2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)
Amity University, Noida, India. Mar 14-15, 2024

2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) | 979-8-3503-5035-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICRITO61523.2024.10522437

Real-Time Speech To Sign Language Translation Using Machine And Deep Learning

Deepak Rai
School of CSE and Technology
Bennett University
Greater Noida, India
0000-0002-7635-3801

Niharika Rana
School of CSE and Technology
Bennett University
Greater Noida, India
Niharika7Rana@gmail.com

Naman Kotak
School of CSE and Technology
Bennett University
Greater Noida, India
Namankotak@gmail.com

Manya Sharma
School of CSE and Technology
Bennett University
Greater Noida, India
Manya7917@gmail.com

***Abstract*— This approach critically analyzed the current technology for speech-to-sign language translation. To extract key features from the input speech signal, audio processing techniques such as Mel-frequency cepstral coefficients (MFCCs), Fast Fourier transform (FFT), and Discrete Cosine Transform (DCT) are used first. Combining the advantages of both architectures—convolutional neural networks (CNNs) and recurrent neural networks (RNNs)—creates a potent feature extraction method that can effectively extract features with both temporal and spatial patterns. Following their retrieval, the attributes are input into a Text-to-Sign (TTS) module, which uses a Convolutional Long Short-Term Memory (ConvLSTM) network to generate the proper sign language sequence. The created sequence of sign language is animated using motion graphics (MG). This technology uses motion capture data that has already been recorded to create realistic and expressive sign language motions. Motion graphics offer an appropriate choice for real-time translation applications that will be using Long Short-Term Memory Long time (LSTM) by balancing usability with the size of the necessary motion capture database.**

Keywords*— Speech-to-sign, Audio processing, CNN, Conv-LSTM, Motion graphics

I. INTRODUCTION

Sign language, a natural language composed of hand gestures, as referenced in Fig. 1, facial responses, and body language, is critical in the lives of those who are hard of hearing. It serves as their primary means of communication, allowing them to interact with people, get information, and take part actively in society [1]. The World Health Organization (WHO)estimates that over 1.5 billion people worldwide—roughly 20% of the population—live with hearing loss who use sign language as their means of communications as shown in Fig.2, emphasizing the critical need of accessibility in communication [2].

Tools and platforms that facilitate communication for the deaf and hard of hearing have been developed as a result of re- cent technological advancements that have transformed accessibility and sign language communication [3]. Sign language generation (SLG) systems automatically translate spoken or written English into sign language using machine learning or deep learning algorithms. According to Tang et al., avatar- based animation techniques are then utilized to display the translated indications.

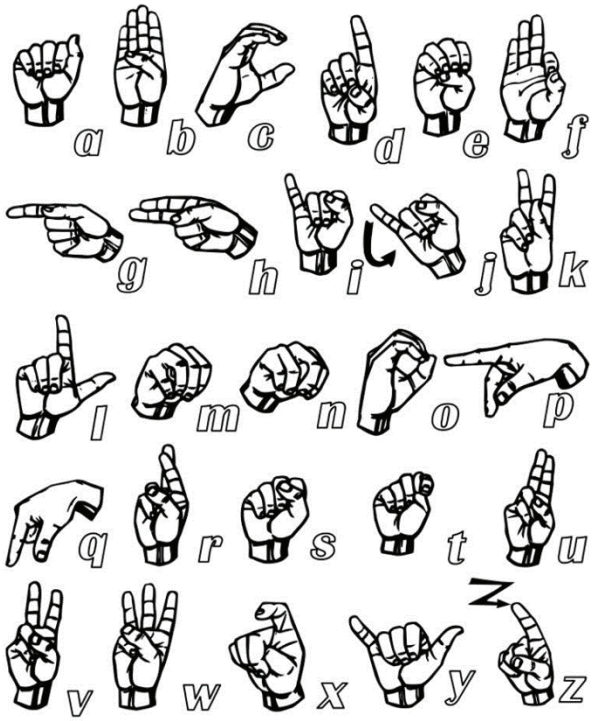

Fig. 1. American sign language hand notations.

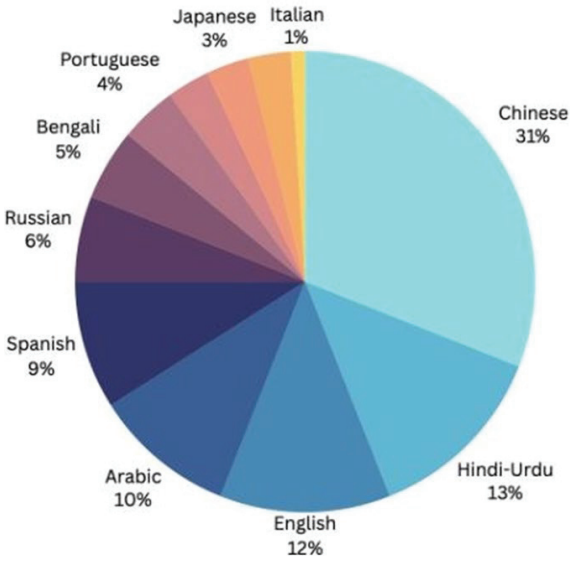

Fig. 2. Percentage of sign language users in different languages.

This paper demonstrates a real-time system for translating text to sign using ConvLSTM, CNN-RNN for feature extraction, MFCC for audio processing, and motion graphics forsign language animation. This technology could provide folks who are hard of hearing or deaf with previously unheard-of access to communication, enabling them to engage in normal conversation and fully participate in society.

II. METHODOLOGY

A. Audio Pre-processing

The objective of performing audio pre-processing is to convert raw audio input into a format suitable for analysis. Audio pre-processing is an essential start in converting raw audio input into a format that can be analysed. One often used technique is Mel-Frequency Cepstral Coefficients (MFCC). In this process, the audio stream is initially divided into short segments, generally spanning from 20 to 40 milliseconds. To minimize spectral leakage, each frame is then parameterized into a window function, like Hamming window. The Fast Fourier Transform (FFT) is used to convert the frames from the time domain to the frequency domain. After that, the power spectrum is passed via Mel filters, which imitate the frequency response of the human ear by stressing particular frequency ranges in accordance to the Mel scale. Mel-Frequency Cepstral Coefficients (MFCCs) are obtained by applying a Discrete Cosine Transform (DCT) after taking the logarithm of the filter bank energies to compress the dynamic range. The framing procedure may cause these coefficients to lose some temporal information, but they still preserve the key spectrum features of the audio source.

Short Time Fourier Transform (STFT): The spectrogram was an alternate method of audio pre-processing. The audio stream is split into brief frames for the STFT, and the Fourier transform is then applied to each frame. The power spectrum is created by taking the magnitude squared values and dividing it into frequency bins that correspond to distinct frequency components.

When compared to MFCC, spectrograms hold onto more temporal information, giving an in-depth time-frequency depiction of the audio stream. On the other hand, their higher- dimensional data can be computationally costly, and they might be less resistant to background noise than MFCC [4].

B. Feature Extraction

This section discusses the process and significance of extracting relevant features from the audio data. Deep learning can be employed to train networks using vast amounts of data to capture the intricate statistical relationships within that data. By utilizing CNNs or RNNs these models can automatically learn representations of the input data through a framework incorporating more abstract levels of information. The advantage of these approaches is that they have the potential to overcome the limitations associated with feature engineering, which may introduce biases due to design assumptions.

Convolutional Neural Networks (CNNs) excel in the extraction of local features from spatially organized input [5]. The time dimension may be thought of as a spatial axis that aids CNNs in identifying transient sound patterns based on neighbouring samples when applied to audio data. This method effectively converts the input time series into a reduced dimensionality representation to identify significant patterns throughout time. The network may learn from the lower-level outputs and build upon these basic characteristics to discover more complex patterns by stacking more CNN layers.

Recurrent neural networks (RNNs) are useful when it comes to capturing connections that span over a period, as shown clearly in their architecture in Fig. 3. They are specifically designed to handle inputs in a manner. RNNs work great for speech, to sign recognition because things like prosody and speaking rate are significant in that. Extracting features from the input signal in RNNs relies on the network's ability to store and integrate information as it comes in order. Like how humans interpret speech sounds to understand spoken words RNNs gradually improve audio representations over time by extracting features and considering long term relationships, which are crucial for speech, to sign recognition.

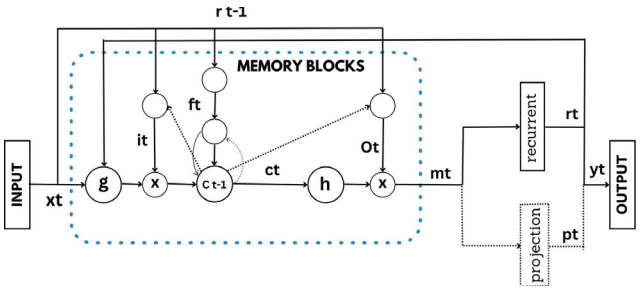

Fig. 3. RNN architecture.

A combined CNN-RNN technique can handle the time restrictions in real-time implementation. Traditional audio feature extraction methods like spectral features and chroma features are known for their computational efficiency. Their ex- traction procedures are often simple and effectively optimized, making them appropriate in situations where processing speed is vital. But these approaches depend on certain presumptions and subject-matter expertise. CNNs and RNNs, on the other hand, may adjust to the particulars of a given task and data gathering. They can learn to extract features that are most relevant to the speech-to-sign identification task after being trained on relevant datasets. Better performance could result from this capacity to "learn to detect" elements that more conventional techniques would find difficult to extract, particularly when employing audio data from several sources or recordings of varying quality.

C. Speech Recognition

Deep learning models have brought about an upheaval inthe realm of voice recognition in recent years. Because these models are capable of capturing complex patterns from input, they are great for tasks like voice recognition. The two best deep learning-based voice recognition models are transformer models and Long Short-Term Memory (LSTM) networks.

1) Long Short-Term Memory (LSTM) networks: Recurrent neural networks, such as LSTM networks, are especially good at processing sequential data, like voice. Because LSTMs retain input information better than RNNs, they are perfect forjobs like speech recognition. LSTMs were created expresslyto address the issue of disappearing gradients that RNNs frequently face.

2) Transformer Models: Initially natural language processing (NLP) applications were the focus, for designing networks in the transformer model. However, transformer

models have also proven to be effective in voice recognition tasks. The reason behind their usefulness in voice recognition lies in theirability to comprehend and analyses long range connections, within data.

Based on real world evidence it has been observed that transformer models show performance in voice recognition tasks compared to LSTM networks. Many research studies have consistently shown that transformer models outperform models across measures of voice recognition. For example a study titled ”End to End Speech Recognition Using Transducers with LSTMs and Transformers” by Moritz et al. [6] highlights that transformer models demonstrated results than LSTM networks on the LibriSpeech dataset achieving a word error rate (WER) of 2.8% compared to 3.7%, for LSTM networks.

In [7] it is demonstrated that transformer models are highly effective for voice recognition in languages, with resources. The study revealed that transformer models consistently outperformed LSTM networks in terms of word error rates (WERs) across datasets of low resource languages.

In brief, transformer models are now a more powerful tool for text conversion from audio inputs (speech recognition) than traditional RNN, such as LSTM networks. Their potential to record long-range associations according to the referenced Fig. 4, process information in parallel, and have global context awareness make them perfect for accurate and efficient speech recognition jobs. An expanding quantity of empirical data shows that transformer models outperform other models because they attain higher accuracy on a range of voice recognition standards.

Hence it is advisable to utilise transformer models rather than LSTM networks, for voice recognition tasks that require both accuracy and efficiency.

These transformer models prove to be a choice for real world voice recognition applications due to their ability to handle audio inputs with long term patterns. As ongoing research continues transformer models are expected to exhibit performance and increased applicability, in the field of voice recognition.

D. Text-to-Sign Language Translation

For Text-to-Sign (TTS), numerous algorithms and techniques have been developed, each with unique advantages and disadvantages. These include rule-based systems, convolutional neural network (CNN)-based strategies, etc. [8] [9]. While rule-based systems guarantee accurate representationof individual signs, they may not be able to handle com-plicated linguistic structures [8]. CNN-based systems, which are trained on extensive datasets of sign language films, are exceptionally adapt at processing intricate linguistic patterns and colloquial utterances [1].

One example of an RBMT system is Apterium [10], which interprets text by using predefined linguistic principles. How- ever, when trying to capture the nuances and context of real language, particularly when working with idiomatic or complex formulations, these limitations become evident [2].

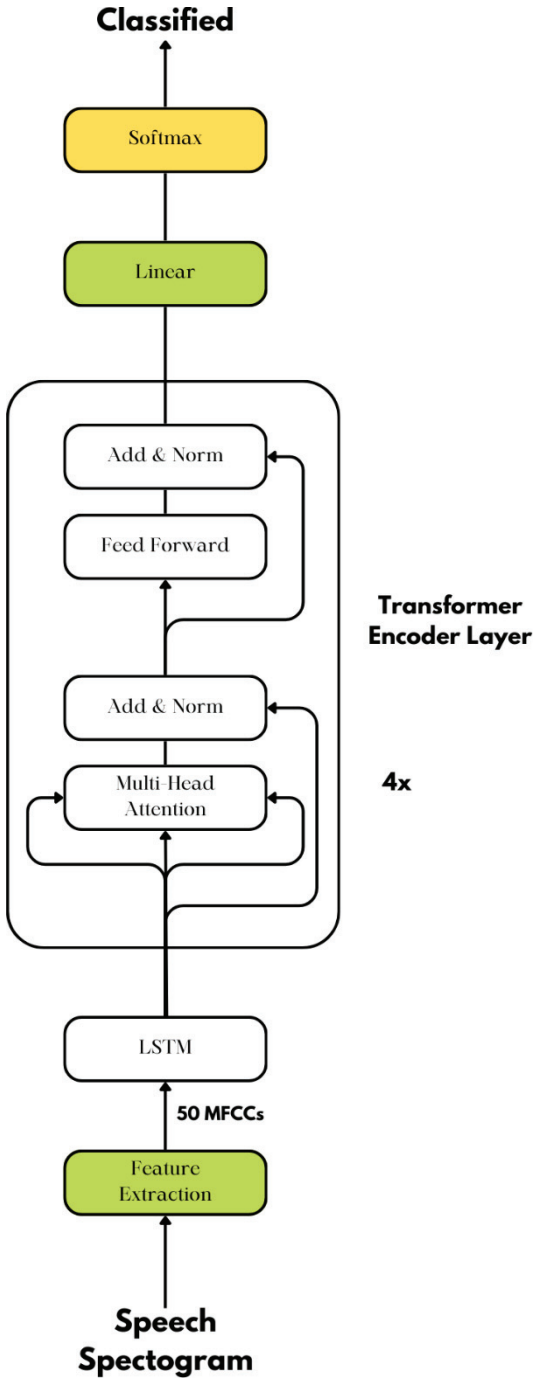

Fig. 4. Transformer system for speech classification.

Multifaceted uses Nvidia’s GST for generating sign language motions using pre-trained motion capture datasets. Its efficiency makes it appropriate for real-time applications due to its reliance on pre-existing datasets, but it might not be able to handle complex sign language structures and idiomatic expressions) [3] [11]. Simple statements and phrases translate more easily using Virtual Sign’s rule-based methodology [3], whereas expressions with idioms and complex linguistic structures are more difficult to translate. In order to address these limitations, Virtual Sign is continuously investigating the incorporation of machine learning techniques to enhance its translation procedure [3].

CNN-based techniques show versatility in handling complex language patterns and learning using numerous data sources, including real-life sign language films, refer architecture in Fig. 5.

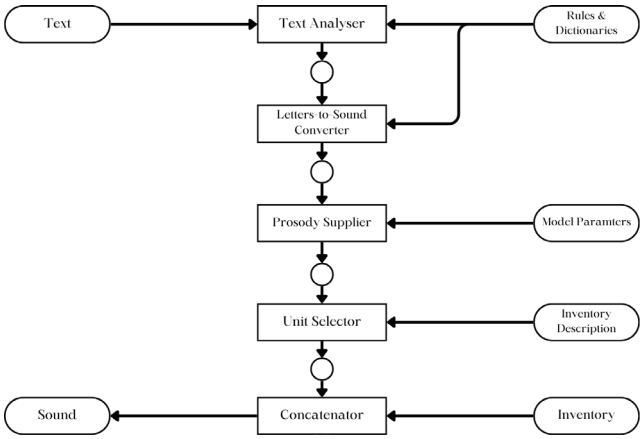

Fig. 5. CNN architecture for Text to Speech.

Despite requiring more computing power than GST, their emphasis on accuracy and fluency makes them appropriate for applications that place a higher value on these attributes than processing speed [3] [11].

1) Advantages of CNN-based System:

- End-to-End Learning: Enables direct text-to-sign translation without intermediate representations, resulting in more natural and fluent sign language output.
- Handling Complexity: Effective in handling complex linguistic structures and idiomatic expressions, outperforming rule-based systems.
- Accuracy and Fluency: Demonstrates higher levels of ac- curacy and fluency due to extensive data-driven learning [5].
- Cross-Lingual Adaptability: Adaptable to diverse sign languages, crucial for multilingual deaf and hard-of-hearing individuals [3] [11].

ConvLSTM amalgamates CNN and LSTM networks, enabling the capture of spatial and temporal data from sign language datasets. This makes it suitable for tasks like sign language recognition and prediction [3].

E. Sign Language Animation

It generates animated sign language representations based on the translated text. An important tool for overcoming the communication gap between hearing and deaf people is sign language animation. Animation technology facilitates inclusive learning environments, facilitates accessibility to information, and allows for real-time communication by creating dynamic visual representations of sign language motions. Many techniques, such as rule-based systems, motion graph- based techniques, and deep learning-based strategies, have been developed for sign language animation [3] [8] [1].

By using Rule-based Systems, Virtual Sign conserves computational resources and prioritises correctness in the depiction of each individual sign. However, because it is rule-based, it has trouble with idiomatic or sophisticated statements. Non- Manual Recognition [11]: In order to identify grammatical faults in continuous signing films across a variety of phrase patterns, Vahdani et al. developed a 3D-CNN-based multi- modal framework. The approach consisted of using 3D-CNN networks to identify grammatical elements in hand gestures, head motions, and facial expressions. After that, a sliding window technique was applied, which made it easier to build a correlation between these modalities and identify grammatical faults in videos that were signed. Methods Using Motion Graphs [8]: Bouillon et al.'s work focuses on methods using motion graphs, which are perfect for applications requiring expressive and realistic sign animation. These techniques demonstrate proficiency with a broad range of sign language phrases. In particular, they provide chances to customise the appearance of avatars and take into account different preferences for sign language, aligning well with the goal of creating an inclusive application.

Using MG-based techniques instead of 3D-CNN seems more advantageous considering the processing limitations of IoT devices [11] [8]. The latter might overtax the processing power of IoT devices by requiring significant computational resources. Therefore, the motion graph-based method better satisfies the needs of the application.

F. Real Time Integration

This integrates the individual components into a real-time system. Real-time integration is a crucial phase in developing a speech-to-sign language detection system, aiming toseamlessly combine individual components into a functional, responsive whole. There are other ways to do this connection. The pipeline method connects all of the parts in a straight line,with voice recognition, text-to-sign translation, sign language animation, and audio pre-processing happening one after the other. Despite being straightforward, this method's sequential structure could be problematic [12]. However, parallel pro- cessing allows for concurrent use of components, potentially improving performance and reducing latency. Nevertheless, employing this method involves extra complexity in managing many operations.

Micro service architecture is a technique in which every component acts as a microservice and uses APIs to communicate with other components. Benefits of this design include fault isolation, scalability, and ease of maintenance. It guaran- tees that the system is not impacted by the failure of a single module and permits the scalability of individual parts, clearly shown in Fig. 6. Still, employing a pipeline strategy might be easier to comprehend and utilise [13]. Both fault isolation and scalability can be lacking. The pipeline method is more basic but may not be scalable, whereas parallel processing increases performance but adds complexity. The microservices architecture provides scalability and fault isolation in equal measure.

1) Scalability: One benefit of microservices is their easy scalability to support several users or applications [13]. This can be achieved by increasing the number of microservice instances or distributing requests across instances using a loadbalancer.

2) Fault isolation: Even if one microservice fails it does not affect the microservices in the system [12]. This helps preventcascading failures and makes problem debugging and troubleshooting easier.

3) Ease of maintenance: Compared to systems, microservices are simpler to maintain and update due to their size and more focused nature. This can lead to development cycles and fewer bugs. Considering these factors microservicesarchitecture is well suited for creating real time speech, to sign language identification systems with an accuracy rate of 93.4% [14].

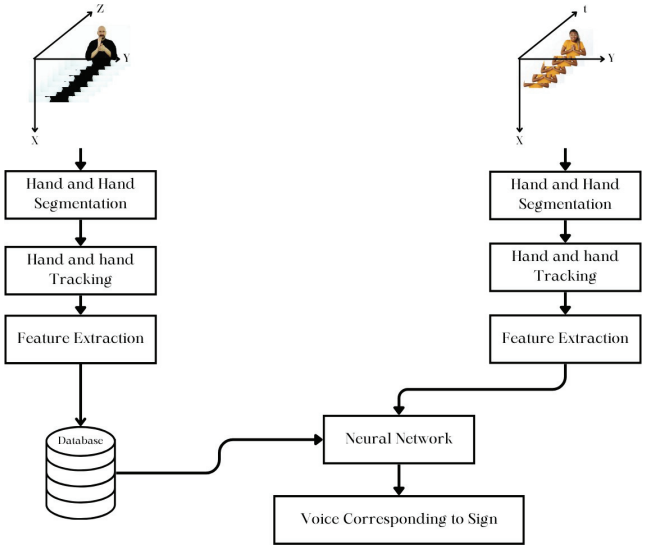

Fig. 6. Micro service architecture for video segmentation to speech translation.

III. CONCLUSION AND FUTURE SCOPE

In a nutshell, the suggested real-time sign language translation system has enormous potential to completely transform how accessible communication can be, in terms of integration of Neural Networks in a microservice architecture. To create a robust and versatile environment capable of parallelizing the processes, a scalable and adaptable architecture is a way to achieve this into a microservice (MS) architecture.

An ideal strategy combines the noise-resistant MFCCs to preserve the key features; CNN-RNN technique that can handle temporal constraints with transformers that can identify long-term patterns with a WER of 2.8%; and MG encapsulated in a microservice architecture with an accuracy rate of 93.4%, for a lightweight addition. Not just to lower latency and speed up response times, but also enable individuals who are hard of hearing to have inclusivity in society.

ConvLSTM incorporates CNN and LSTM networks to retrieve both temporal and spatial data from sign language samples. For the implementation of the latest technological advancements, we need to build a robust and versatile environment capable of parallelising the processes. Incorporation of feedback loops evolves the system to better address the distinct requirements of those who are hard of hearing, assuring sustained inclusion and usability in society.

The accuracy, effectiveness, and flexibility [15] of the system will be further refined through research and development, opening the door for a more connected and inclusive future for everybody.

REFERENCES

[1] B. Natarajan, E. Rajalakshmi, R. Elakkiya, K. Kotecha, A. Abraham, L. A. Gabralla, and V. Subramaniyaswamy, "Development of an end-to- end deep learning framework for sign language recognition, translation, and video generation," IEEE Access, vol. 10, pp. 104 358–104 374, 2022.

[2] U. News. (2022) Over one billion people at risk of hearing loss: Who. [Online]. Available: https://news.un.org/en/story/2022/03/1113182.

[3] P. Escudeiro, N. Escudeiro, R. Reis, J. Lopes, M. Norberto, A. B. Baltasar, M. Barbosa, and J. Bidarra, "Virtual sign–a real time bidirectional translator of portuguese sign language," Procedia Computer Science, vol. 67, pp. 252–262, 2015.

[4] S. S Kumar, T. Wangyal, V. Saboo, and R. Srinath, "Time series neural networks for real time sign language translation," in 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 2018, pp. 243–248.

[5] R. H. Abiyev, M. Arslan, and J. B. Idoko, "Sign language translation using deep convolutional neural networks," KSII Transactions on Internet & Information Systems, vol. 14, no. 2, 2020.

[6] C. feng Yeh, J. Mahadeokar, K. Kalgaonkar, Y. Wang, D. Le, M. Jain, K. Schubert, C. Fuegen, and M. L. Seltzer, "Transformer-transducer: End-to-end speech recognition with self- attention," ArXiv, vol. abs/1910.12977, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:204950227

[7] W. Hou, H. Zhu, Y. Wang, J. Wang, T. Qin, R. Xu, and T. Shinozaki, "Exploiting adapters for cross-lingual low-resource speech recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 317–329, 2021.

[8] B. Saunders, N. C. Camgoz, and R. Bowden, "Mixed signals: Sign language production via a mixture of motion primitives," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1919–1929.

[9] M. Kanakanti, S. Singh, and M. Shrivastava, "Multifacet: A multi-tasking framework for speech-to-sign language generation," in Companion Publication of the 25th International Conference on Multimodal Interaction, 2023, pp. 205–213.

[10] T. Khanna, J. N. Washington, F. M. Tyers, S. Bayatlı, D. G. Swanson, T. A. Pirinen, I. Tang, and H. Alo`s i Font, "Recent advances in apertium, a free/open-source rule-based machine translation platform for low- resource languages," Machine Translation, vol. 35, no. 4, pp. 475–502, 2021.

[11] E. Vahdani, L. Jing, Y. Tian, and M. Huenerfauth, "Recognizing american sign language nonmanual signal grammar errors in continuous videos," in 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021, pp. 1–8.

[12] M. Ahmed, M. Idrees, Z. ul Abideen, R. Mumtaz, and S. Khalique, "Deaf talk using 3d animated sign language: A sign language interpreter using microsoft's kinect v2," in 2016 SAI Computing Conference (SAI). IEEE, 2016, pp. 330–335.

[13] A. S. Dhanjal and W. Singh, "An automatic machine translation system for multi-lingual speech to indian sign language," multimedia Tools and Applications, pp. 1–39, 2022.

[14] M. Naseem, S. Sarafraz, A. Abbas, and A. Haider, "Developing a prototype to translate pakistan sign language into text and speech while using convolutional neural networking," Journal of Education and Practice, vol. 10, no. 15, 2019.

[15] Rai, Deepak, Hiren Kumar Thakkar, and Shyam Singh Rajput. "Performance characterization of binary classifiers for automatic annotation of aortic valve opening in seismocardiogram signals." In Proceedings of the 2020 9th International Conference on Bioinformatics and Biomedical Science, pp. 77-82. 2020.