin, K. P.; et al. 2017. Emotion detection deficits and changes in personality traits linked to loss of white matter integrity in primary progressive aphasia. *NeuroImage: Clinical*, 16: 447–454.

Muñoz, M. 2013. AphasiaBank English Protocol PWA TCU Corpus.

Noda, K.; Yamaguchi, Y.; Nakadai, K.; Okuno, H. G.; and Ogata, T. 2015. Audio-visual speech recognition using deep learning. *Applied intelligence*, 42: 722–737.

Novack, M. A.; and Goldin-Meadow, S. 2017. Gesture as representational action: A paper about function. *Psychonomic Bulletin & Review*, 24: 652–665.

Oomen, C. C.; Postma, A.; and Kolk, H. H. 2001. Prearticulatory and postarticulatory self-monitoring in Broca's aphasia. *Cortex*, 37(5): 627–641.

Poggi, I. 2008. Iconicity in different types of gestures. *Gesture*, 8(1): 45–61.

Pritchard, M.; Dipper, L.; Morgan, G.; and Cocks, N. 2015. Language and iconic gesture use in procedural discourse by speakers with aphasia. *Aphasiology*, 29(7): 826–844.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, 28492–28518. PMLR.

Ramage, A. 2013. AphasiaBank English Protocol PWA UNH Corpus.

Richardson, J. 2008. AphasiaBank English Controls Richardson Corpus.

Rohlfing, M. L.; Buckley, D. P.; Piraquive, J.; Stepp, C. E.; and Tracy, L. F. 2021. Hey Siri: How effective are common voice recognition systems at recognizing dysphonic voices? *The Laryngoscope*, 131(7): 1599–1607.

Salama, E. S.; El-Khoribi, R. A.; and Shoman, M. E. 2014. Audio-visual speech recognition for people with speech disorders. *International Journal of Computer Applications*, 96(2).

Sampson, M.; and Faroqi-Shah, Y. 2011. Investigation of self-monitoring in fluent aphasia with jargon. *Aphasiology*, 25(4): 505–528.

Sanguedolce, G.; Naylor, P. A.; and Geranmayeh, F. 2023. Uncovering the Potential for a Weakly Supervised End-to-End Model in Recognising Speech from Patient with Post-Stroke Aphasia. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, 182–190.

Schwartz, M. F.; Middleton, E. L.; Brecher, A.; Gagliardi, M.; and Garvey, K. 2016. Does naming accuracy improve through self-monitoring of errors? *Neuropsychologia*, 84: 272–281.

Silverman, M. 2013. AphasiaBank English Protocol PWA TAP Corpus.

Simmons-Mackie, N.; and Cherney, L. R. 2018. Aphasia in North America: highlights of a white paper. *Archives of Physical Medicine and Rehabilitation*, 99(10): e117.

Stark, B. 2022. AphasiaBank English CC Stark Corpus.

Stark, B. 2023. AphasiaBank English Aphasia NEURAL Corpus.

Stark, B. C.; and Cofoid, C. 2022. Task-specific iconic gesturing during spoken discourse in aphasia. *American Journal of Speech-Language Pathology*, 31(1): 30–47.

Stark, B. C.; and Oeding, G. 2023. Demographic, neuropsychological, and speech variables that impact iconic and supplementary-to-speech gesturing in aphasia. *Gesture*, 22(1): 62–93.

Strik, H.; Truong, K.; De Wet, F.; and Cucchiarini, C. 2009. Comparing different approaches for automatic pronunciation error detection. *Speech communication*, 51(10): 845–852.

Szabo, G. 2013. AphasiaBank English Protocol PWA Adler Corpus.

Thompson, C. 2013. AphasiaBank English Protocol PWA Thompson Corpus.

Tompkins, C. A.; Scharp, V. L.; and Marshall, R. C. 2006. Communicative value of self cues in aphasia: A re-evaluation. *Aphasiology*, 20(7): 684–704.

Tong, H.; Sharifzadeh, H.; and McLoughlin, I. 2020. Automatic assessment of dysarthric severity level using audio-video cross-modal approach in deep learning.

van Nispen, K.; van de Sandt-Koenderman, W. M. E.; and Krahmer, E. 2017. Production and comprehension of pantomimes used to depict objects. *Frontiers in Psychology*, 8: 1095.

Wasay, M.; Khatri, I. A.; and Kaul, S. 2014. Stroke in south Asian countries. *Nature reviews neurology*, 10(3): 135–143.

Whiteside, J. 2013. AphasiaBank English Protocol PWA Whiteside Corpus.

Williamson, D. 2013. AphasiaBank English Protocol PWA Williamson Corpus.

Wilson, S. 2021. AphasiaBank English Protocol PWA APROCSA Corpus.

Wozniak, L. 2013. AphasiaBank English Protocol PWA Wozniak Corpus.

Wright, H. 2013. AphasiaBank English Protocol PWA Wright Corpus.

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6): 82–88.