A Performance Comparison Between Two Speech-to-ASL-Gesture-Projection Translation Implementations

Alexandra Kashani Motlagh
Computer Science and Engineering
The University of Texas at Arlington
Arlington, Texas, USA
alexandra.kashanimotlagh@uta.edu

Shikha Mehta
Computer Science and Engineering
The University of Texas at Arlington
Arlington, Texas, USA
shikha.mehta@uta.edu

Ahmed Shah Rahid Syed
Computer Science and Engineering
The University of Texas at Arlington
Arlington, Texas, USA
axs0743@mavs.uta.edu

Ishfaq Ahmad
Computer Science and Engineering
The University of Texas at Arlington
Arlington, Texas, USA
iahmad@cse.uta.edu

Addison Clark
Computer Science and Engineering
The University of Texas at Arlington
Arlington, Texas, USA
addison.clark@mavs.uta.edu

Abstract— **Millions of people with hearing disabilities use sign language for communication, creating a communication gap with those who are not fluent in ASL (American Sign Language). This paper aims to introduce an ASL interpreter system using a smart-glasses-based augmented reality system. We begin by introducing and comparing two models that translate spoken language into ASL poses. The first system translates spoken text to ASL Gloss, an intermediate representation, before generating ASL poses. The second system directly translates the text to ASL poses. Our analysis shows that using ASL Gloss as an intermediate step significantly improves the translation speed. We then explore a system of encoding ASL pose videos for display on smart glasses. The chosen translation method has a BLEU score of 66.5801 and a rate of 1.825 milliseconds per gloss translation. Our algorithm for mapping gloss text to ASL videos obtained a mean squared error of 0.05, indicating that our system has good translational accuracy and a low mapping error.**

Keywords—Sign language, augmented reality, AR, smart glasses

I. INTRODUCTION

Hearing loss is a global phenomenon that affects millions worldwide and in the United States. One in eight people in the U.S. aged 12 and older has hearing loss in both ears, which equates to 30 million people [11]. Many individuals with permanent hearing loss may choose to learn and communicate in ASL (American Sign Language) as their natural first language. According to a 2018 study, adult sign language use was substantial (2.80%), with respondents with complete hearing loss having a far higher rate of sign language use than any other hearing acuity group [9].

The World Health Organization (WHO) projects that over 1 billion young adults are at risk of permanent, avoidable hearing loss due to unsafe listening practices [4]. To the effect of growing cases of hearing loss nearing the future, the U.S. Bureau of Labor Statistics expects the employment of interpreters and translators to grow 4% from 2020 to 2032 [8].

Moreover, hearing loss is likely to impact many areas of life on an individual and national basis. Individually, hearing loss can impact communication, cognition, education, employment, social interaction, and mental health. On a larger scale, hearing loss incurs economic losses in healthcare, education, and productivity. Specifically, WHO has calculated a $980 billion annual cost due to neglected hearing impairment issues [4]. Therefore, there can be both personal and national implications to a lack of effective assistive technologies for people with hearing impairments, suggesting a strong need for improving our current sign language-enhanced devices. Optimizing our current state-of-the-art speech to ASL pose translators is relevant, as a handy artificially intelligent ASL interpreter proves a cost-effective option for many.

The population of individuals with permanent hearing loss who use ASL as their natural language and special educators who would like an applicable way to learn ASL will benefit from Smart Glasses equipped with a state-of-the-art, real-time speech-to-ASL pose translator rendered in their augmented reality (AR) view.

Individuals who want to learn ASL to communicate with their ASL-speaking peers or loved ones can benefit from this software. On the same token, ASL-speaking individuals can use this software as a personal on-the-go interpreter in case they prefer audio translated to ASL poses rather than captioned text.

As the rate of total hearing loss is expected to rise in the coming decades and employment of interpreters begins to increase, the hiring prices of interpreters are also expected to rise, or a price market will form in turn. A personal free speech to ASL pose translator could succeed in hiring a real interpreter.

This project was funded by the National Science Foundation under Award Number:1757641.

The layout of the paper is as follows: Section II gives a brief overview of related work. Section III outlines our proposal. Section IV describes our experimental setup and results. Section V provides our conclusions and potential future work.

II. Related Work

In [3], Othman and Jemni implement the state-of-the-art text-to-ASL pose translator, utilizing an ASL Gloss translation intermediate step. Other researchers, such as Stoll [6], implement translating text to signed poses, also with an embedded text to gloss stage for performance speedup.

Several other works explore sign language applications for smart glasses or AR. [1], [10], and [13] evaluate technology as a learning tool for sign language students. Other research, such as [14], performs sign language recognition using smart glasses cameras. There is much additional research in using computer vision or wearable devices to translate sign language to text, but this is outside of the scope of this paper. Our focus is on translating speech to ASL to develop an AR interpreter.

III. Proposal

Our proposal is a Smart Glasses system that will translate speech to ASL poses, which will be displayed in an AR view. To this goal, a performance comparison between two speech-to-ASL pose translations is done. To preface, a sign language gloss, here “ASL gloss,” is an abbreviation of text such that the remaining words each correlate to a sign language pose. For example, “of,” “the,” and other terms would be filtered out in the gloss translation process, as they do not map to a signed gesture. The first implementation we compare converts speech to text to ASL Gloss to ASL poses. The second implementation converts speech to text to ASL poses, skipping the ASL gloss step. The latter will have more cluttered input in handling ASL pose translation, while the former will experience speedup while reading information due to its shorter length. Thus, the first implementation's text-to-gloss translation should have a duration faster than the second’s text-to-pose translation. The system with the best performance will then be used for a Smart Glasses AR translation module, which will allow sign language users and learners to translate speech to ASL in everyday scenarios.

IV. Experiments

Our experiment is a performance comparison between two implementations of speech to ASL pose translators, labeled implementations 1 and 2 for this paper. Implementation 1 translates from audio to text to ASL Gloss to ASL poses. Implementation 2 translates from speech to text to ASL poses and skips ASL gloss generation. Our experiments show that using an English text-to-ASL gloss Natural Language Processing (NLP) learning model, specifically t5-small, speeds up text-to-ASL pose translation. Our proposal is a cumulative speech-to-ASL pose software package that employs the better implementation of implementations 1 and 2 listed above.

A. Speech to Text and Gloss Modules

Our first module is for speech to text. We used the state-of-the-art speech-to-text software Speech Recognition and pyttsx3 for generating text from speech input. In our project, we intend to pass the text as a parameter for interpreting the ASL gloss model in Implementation 1 and the ASL pose interpreter for Implementation 2.

After generating text from the PyAudio speech-to-text translator, the text in Implementation 1 must be transformed into an ASL Gloss or a contextual abbreviation of the text. In selecting the best text-to-gloss transformer, we favored a high BLEU score to indicate high similarity between the model-generated and reference texts [5] and a shorter duration average processing time in seconds per gloss. These two performance metrics are necessary for defining the model's latency and text accuracy.

We first implemented an open-source pretrained text-to-gloss transformer model from [7] to determine the best model for translating text to ASL Gloss. Yet, due to a low BLEU score and poor unimproved average time post-fine-tuning, we trained the same t5-small model from scratch. Post-fine-tuning, the new model yielded an improved BLEU score and average time per gloss.

Thus, this section evaluates the performance of two t5-small models on a shared dataset. For simplicity, in this section, the open-source model provided by [7] is referred to as 'Model 1', and the t5-small model supplied by us is referred to as 'Model 2'.

A concluding additional contribution to the proposal is a performance-improved finely tuned t5-small text-gloss model, or Model 2.

HuggingFace's Synthetic English-ASL Gloss Parallel Corpus 2012 (ASLG_PC12) [2] was used for training both Model 1 [7] and Model 2.

As mentioned, we first implemented Model 1 from [7]. This model is a fine-tuned version of the t5-small learning model for Text2Text Generation. White’s paper reports a loss of 0.5811 using the Cross-Entropy Loss function [7], indicating a converging training process and thus improved responses. The reported BLEU score is 56.4281 [7], indicating average similarity to the target text. White confirmed the model's Generation Length or average number of tokens per generated text as 15.5526 [7]. The generation length is consistent throughout all trained t5-small models in the text-gloss experiment.

We implemented extra fine-tuning for both models, and for training Model 2. After importing the ASLG_PC12 dataset from HuggingFace, the first step to data preprocessing is making all columns in the dataset lowercase, as all text samples matching in case make it easier for the model to detect and learn differences when processing the text. Thus, [english("adjournment of the session"), ASL Gloss("ADJOURNMENT SESSION")] becomes [english("adjournment of the session"), ASL Gloss("adjournment session")].

The 'batched' flag is also set to true as a data preprocessing step, which groups the data into batches of dictionaries

containing fields: "input_ids" for numerically representing the tokens in the "labels" field, "attention_mask" for prioritizing learning performance of specific tokens, and "labels" for storing the target output text as a tokenized sequence. Moreover, it enables the data collator class to pad each element's token sequence for each batch dynamically. Dynamic Padding sets the length of every element in the batch to the size of the longest string. T5-small models process each example of data as a list of 512 tokens. The first tokens in the list are the sequentially tokenized words, followed by null tokens as fillers to the maximum number of tokens. Since the reported average generation length of the dataset is 15.5526 [7], only approximately 16 tokens need to be evaluated. Thus, Dynamic Padding is employed as a data preprocessing step to speed up training and evaluation by efficiently using GPU resources.

B. ASL Gloss Model Training

After the aforementioned preprocessing for both Model 1 and Model 2, the dataset undergoes tokenization and initialization with an instance of a Data Collator after getting converted to PyTorch format.

For evaluating and comparing text-to-gloss models 1 and 2, we are using average generation time and their BLEU scores. We favor average generation time for determining translation stage duration, such as translating text to ASL Gloss. We use the BLEU Score to observe the similarity between the generated and target texts. After all data is processed, the total time is divided by the number of elements in the dataset, yielding the average generation time in seconds per ASL Gloss. A plotting of loss per epoch using the Cross-Entropy Loss function is obtained post-training.

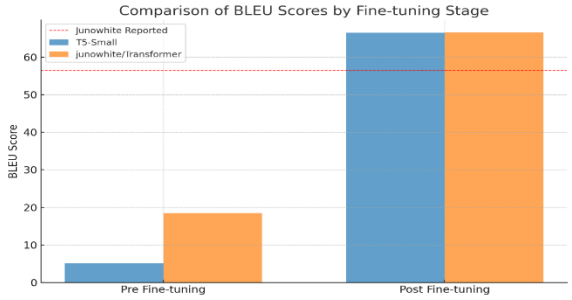

Fig. 1. Comparison of BLEU scores by fine-tuning stage.

This text-to-gloss portion of the paper explores the performance of trained t5-small models. It contributes a finely tuned text-to-gloss t5-small model or an open-source text-to-gloss competitor alongside Model 1 and an optimized version of Model 1.

Fig. 1 summarizes the performance results of the optimized Models' BLEU scores before and after fine-tuning. The BLEU score reported by the author for Model 1 was 56.4281, shown as the Junowhite Reported line on the plot. Our system observed a BLEU score of 18.4829 pre-fine-tuning and 66.5801 post-fine-tuning. The contributed optimizations generated from this experiment raised Model 1's BLEU Score by 10.152 points, shown in Fig. 2. The BLEU score of Model 2 is 5.0671 pre-fine-tuning and 66.5801 post-fine-tuning.

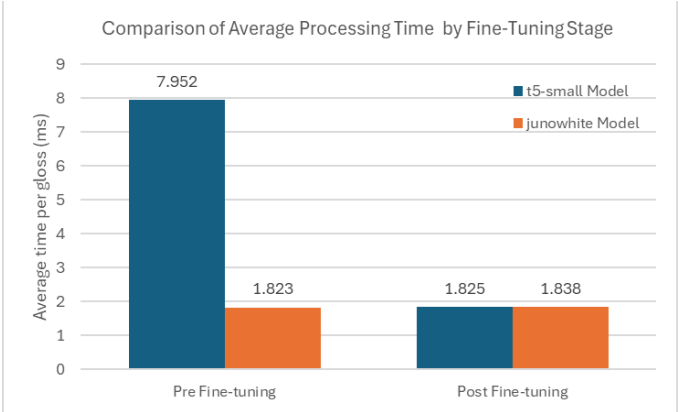

Fig. 2. Comparison of average processing time by fine-tuning stage.

Fig. 2 summarizes the performance results of the Models' average generation time in seconds per ASL Gloss before and after fine-tuning. The average generation time is relatively the same across both trained models and their pre- and post-fine-tuning stages. Model 1 had an average generation time of 1.823ms/gloss before and 1.838ms/gloss after fine-tuning. Model 2 had an average generation time of 7.952ms/gloss before and 1.825ms/gloss after fine-tuning. On both models, the average generation time in seconds per gloss slightly increased after fine-tuning as the BLEU score increased and thus reflected more similarity to the target.

Fig. 3 plots the loss of the contributed, optimized version of Model 1 and Model 2 using the t5-small built-in Cross-Entropy-Loss function. The models share similar convergence rates.

In conclusion, our text-to-ASL Gloss findings and performance tests contribute an optimized version of a T5-small model for text-to-text generation. The additional preprocessing step of making each text lowercase improved the BLEU Score by approximately 10 points. The custom-trained t5-small model yielded a lower BLEU score than the pre-trained Junowhite model, yet both models resulted in the same BLEU score post-fine-tuning. The average generation time was consistent across both models, before and after fine-tuning. Lastly, the plots of both models converge in sync to 0. The final contribution to our more significant research from this section is providing an improved open-source text-to-gloss model.

Fig. 3. Loss over epochs for two T5-Small models.

C. Text to ASL Pose Module

"World-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison" (WLASL) provided an effective source from which to map gloss words to ASL gestures through videos [12]. The dataset includes about 64,000 ASL gesture videos with a file providing its corresponding URL, bounding box, frame rate, starting frame, ending frame, instance id, signer id, subset, source site, dialect variation id, and unique video identifier. Considering the large size of the dataset, we decided to focus on encoding the video URLs, so that it was not integral to download all of the videos. After collecting the data, we analyzed our set to determine steps for processing.

Processing the WLASL dataset for a recurrent neural network (RNN) model was twofold in that it required cleaning and encoding. Once the dataset was cleaned to better serve our needs while encoding the glosses and video URLs, we researched and selected new libraries to begin encoding our data. The URL and gloss data was then encoded into vector and numerical values in preparation for training.

TABLE I. KERAS VGG16 MODEL'S VIDEO TO VECTOR ENCODING RATE

Trials	Average Video Encoding Rate (s/step)
1	5.7
2	8.39

Table 1 describes the multiple trials involved in determining the efficiency of using Keras' VGG16 model to encode video URLs into vector values in order to create a numerical array to feed into the RNN model. Descriptive statistics and terminal output were used to collect data in s/step and determine the mean value.

D. Text to ASL Pose Model Training

In order to map English to glosses to ASL gestures, we developed an RNN model to map the words to videos. We used Keras' built-in RNN layers, specifically employing the Sequential model to process the integer encodings into 64-dimensional vectors [7]. This model is usually evaluated using mean squared error.

TABLE II. TEXT TO ASL POSE MEAN SQUARED ERROR

Test Set Size	Mean Squared Error
5	0.12
15	0.09
30	0.09
50	0.05

Table 2 highlights how the mean squared error decreased for larger test sets, indicating more accuracy in the RNN model's prediction with larger sample sizes. The mean squared error for the RNN model was relatively low, considering that a 0 mean squared error value indicates a perfect model. Additionally, it was promising to note that the mean squared error generally functions inversely with sample size, indicating that this model could work on even large amounts of input data.

E. ASL Pose to Vuzix Blade Projection

In this section, we present the tangible results obtained from our experiments, which focus on seamlessly translating American Sign Language (ASL) signs into immersive visual projections using Vuzix Blade augmented reality (AR) glasses. This innovative approach not only enhances the accessibility of ASL communication but also opens new possibilities for immersive language learning and communication experiences.

Before the ASL signs from the WLASL dataset could be transformed into visually immersive projections, a series of preprocessing steps was performed to optimize the data. First, we converted the video data from the WLASL dataset into a format compatible with Vuzix Blade glasses, ensuring seamless playback. Real-time synchronization of video data with the Vuzix Blade's display frame rate was essential for seamless projection. The synchronization process minimized any perceivable delays between sign selection and projection display. We then identified the start and end points of individual ASL signs within the video streams. This segmentation allowed for precise projection of each sign.

The final piece of our experiments lay in the pose-to-projection mapping module, the first component of which is the video playback mechanism. Using the playback mechanism, ASL sign videos are able to be overlayed onto the user's real-world view. Finally, Augmented Reality Markup Language (ARML) was used to encode instructions on how and where to project the ASL videos. ARML allowed for precise placement and scaling of sign videos based on user interactions.

V. CONCLUSIONS AND FUTURE WORK

The proposal discussed in this paper compares two implementations of translating speech to ASL gestures, rendered in real-time, utilizing real-time latency and accuracy to assess performance. Both implementations used Speech Recognition and pyttsx3 for speech-to-text translation. In implementation 1, this was input into the t5-small-trained model to generate the ASL gloss. After, the text is translated into ASL poses using an RNN.

We tested the two implementations of speech-to-ASL pose translation methods. Implementation 1 including the ASL gloss intermediary step, and implementation 2 translating English text directly to ASL poses. Our findings show that implementation 1 outperformed implementation 2 in terms of speed and accuracy. We additionally tested two text-to-ASL translation models and found that our fine-tuning steps for the models improve the translation time in s/gloss and the model's BLEU score. Using the better performing implementation and translation models, we also develop a method of displaying the ASL pose videos on AR glasses so that the translation can be viewed in real time.

The future work for this proposal is fully testing the first implementation as a smart glasses application to determine the feasibility and usability of a seamless real-time interpreter via AR glasses. The future work concerning the speech-to-ASL pose translation problem is improving the text-to-ASL gloss learning model so that it can convert long text into

grammatically correct and correlating ASL glosses for ASL pose generation.

ACKNOWLEDGMENT

This project was funded by the National Science Foundation under Award Number:1757641.

REFERENCES

[1] A. Miller, J. Malasig, B. Castro, V. Hanson, H. Nicolau, and A. Brandao, "The Use of Smart Glasses for Lecture Comprehension by Deaf and Hard of Hearing Students." *ACM,* 2017, pp. 1909-1915.

[2] A. Moryossef, "Synthetic English-ASL Gloss Parallel Corpus 2012," HuggingFace, https://huggingface.co/datasets/aslg_pc12 (accessed Sep. 15, 2023).

[3] A. Othman and M. Jemni, "Designing high accuracy statistical machine translation for sign language using Parallel Corpus," *Journal of Information Technology Research*, vol. 12, no. 2, pp. 134–158, 2019. doi:10.4018/jitr.2019040108

[4] "Deafness and Hearing Loss," *World Health Organization,* February 2023.

[5] E. Reiter, "A structured review of the validity of BLEU," *Computational Linguistics - Association for Computational Linguistics*, vol. 44, no. 3, pp. 393–401, Sep. 2018, doi: 10.1162/coli_a_00322.

[6] G. Varol, L. Momeni, S. Albanie, T. Afouras, and A. Zisserman, "Scaling Up Sign Spotting Through Sign Language Dictionaries," *Springer International Journal of Computer Vision*, vol. 130, 2022, pp. 1416–1439.

[7] "junowhite/transformer_model - Hugging Face," Jan. 09, 2001. https://huggingface.co/junowhite/transformer_model (accessed Sep. 15, 2023).

[8] J. Wu, L. Sun and R. Jafari, "A Wearable System for Recognizing American Sign Language in Real-Time Using IMU and Surface EMG Sensors," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, September 2016, no. 5, pp. 1281-1290.

[9] R. E. Mitchell and T. A. Young, "How many people use sign language? A National Health Survey-Based estimate," *Journal of Deaf Studies and Deaf Education*, vol. 28, no. 1, pp. 1–6, Nov. 2022, doi: 10.1093/deafed/enac031.

[10] S. Al-Megren and A. Almutairi, "Assessing the Effectiveness of an Augmented Reality Application for the Literacy Development of Arabic Children with Hearing Impairments," *Cross-Cultural Design. Applications in Cultural Heritage, Creativity and Social Development Lecture Notes in Computer Science()*, vol. 10912, June 2018 .

[11] S. R., S. M., S. K., and A. Thilagavathy, "AI-Powered Smart Glasses for Blind, Deaf, and Dumb," *2022 5th International Conference on Advances in Science and Technology (ICAST)*, Mumbai, India, 2022, pp. 280-285.

[12] S. Zhu and F. Chollet, "Working with RNNs," *TensorFlow*. https://www.tensorflow.org/guides/keras/working_with_rnns (accessed Sep. 15, 2023).

[13] V. Falvo, L. P. Scatalon and E. Francine Barbosa, "The Role of Technology to Teaching and Learning Sign Languages: A Systematic Mapping," *2020 IEEE Frontiers in Education Conference (FIE*), Uppsala, Sweden, 2020, pp. 1-9.

[14] Y. Jin, S. Choi, Y. Gao, J. Li, Z. Li, and Z. Jin, "TransASL: A Smart Glass Based Comprehensive ASL Recognizer in Daily Life Proceedings of the 28th International Conference on Intelligent User Interfaces," *ACM*, March 2023, pp. 802-818.