Speech-to-ASL-like Sign Output System

1/11/2026

Abstract

This project develops a rule-based prototype system that translates spoken English into American Sign Language (ASL) gloss and corresponding sign videos. The system first converts

speech to text using OpenAI's Whisper (medium model) and then normalizes the text by removing punctuation and standardizing formatting. SpaCy is used to parse the grammatical structure of each sentence, identifying subjects, verbs, objects, and time expressions. A rule-based algorithm rearranges these components into ASL-like gloss following the Time + Topic + Comment structure. Each gloss token is matched to a sign video in a JSON dictionary using fuzzy string matching with an 85% similarity threshold. Real Gloss Coverage measures the percentage of English words correctly converted to gloss tokens and matched to videos. In contrast, Dictionary Coverage measures the percentage of gloss tokens present in the dictionary. A trial is considered successful if both metrics are at least 70%. The system was tested on 20 simple sentences and successfully passed 15 trials. Results indicate that the prototype works well on simpler sentences but is limited by dictionary size and simplified gloss rules. Overall, the system demonstrates potential as an accessible communication tool for Deaf and Hard-of-Hearing individuals and provides a foundation for future improvements.

Introduction

As of 2021, an estimated 11 million individuals living in the United States consider themselves deaf or are hard of hearing (National Deaf Center, 2024). On a global level, the World Health Organization (WHO) estimates that by 2050, more than 700 million people will have "disabling hearing loss". Deafness refers to the condition of profound or complete hearing loss, resulting in little to no functional hearing in affected individuals (WHO, 2025). It is a severe impairment, often necessitating the use of sign language, lip reading, and cochlear implants (used when traditional hearing aids prove inadequate) for communication. It is important to note that every Deaf individual has their own preferences regarding communication (National Deaf

Center, 2025). However, for this project, communication through American Sign Language (ASL), is being considered.

Around 500,000 people throughout the United States use ASL as their native language. Although lip reading can be an alternative to sign language communication, it can be considered ineffective. It is estimated that only about 30% of spoken English can be understood by lip-reading. Furthermore, it is unlikely that every individual participating in a conversation is skilled at lip-reading, making it an unreliable alternative (Commission on the Deaf and Hard of Hearing [CDHH], 2023). Another method of communication is text. A study conducted by Nicholas Epley and Amit Kumar showed that people often prefer texting for convenience or to avoid awkwardness. However, the study's findings have shown that communicating face-to-face improves the quality of connection and interaction, while having no tangible effect on perceived awkwardness (Kumar & Epley, 2020). Therefore, it follows that the Deaf or Hard-of-Hearing community might feel more recognized and accepted if individuals choose to communicate with them in sign language (Silent Sounds, 2025).

ASL is an incredibly visual language, with facial and spatial cues (Academic Center for Excellence, 2023). In ASL, the topic (noun) of the sentence, and any descriptions of the topic (adjectives), precede the comment (rest of the sentence). When forming cause-and-effect sentences, the cause is placed before the effect. When setting a scene, the signer goes from the most general to the specific details. Unlike spoken English, verbs are not conjugated based on tense. Instead, the tense is indicated by including a time indicator (e.g., yesterday, tomorrow, etc.), usually at the beginning of the sentence before the topic. The Object-Subject-Verb (OSV) order follows the "Time + Topic + Comment" structure. A sentence may also be arranged in the Subject-Verb-Object (SVO) order, in which case the subject comes first in the sentence. When

signing questions in ASL, the WHO, WHAT, WHEN, WHERE, WHY, WHICH, and/or HOW sign is placed at the end of the sentence. In ASL, these signs are used only for questions. Negating signs, such as NOT or NONE, generally follow the words they are negating (Academic Center for Excellence, 2023).

A non-manual marker (NMM) is similar to how tone of voice adds meaning to spoken words (Academic Center for Excellence, 2023). NMMs add context and meaning to signs, and include facial expressions and body language. These are particularly important when it comes to asking questions. Signs expressing emotion are also accompanied by facial expressions. Facial expressions can also provide extra detail to the sign or indicate negation. ASL contains three types of verbs: non-directional, one-directional, and multi-directional. Non-directional verbs are signs where movement does not indicate the subject/object of the sentence. One-directional verbs do not indicate the subject of the sentence, but do indicate the direct object, and multi-directional verbs express both.

Methods

Due to initial set limitations of speech-to-ASL and the numerous ways in which ASL differs from traditional languages (i.e., facial, spatial, and specific hand signs are crucial), although these 3D cues are extremely important, comprising around 50% of the ASL language, it is not within the current scope and is very expensive and difficult to achieve. As a result, these factors are being omitted for this project. Instead, this system is focused on attempting to mimic the remaining 50% of ASL (natural grammar order and basic hand signs).

The primary goal of this project is to construct a system that is capable of taking speech as input, translating (matching) it to ASL-like gloss, and displaying the corresponding sign

videos. The medium model of OpenAI's Whisper was imported into Google Colab. Whisper utilizes Automatic Speech Recognition (ASR) and is a valuable tool for converting spoken language into digital transcriptions. First, the audio from the user is converted into a log-Mel spectrogram. This representation allows the model to capture relevant time–frequency characteristics of the signal, which are then used for training and testing. In the spectrogram, the x-axis represents time, the y-axis corresponds to frequency bands, and the intensity (brightness) of each point indicates the energy present at a given frequency. In the creation of this system, the medium model of Whisper AI was used. It is nearly as efficient as the "large" model and is capable of transcribing multiple languages into English. Furthermore, this model is considerably more accurate than the "base" or "tiny" model. Although the use of Whisper AI has the disadvantage of higher latency–especially when the model is given larger, more complex sentences–it ensures the initial transcription is accurate, which is crucial for the subsequent steps.

Once the text is transcribed, it is then normalized to ensure the text is standardized before being sent into the natural language processing (NLP) system. During normalization, the transcribed phrase or sentence is converted into lowercase, punctuation is removed, and spacing is corrected. Although ASL is a visual language, its sentence structure can be mimicked or approximated (e.g., OSV and SVO orders). To accomplish this, spaCy is used as an NLP parser to analyze the grammatical structure of each sentence. It identifies parts of speech such as nouns, verbs, and subjects, and determines how words relate to each other in the sentence (e.g., which word is the subject or object of a verb). This allows the system to extract important components like time words, locations, subjects, and actions, and then reorganize them into an ASL-like gloss order. It is important to note that spaCy does not translate the sentence itself, but rather, provides the linguistic structure needed for the rule-based system to rearrange and simplify the sentence.

After the sentence is processed by the NLP parser, the system converts the English sentence into a list of ASL gloss tokens. ASL gloss is a simplified, word-based representation of sign language that removes unnecessary English grammar, such as verb tenses, articles, and filler words, and rearranges the remaining words into a structure closer to ASL. The program extracts key elements such as time expressions, locations, subjects, verbs, and objects, then reorders them into an ASL-like sequence. Each of these elements becomes a "gloss token," representing one concept that should be signed. This step is rule-based and approximate, meaning it does not produce perfect ASL, but it creates a structured, simplified representation that can be mapped to sign videos.

Once the ASL gloss tokens are generated, each token must be matched to a corresponding sign video stored in a dictionary (JSON file). Taking into consideration that real-world text may contain spelling variations, small errors from speech recognition, or slightly different word forms, the system uses fuzzy string matching instead of exact matching. Fuzzy matching computes a similarity score between the gloss token and every available key in the sign database and selects the closest match above a chosen threshold. This allows the system to find the correct sign even if the token is not an exact match (e.g., matching "hyngry" to "hungry"). However, this also introduces a trade-off: if the threshold is too low, incorrect signs may be chosen; if it is too high, some valid signs may not be found. To combat this, the system has a set threshold of 85%, meaning most of the phrase matching is close to the original string or word. However, it's important to note that incorrect words may occasionally be chosen, especially if the speech-to-text transcription contains errors.

After the system finds the best matching sign for each ASL gloss token, it plays the corresponding sign language video file. The videos are displayed one-by-one in the order of the

generated ASL gloss sentence, creating the impression of a continuous signed sentence. Each video represents a single word or concept in sign language. The program uses a video player (OpenCV) to load and display each clip on the screen. In this way, the system converts a text or speech input into a sequence of visual sign language demonstrations.

The model will be given 20 simple sentences (one sentence per trial). Its success will be evaluated using two values: Real Gloss Coverage and Dictionary Coverage. The Real Gloss Coverage value is the percentage of English words in a sentence that were correctly mapped to their corresponding ASL gloss tokens and matched to a sign video in the system. Dictionary Coverage will be determined by the percent of words in the model's predicted ASL gloss that exist in the JSON dictionary file. For a trial to be considered successful, both the Real Gloss Coverage and Dictionary Coverage must be 70% or greater.

Results/Discussion

The model passed 15/20 trials. Of the trials that failed, 60% had a Real Gloss Coverage value greater than 70%. Only one of the unsuccessful trials passed Dictionary Coverage but failed Real Gloss Coverage. The lowest value recorded for Real Gloss Coverage was 60%, whereas the lowest value for Dictionary Coverage was 0%. The highest value for both variables was 100%, which occurred 11 times for Real Gloss Coverage, 12 times for Dictionary Coverage, and 8 times simultaneously.

Based on these results, it can be determined that neither the system's dictionary coverage nor ASL gloss is satisfactory given the criteria established. Since both variables are considered simultaneously, it is reasonable to state that the low dictionary coverage values generated in 60% of the failed trials were the reason the system was unsuccessful. If the model had been given

access to a wider range of sign language vocabulary, it may have performed successfully in these trials. However, if the variables are taken separately, there is a slightly higher occurrence of 100% values in Dictionary Coverage as opposed to Real Gloss Coverage. This indicates that the algorithm used to convert text into ASL gloss should be improved as well.

Conclusion

While the system didn't successfully pass 17 out of 20 trials, it came close to meeting the expected accuracy, with 15 out of 20 trials passed. Therefore, the model shows promise as a speech-to-ASL system. Currently, the limitations of the model include a limited dictionary, modest grammar and sentence structure, and an inability to communicate in some contexts. These shortcomings can be addressed in future iterations of this model by expanding the JSON dictionary and simplifying the system's pipeline by using and training an AI model (such as T5). This will not only improve the model's accuracy, but also its flexibility and ability to handle complex sentences. Overall, despite its limitations, this model has numerous real-world applications as it can be used as an emergency communication device in the absence of reliable internet or Wi-Fi.

References

American Sign Language | Commission on the Deaf and Hard of Hearing (CDHH). (2023, September 8). Cdhh.ri.gov. https://cdhh.ri.gov/information-referral/american-sign-language.php

Center, N. D. (2024, September 30). *How many deaf people live in the United States? - National Deaf Center Knowledge Base*. Nationaldeafcenter.org. https://help.nationaldeafcenter.org/article/51-how-many-deaf-people-live-in-the-united-states

Communicating with Deaf People. (2025, May). National Deaf Center. https://nationaldeafcenter.org/resources/deaf-awareness/communicating-with-deaf-people/

Kumar, A., & Epley, N. (2020). It's surprisingly nice to hear you: Misunderstanding the impact of communication media can lead to suboptimal choices of how to connect with others. *Journal of Experimental Psychology: General*, *150*(3). https://doi.org/10.1037/xge0000962

Mitchell, D. (2013). *American Sign Language ASL Dictionary*. Signasl.org. https://www.signasl.org/

OpenAI. (2022). *Introducing Whisper*. Openai.com. https://openai.com/index/whisper/

Provided by ASL Grammar Guide The Academic Center for Excellence 1 ASL Grammar Guide. (2023). https://germanna.edu/sites/default/files/2023-07/ASL%20Grammar%20Guide%20%28edit%207-24-23%29.pdf

The importance of everyday communication with Deaf people | Silent Sounds. (2025). Silent-Sounds.co.uk. https://silent-sounds.co.uk/the-importance-of-everyday-communication-with-deaf-people/

World Health Organization. (2025, February 26). *Deafness and hearing loss*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss