INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 10, October 2023, pp : 160-164

e-ISSN : 2583-1062

Impact Factor : 5.725

A MACHINE LEARNING FRAMEWORK AND METHOD TO TRANSLATE SPEECH TO REAL-TIME SIGN LANGUAGE FOR AR GLASSES

Rahul Solleti[1]

[1]North Carolina School of Science and Mathematics.

DOI: https://www.doi.org/10.58257/IJPREMS32125

ABSTRACT

Communication challenges persist among individuals with hearing disabilities, given the limited prevalence of sign language proficiency within the general population. This research endeavors to devise an approach, with the overarching goal of ameliorating communication hurdles faced by those with hearing impairments by translating the speech into configurable Sign Language (cSL). Acknowledging the arduous nature of obtaining sign language skills, this study introduces a viable solution through the amalgamation of speech recognition, image processing technologies and machine learning solutions. The evolution of sign languages has significantly augmented communication accessibility for the deaf and hard of hearing communities. Within this scholarly endeavor, we propose the implementation of a real-time system proficient in recognizing speech through Recurrent Neural Networks (RNN). This system further embraces the sequence-to-sequence learning to transmute recognized speech into textual form. Subsequently, the text undergoes translation into cSL using machine learning framework, culminating in its manifestation as a series of images, seamlessly presented through augmented reality glasses.

1. INTRODUCTION

Over 1.5 billion people (nearly 20% of the global population) live with hearing loss, and about 430 million of them have disabling hearing loss [1]. Sign language is a primary means of communication for deaf and hard of hearing people, and there are over 300 different sign languages used throughout the world today [2]. Sign language empowers disabled people to participate in various aspects of social life, but it relies on trained sign language interpreters, who often charge for their services and have limited availability.

This work presents a solution to this problem by recognizing speech and translating it to sign language before presenting in augmented reality (AR) glasses. Research and interviews with deaf people have shown that they can better understand translation if they can simultaneously see the speaker's mouth and sign language [3]. Additionally, around 30% of what is said can be read from the speaker's lips if the mouth movements are clear [4].

These findings led us to choose the approach of using AR glasses to display translated sign language. This allows deaf people to follow the situation while translation is being provided in their field of view, for example through an avatar. An AR avatar provides additional benefits for deaf people by allowing them to follow the translation as well as the speaker's facial expressions and gestures. This allows for more natural and engaging conversations, with eye contact and facial expressions.

2. RELATED WORK

Researchers have developed a variety of approaches to build or recognize sign languages and hand gestures from different parts of the world. Often researchers have focused on developing systems that can recognize sign languages that are unique to specific continents, countries, or regions. One such example is the Virtual Signing, Animation, Capture, Storage, and Transmission (ViSiCAST) translator, proposed by Bangham et al. [4]. This system translates English to British Sign Language (BSL) using Head-driven Phrase Structure Grammar (HPSG) and the Prolog-based freeware system Attribute Logic Engine (ALE) [5]. Another example is the American Sign Language (ASL) Workbench, developed in 2001. This text-to-ASL system uses Lexical Functional Grammar (LFG) to represent English in ASL [6, 8]. It is one of the most sophisticated machine translation systems and uses a phonological model based on the Movement-Hold principle of ASL phonology. imilarly, another text-to-ASL translation system that represents the source text in ASL using Synchronous Tree Adjoining Grammar (STAG) [7]. To recognize the correct word-sign pair, the system maintains a multilingual lexicon. The output of the system is a written ASL gloss notation. The synthesis module of the system then generates an animated human model.

Researchers in South America have also developed Brazilian Sign Language (LIBRAS), a natural language with its own grammar, syntax, and vocabulary, is used by deaf communities in urban Brazil. It is not simply a visual representation of Portuguese.

3. METHOD AND TECHNIQUES

Existing work on sign language translation has largely focused to enable communication within local deaf communities by targeting specific regional sign languages. However, the global connectivity of today's world creates a need for solutions that allow the deaf and hard-of-hearing to surpass geographical barriers. This work proposes an approach to make spoken communication accessible to sign language users irrespective of their location. We suggest utilizing techniques to automatically detect spoken language and translate corresponding text into the desired sign language. This global approach equips users with personalized communication accessibility as they travel or interact with global audience, enabled by a single pair of AR glasses.

Speech Transcription – Automatic speech recognition (ASR) facilitates real-time transcription of spoken language into text. Robust systems necessitate training general-purpose models on extensive and diverse speech datasets. We propose utilizing recent advances in multitask, multilingual ASR models to enable live transcription and translation. Specifically, Whisper, developed by OpenAI and trained on 680,000 hours of speech across 97 different languages and tasks [10], exemplifies a model capable of automatic language identification, speech recognition across multiple languages, and speech-to-text translation. Adoption of such models addresses the need for widely applicable, accurate, and low-latency speech transcription capabilities. Further research should explore optimizations for streaming speech input, as well as domain-specific customization to maximize accuracy for target use cases. Overall, leveraging the latest multitask ASR models represents a promising approach to enabling ubiquitous speech transcription services.

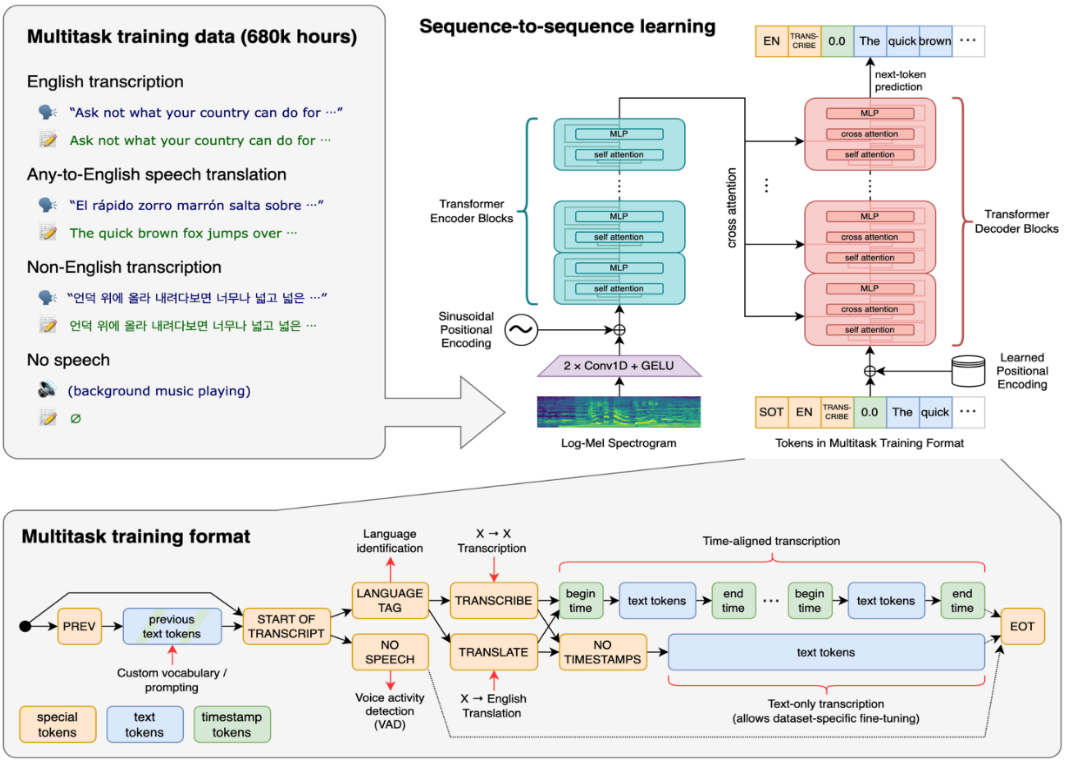
Multitask training data (680k hours)
English transcription
"Ask not what your country can do for ⋯"
Ask not what your country can do for ⋯
Any-to-English speech translation
"El rápido zorro marrón salta sobre ⋯"
The quick brown fox jumps over ⋯
Non-English transcription
"언덕 위에 올라 내려다보면 너무나 넓고 넓은 ⋯"
언덕 위에 올라 내려다보면 너무나 넓고 넓은 ⋯
No speech
(background music playing)
∅
Sequence-to-sequence learning
EN TRANS-CRIBE 0.0 The quick brown ⋯
next-token prediction
MLP
self attention
cross attention
Transformer Encoder Blocks
Transformer Decoder Blocks
Sinusoidal Positional Encoding
2 × Conv1D + GELU
Log-Mel Spectrogram
Learned Positional Encoding
SOT EN TRANS-CRIBE 0.0 The quick ⋯
Tokens in Multitask Training Format
Multitask training format
Language identification
X → X Transcription
Time-aligned transcription
PREV
previous text tokens
START OF TRANSCRIPT
LANGUAGE TAG
TRANSCRIBE
begin time
text tokens
end time
begin time
text tokens
end time
NO SPEECH
TRANSLATE
NO TIMESTAMPS
text tokens
EOT
Custom vocabulary / prompting
Voice activity detection (VAD)
X → English Translation
Text-only transcription (allows dataset-specific fine-tuning)
special tokens
text tokens
timestamp tokens

Figure1: Open AI Whisper Speech Recognition Model [13]

In addition to the Whisper, Meta recently open-sourced SeamlessM4T model represents another option for multilingual speech translation. SeamlessM4T is trained on over 100 languages [12] using a massive multitask framework spanning speech recognition, speech translation, and text translation.

Natural Language Processing (NLP) – The raw output of automatic speech recognition systems requires additional processing to ensure high-quality text. Natural language processing techniques can enhance recognized speech by correcting errors commonly encountered in spoken language. This includes text cleaning to handle disfluencies, punctuation insertion, and capitalization. Grammatical error correction is also beneficial for fixing malformed or unstructured sentences. Furthermore, sentiment analysis models could infer emotional cues to insert appropriate punctuation. Overall, passing ASR transcripts through NLP pipelines to perform tasks such as text cleaning, grammar correction, and sentiment-based formatting improves the cohesiveness, readability, and utility of the resulting text. Further gains can be achieved by optimizing NLP models to handle and correct the types of errors typified by live speech transcription. Targeted development of such speech-tuned natural language processing systems will enable higher-quality usable text output from speech recognition.

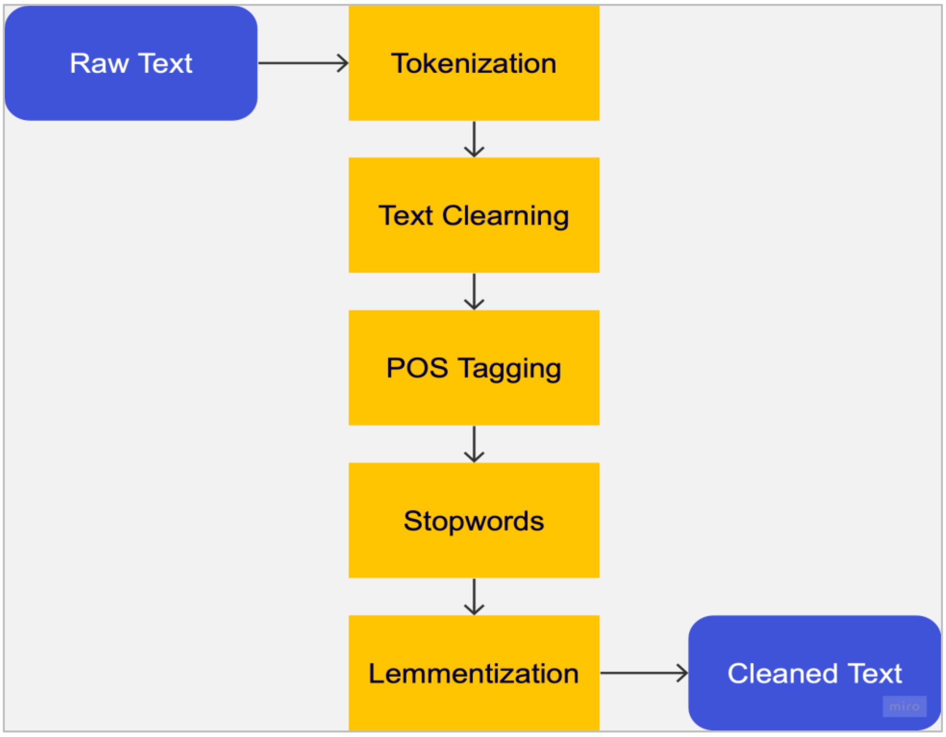

Figure2: Steps Involved in NLP

Sign Language Translation - After processing the speech recognition transcripts through natural language pipelines, the next step is translating the text into regional sign languages (configurable) for accessibility. Regional sign language translation necessitates labeled paired data of text and corresponding sign language sequences. With a curated dataset, machine learning and deep learning models such as recurrent neural networks (RNN) can be trained to map text to a regional sign language animation. Specifically, sequence-to-sequence models with encoder-decoder architectures are well-suited for this text-to-sign translation task. The sign output can be represented using 2D/3D animation libraries or avatar movements compatible with tools like Unity3D. Mapping the model outputs to control avatar motion generates an intuitive visualization of the sign language translation. Tailoring such translation models to each regional sign language and training on available text-sign data enables extending the accessibility of the speech recognition system. Priority should be placed on collaborating with the Deaf community to ensure accurate and representative sign language datasets.

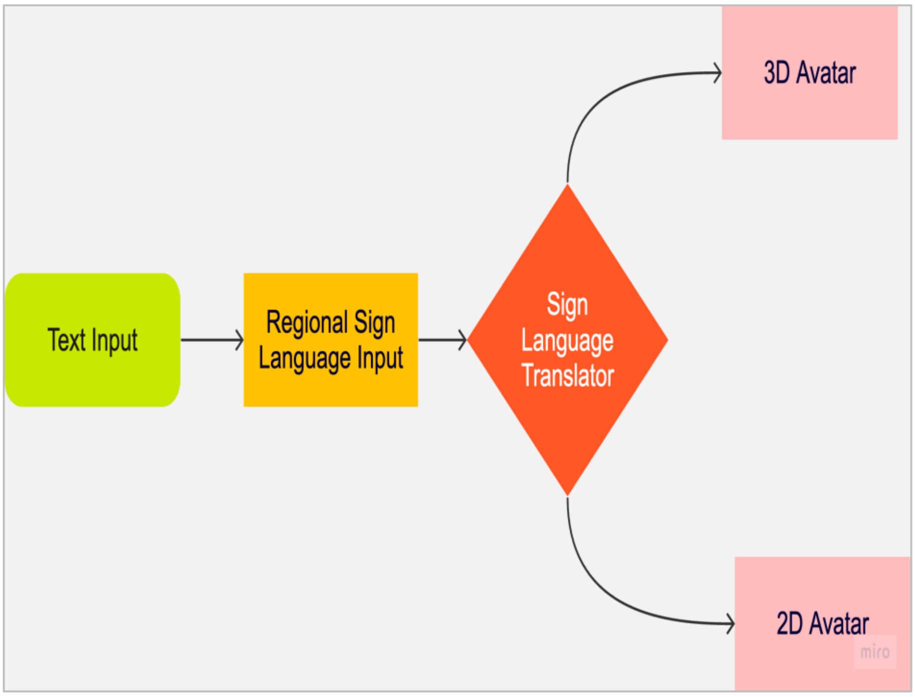

Figure3: Block Diagram in Sign Language Translation

Augment Reality (AR) Glasses Display - Delivering the translated sign language avatars in an AR interface can further enhance comprehension for deaf users. AR overlays computer-generated information onto real-world visuals, transforming the user's field of view into an enhanced display. By overlaying avatars of the speakers mapped to the live speech recognition output, deaf users can simultaneously view lip movements synchronized with the translated sign language animations. Research indicates that combining lip reading and sign language can improve understanding, as facial expressions provide complementary emotional and grammatical cues [3]. Augmented reality provides an ideal medium for seamlessly integrating multiple modes of communication. Overall, an augmented reality interface delivers an inclusive and accessible experience for deaf users that transcends current limits of remote communication.

End to End Methodology - The proposed method converts speech to sign language in four stages. First, audio input from AR glasses is transcribed to text via multilingual speech recognition models. Second, natural language processing corrects errors typical of spoken language transcripts. Third, the cleaned text is translated to regional sign language animations based on detected language. Finally, sign language avatars are rendered in the AR interface for viewing.

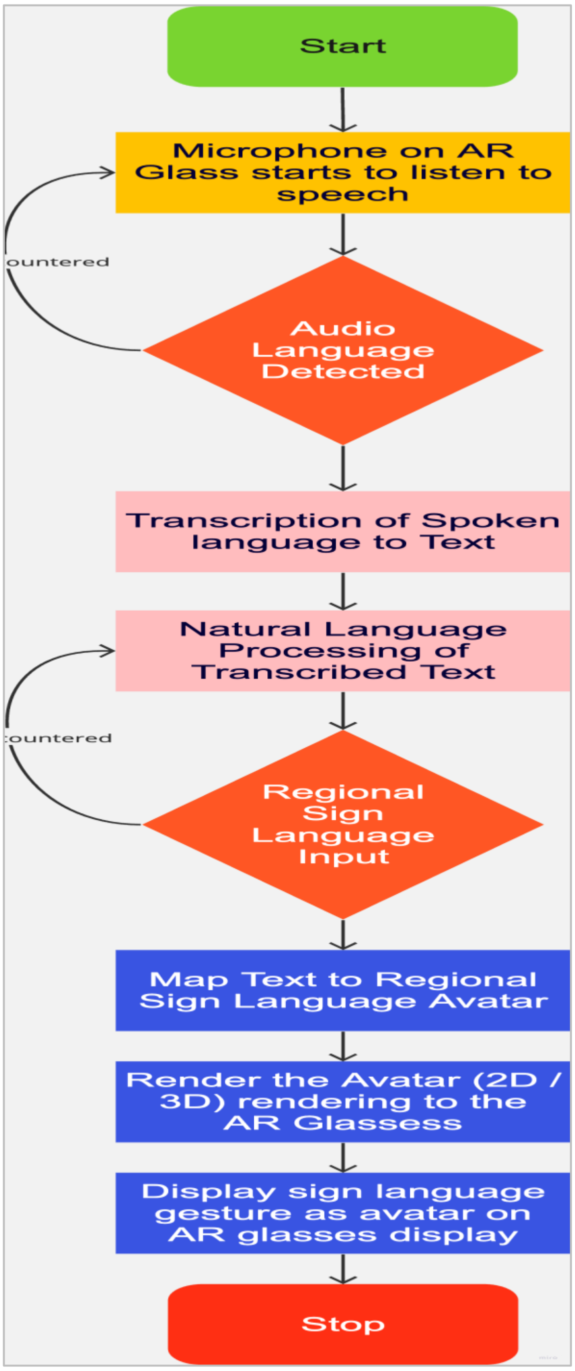
Start
Microphone on AR Glass starts to listen to speech
ountered
Audio Language Detected
Transcription of Spoken language to Text
Natural Language Processing of Transcribed Text
ountered
Regional Sign Language Input
Map Text to Regional Sign Language Avatar
Render the Avatar (2D / 3D) rendering to the AR Glassess
Display sign language gesture as avatar on AR glasses display
Stop

Figure3: Working of the End-to-end System in a Flowchart

4. CONCLUSION

In summary, this work proposes an end-to-end pipeline to provide live sign language translations for improved accessibility. First, automatic speech recognition transcribes spoken audio into text in real-time using recent multitask models like OpenAI Whisper, SeamlessM4T etc. Natural language processing techniques then refine this raw text by cleaning, structuring, and inferring intent. Next, sequence-to-sequence models tailored to each regional sign language translate the text into sign animations. Finally, an augmented reality interface overlays computer-generated avatars performing the translated sign language on top of views of the actual speakers. This allows deaf users to simultaneously integrate information from lip movements with the sign translations. The proposed approach combines state-of-the-art speech recognition, natural language processing, sign language translation, and augmented reality technologies to explore a new paradigm for accessible communication. Overall, this work aims to drive progress towards inclusivity in remote communication for deaf communities worldwide.

5. REFERENCES

[1] Https://Www.Who.Int/Health-Topics/Hearing-Loss#Tab=Tab_2

[2] Https://Education.Nationalgeographic.Org/Resource/Sign-Language/

[3] L. Nguyen, F. Schicktanz, A. Stankowski And E. Avramidis - Evaluating The Translation Of Speech To Virtually-Performed Sign Language On Ar Glasses At Thirteenth International Conference On Quality Of Multimedia Experience, 2021.

[4] Deutscher Geho ̈rlosen-Bund E.V. Geho ̈rlosigkeit. Https://Www.Gehoerlosen- Bund.De /Faq /Geh% C3%B 6rlosigkeit, 2021.

[5] Ja Bangham Et Al. “Virtual Signing: Capture, Animation, Storage And Transmission – An Overview Of The Visicast Project”, Ieee, April 2000 Iee Seminar On Speech And Language Processing For Disabled And Elderly People.

[6] D'armond L. And Speers, M.S. “Representation Of American Sign Language For Machine Translation”, Dissertation Of Ph.D To Georgetown University, Washington Dc, Usa In 2001, Published By Acm Digital Library

[7] Liwei Zhao Et Al. “A Machine Translation System From English To American Sign Language” Published By Springer, Amta 2000: Envisioning Machine Translation In The Information Future Pp 54-67

[8] Kshitij Bantupally And Ying Xie, 2018 “American Sign Language Recognition Using Deep Learning And Computer Vision”.Ieee International Conference On Big Data.

[9] Krunal Sailza, S. Sangeetha, And Viral Shah “Wordnet Based Sign Language Machine Translation: From English Voice To Isl Gloss, 2019 Ieee 16th India Council International Conference (Indicon)

[10] Https://Github.Com/Openai/Whisper

[11] Text Processing Pipeline In A Natural Language Processing Task -Https:// Miro.Medium. Com/V2/ Resize: Fit:1400/1*Cbzccp3xftyvjmwowzlugq.Png

[12] Https://Github.Com/Facebookresearch/Seamless_Communication

[13] Https://Raw.Githubusercontent.Com/Openai/Whisper/Main/Approach.Png