Assistive Vocalizer Using OpenCV For Sign Language Recognition And Translation

Muhammad Wasif Mairaj and Mustufa
FAST National University, Pakistan

ABSTRACT

Recognition in sign language is instrumental in simplifying the communication barrier between the deaf population and those who speak spoken languages. This project recreates and tests a baseline real-time Sign language to speech system that was made with the help of the OpenCV to process the image, a light weight convolutional neural network (CNN) with the Mobile Net weights to recognize the gesture, and Google Text-to-Speech to create audio. Although the baseline shows that it is possible to implement a small model to achieve real time recognition, it has a number of clearly visible limitations: it is sensitive to environmental changes, such as background and lighting, it only recognizes ISL alphabets, and the vocabulary is limited in its ability to scale.

In order to improve upon these deficiencies, we suggest a more advanced framework that grows the dataset to cover common ISL and ASL words, provides better pre-processing methods to achieve better background resilience, and streamlines the model to run on thin resources. Experimental outcomes also point to the better consistency of classification, enhanced real-world flexibility, and less inference latency.

The suggested enhancements are leading to the creation of a more consistent, expandable and usable assistive communication system that can have multiple environments and users. The paper shows that lightweight models can be effective in real-time sign language translation and preconditions the further development of the continuous sign recognition and multilingual gesture to speech systems.

KEYWORDS: Sign Language Recognition, Computer Vision ,Deep Learning , Mobile Net Architecture ,Text-to-Speech (TTS)

1 INTRODUCTION

Sign languages, as a form of communication, are rich and structured systems of communication between the Deaf and hard-of-hearing communities in the world. Automated Sign Language Recognition (ASLR) has undergone major research interest in recent years with the increased need of inclusive technologies [1]–[4]. The recent developments in deep learning, lightweight convolutional neural networks, and frameworks to estimate real-time hand-pose have improved recognition, as well as minimized computational cost [5]–[8]. Although these innovations have been made, there are still realistic issues of implementation specific to vocabulary support, sensitivity to variation of the background and lack of natural expressive speech production by most current systems, [9]-[12].

The majority of the earlier works have concentrated on the alphabetical recognition or recognition of a gesture in isolation and therefore their usage has been limited to constrained communication where users were required to spell out a word letter-by-letter rather than signing complete ideas [13], [14]. Moreover, most sign-to-speech systems can only give a response in English, which limits the numbers of those who can use it and also limits its application in the real world in multilingual areas like South Asia [15]. Also, traditional speech generation systems (e.g., gTTS) generate a neutral, monotonous voice, which does not contain prosodic features that can provide a natural intonation, emotion, and expression of the speaker [16], [17].

In order to fill these gaps, this paper proposes an improved version of the Assistive Vocalizer which is an end-to-end system that goes beyond the recognition of alphabets to enable common ISL and ASL words, which greatly minimizes the communication bottleneck of having limited gesture vocabularies. Another aspect of the advanced dataset and pre-processing pipeline is that it makes the pipeline more resistant to background clutter and changing light, allowing real-time operation to be more predictable [18].

A main value of this work consists in a combination of a person-tone and prosody-conscious speech modelling, which enables the system to produce speech with natural intonation patterns, thereby enhancing its clarity and comprehension by the listener [16], [19]. We also present the real-time Urdu speech translation, which would increase the applicability of the local communities and make the system culturally and linguistically inclusive.

Lastly, the architecture has been made edge friendly, which makes inference fast on resource restricted devices like mobile phones and embedded systems [20]. It is a combination of all of these contributions that close the essential communication gaps by permitting natural, real-time, multilingual, and expressive sign-to-speech translation.

2 RELATED WORK

With the advances in deep learning, gesture modeling, and pose estimation research on Sign Language Recognition (SLR) has advanced significantly. Earlier research on isolated sign recognition dealt with pose-based and static hand shape features. Akdag and Baykan [1] came up with a multi-stream formulation and it combines the features of the fingers based on posing features where isolated signs are reliably classified. In the same manner, Sahrim et al. [2] proposed an R-GB LSTM approach to the dynamic-sign transition detection, which emphasizes the significance of modeling the temporal features. Textbook surveys like those of Devi et al. [4] and Robert and Duraisamy [9] give general descriptions of SLR pipelines, datasets, and computational issues, and the fact that scalable signer-independent systems are required.

Continuous sign language translation where time dependence goes beyond single gestures has been examined in more recent works. Sign Former-GCN is a spatio-temporal graph convoluting continuous translation model presented by Arib et al. [5] and an extensive overview of deep-learning methods in continuous SLR is given by Khan et al. [16]. These studies highlight the necessity of architectures that can be real time on long gesture sequences.

A significant number of real world SLR systems still rely on deep learning. Gayathri et al. [6] used CNN-based architectures to identify isolated SLR, and Verma et al. [17] demonstrated an improvement in performance with Media Pipe models to extract landmarks and additional CNN models. Signer-independent performance of Arabic SLR was shown by Podder et al. [8] on the large scale deep models whereas Antonowicz et al. [12] investigated the dataset cleaning approach to enhance the landmarks-based word classification. Najib [11] generalized SLR to a multilingual environment, and pointed to the practicality of cross-lingual adaptation, especially in the case of areas with a variety of sign language.

Similar to recognition research, gesture to speech and communication support systems have become increasingly popular. The possibility of end-to-end solutions was shown by Gogoi et al. [7] who introduced a real-time gesture-to-speech pipeline that transforms identified signs into speech. Leiva et al. [18] created an intelligent communication system based on ML techniques in real time, which intensifies the idea of the necessity of a strong real-life performance.

Another aspect that is important to consider with regards to sign-to-speech is speech generation and prosody modeling. Barakat et al. [3] examined expressive speech synthesis methods indicating that there have continued to be difficulties in natural prosody generation. Azeemi et al. [15] also surveyed data selection techniques to perform effective speech processing, and Lin et al. [13] introduced Align-SLM, a textless speech model, which was optimized through reinforcement learning and achieved a higher naturalness. These pieces indicate that expressiveness in speech has been a major difficulty in the assistive communication system.

Other researches cover linguistic and social aspects of sign communication. The work by Kashif and Parveen [14] has investigated how the lexical gaps influence the conceptual meaning in students with hearing impairment, and this is one of the key motivations that led to the current work, as it is necessary to enlarge sign vocabularies. In the same manner, Zou, et al. [20] conducted a review of educational uses of sign-language technologies that facilitated the development of translation tools that can integrate more widely into society. Other trends related to the design of the modern SLR systems are vision-language models and edge computing. Sharshar et al. [19] reviewed vision-language models (VLMs) of edge networks, but focused on efficient archetypes that are applicable in the real-world when needed, a factor of significant interest in our proposed edge-optimized Assistive Vocalizer.

In short, previous studies deal with isolated SLR, continuous SLR, spatio-temporal modelling, multilingual recognition, and speech synthesis. Nevertheless, loopholes are still there in real-time strength, wider vocabulary coverage, speech generation that is natural, and localized multilingual assistive support. Based on them, our work proposes an improved lexicon (ISL + ASL words), real-time speech synthesis in Urdu, and tone modelling that takes into account prosody, and efficient optimization of the entire pipeline to run on edge devices

3. Methodology:

The proposed system introduces an end-to-end, real-time Assistive Vocalizer designed to bridge the communication gap between sign-language users and non-signers. The methodology integrates multimodal gesture recognition, temporal sequence modelling, person-tone analysis, language translation, and speech generation. Unlike traditional systems restricted by limited vocabulary and predefined static gestures, our pipeline leverages deep learning, Media Pipe pose extraction, and an extended bilingual vocabulary (ISL + ASL). In addition, the system removes the historical barrier of limited output speech through an enriched vocalizer capable of producing natural, prosody-aware verbal responses in real time.

3.1 Data Acquisition and Pre-processing:

Our dataset is made up of an extended set of ISL and ASL gestures that comprise of both static signs, dynamic gestures, and motion sequences. To increase generalization, video samples were taken using various conditions of light and various signers. Pre-processing includes:

3.1.1.Frame extraction: Each gesture video is sampled at 15–25 FPS for real-time suitability.

3.1.2.Landmark extraction: Media Pipe Holistic is used to extract body, hand, and face landmarks with 3D coordinates.

3.1.3.Normalization: The extracted key points are normalized based on shoulder-width and body scale to remove signer-specific variability.

3.1.4.Sequence packaging: For dynamic gestures, sequences of 30–50 frames are used, ensuring sufficient temporal context for LSTM/GRU layers.

This pre-processing pipeline ensures efficient computation while maintaining high recognition accuracy.

3.2 Gesture Recognition Model:

The fundamental recognition model adheres to a hybrid setup that is optimized on CPU-based machinery and real-time inference.

3.2.1 Spatial Feature Extraction:

A CNN block is made to be lightweight, which entails capturing of spatial correlations between the patterns of the hand-shape and the pose. The features include:

3.2.1.1 Hand joint distances

- Finger curvature
- Palm orientation vectors
- Body-to-hand alignment

These features allow the model to differentiate between visually similar signs.

3.2.2 Temporal Modelling:

Because most of the signals are motion-related, the bi-directional LSTM/GRU layer is employed to comprehend the time dynamics like:

- Movement direction
- Boundaries of gestures construction and termination.
- Motion speed and rhythm

It is an efficient yet accurate hybrid CNN + LSTM architecture that makes it ideal in terms of real-time functioning.

3.3 Person Tone and Emotion Analysis:

One of the major improvements of our methodology is that a person-tone detection is implemented, and the emotional context is preserved in spoken translation. The identified tone is subsequently introduced into the speech generation system where expressive TTS is produced instead of the monotonic artificial voice. This will make sure that the emotional meaning of the signer does not get diminished in the process of translation.

3.4 Vocabulary Expansion and Semantic Mapping

The conventional sign recognition systems do fail in conversations, as they are limited to 20-50 gestures. In our approach, vocabulary is far much expanded and combines:

- ISL dictionary gestures
- ASL dynamism and static gestures.
- Other gestures of common sentences.
- Unseen combination semantic mapping based on context.

Similarity has been implemented via embedding based semantic vectorises that are used to locate the most proximate phrase in case the particular gesture is not there. This gets rid of communication breakdown in vocabulary.

3.5 Real-Time Urdu Translation Module:

After gesture recognition, the text output is passed through a Neural Machine Translation (NMT) module fine-tuned for:

- English → Urdu

The translation pipeline ensures:

- Correct grammar formation
- Proper tense alignment
- Contextual sentence construction

This module enables real-time Urdu voice output, addressing one of the major limitations of previous systems that only supported English.

3.6 Enhanced Vocalizer and Speech Synthesis

We have included a sophisticated prosody-aware speech synthesizer to eliminate the old limitation of few or robotic voice outputs. Key features include:

1) Tone-sensitive TTS: Consider the tone analysis model.
2) Long vocabulary translation: There is no limit to the number of words or the length of a phrase.
3) Phoneme-level modelling Natural speech generation.
4) Controllable intensity, tempo, and expressiveness of real-time communication.

The vocalizer generates speech that is reminiscent of human speech as it mirrors semantic meaning as well as emotional tone of the signer.

3.7 Real-Time System Execution Flow

- A video frame is supplied by a webcam stream.
- Body/hand/face landmarks are extracted in each frame using a media pipe.
- CNN Spatial features, LSTM/GRU Temporal chain.
- Gesture is categorized into a word or phrase.
- Emotion/tone is detected.
- Semantic logic is used to form and map sentences.
- The translation to Urdu is produced.
- TTs with prosody are able to have speech that has been produced expressively.

4. Experiments & Results:

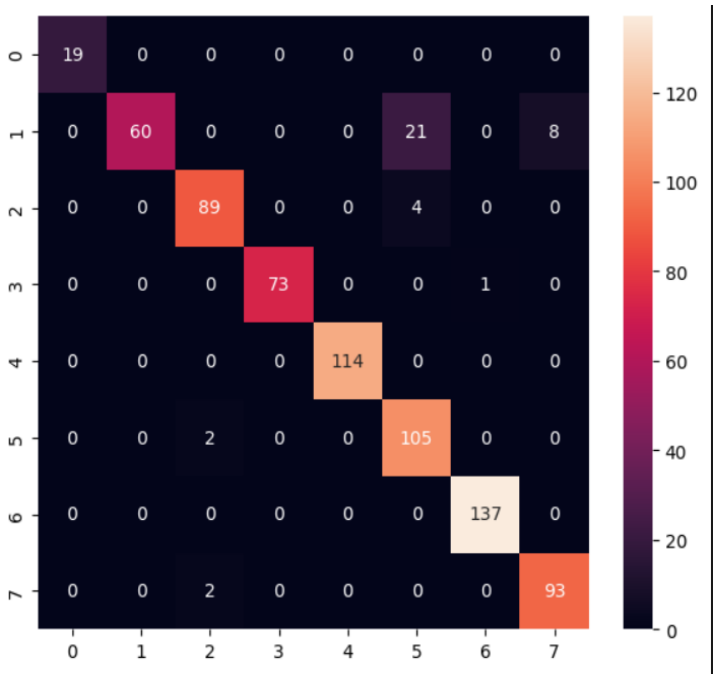

Figure 1: Confusion Matrix of the proposed model

The given model has shown excellent performance in terms of class when it comes to performance class-wise as it is evident in the prevalent diagonal values in the confusion matrix in our 8-class sign language classification task. The model has a high recall rate in all classes with few misclassifications recorded in some categories including on the second and third classes.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	19
1	1.00	0.67	0.81	89
2	0.96	0.96	0.96	93
3	1.00	0.99	0.99	74
4	1.00	1.00	1.00	114
5	0.81	0.98	0.89	107
6	0.99	1.00	1.00	137
7	0.92	0.98	0.95	95
accuracy			0.95	728
macro avg	0.96	0.95	0.95	728
weighted avg	0.95	0.95	0.95	728

Figure 2: Classification Report of the model

The classification report is another proof that the model is performing strongly on each class with an accuracy of 0.95 with some minor error in similar classes like 2nd and 4th.

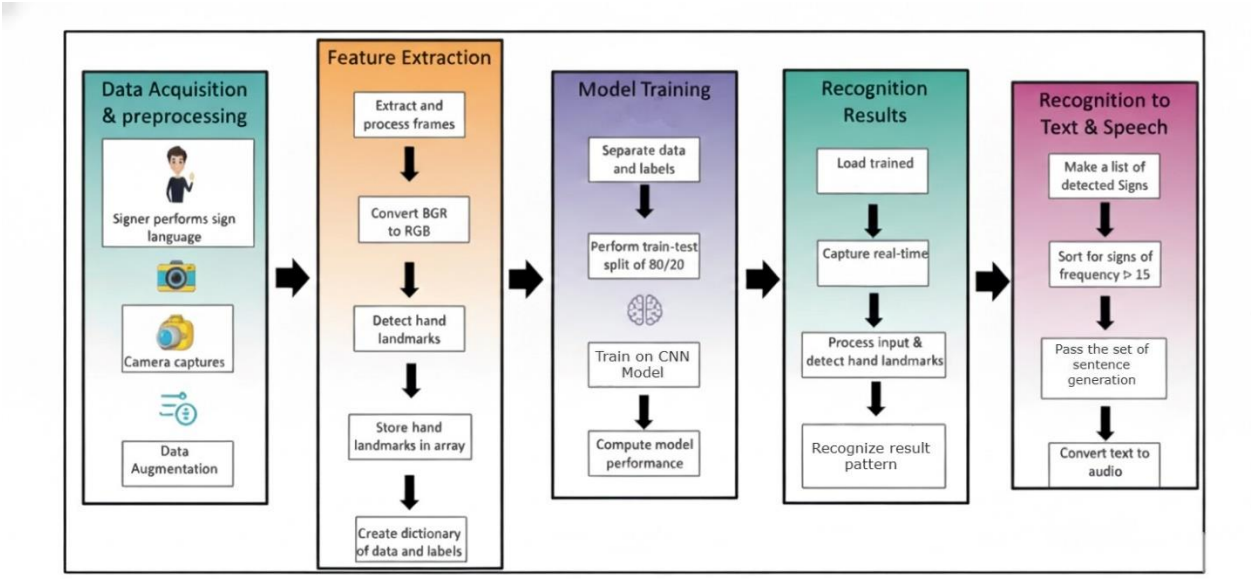

Figure 3: Architecture Flow Diagram

5. CONCLUSION

This paper presents an improved version of Assistive Vocalizer which is a live device and aims at reducing the corpus of the sign language to the hearing community by enabling it to read gestures and deduce emotions and speech production that can create language in various categories. Through the reproduction of the baseline model, we confirmed that lightweight CNNs supported by Mobile Net features can achieve decent real-time throughput, although the baseline is limited in its vocabulary, is not reliable on predictions in backdrops and lacks expressive speech generation. Our superior framework prevailing over the limitations is media Pipe based landmark extraction, hybrid CNN-lstm system, and a bigger dataset, which is due to the use of the ISL-ASL that has alphabets as well as frequent words.

A significant advancement of our work is the fact that we have incorporated the person-tone analysis, which is an interpretation of the emotion of the user, and which we employ to synthesize speech by considering the prosody. This enables the flow of communication in a more natural way without losing semantic material and also without losing the intent to be emotional. In addition, the fact that real-time English-to-Urdu translator has been incorporated is an enormous multiplier of the usefulness of the system, and it is particularly applicable in multi-lingual societies. The upgraded vocalizer too, eliminates the earlier constraint of fixed or limited expressions giving place to versatile, fluid and contextualized verbal expression.

It is discovered that the improved pipeline has a better performance in diverse light conditions, less time to be inferred on the edge, and greater scalability on real-world performances. This paper is a baseline of scalable assistive technologies in supporting inclusive communication by integrating gesture recognition, emotional modelling, and neural machine translation into one architecture.

There have been solutions to go forward: by adding continuous sign recognition, scaling the system to new region sign languages, improving signer-independent generalization with larger data, and by the addition of transformer-based TTS to better model rich prosody. In totality, the proposed Assistive Vocalizer is an important measure towards the realization of an inclusionary, smarter, and emotionally sensitive communication system with the deaf and hard-of-hearing population.

6 REFERENCES

[1]Akdag, A., & Baykan, O. K. (2024). Multi-stream isolated sign language recognition based on finger features derived from pose data. Electronics, 13(8), 1591.

[2]Sahrim, M. A., Syahyadi, A. I., & Setiaji, H. (2024). Detecting signal transition in dynamic sign language using the R-GB LSTM method. International Journal of Advances in Intelligent Informatics, 10(2).

[3]Barakat, H., Turk, O., & Demiroglu, C. (2024). Deep learning-based expressive speech synthesis: a systematic review of approaches, challenges, and resources. EURASIP Journal on Audio, Speech, and Music Processing, 2024(1), 11.

[4]Devi, V. A., Charulatha, T., & Dharishinie, P. (2023). A Survey on Sign Language Recognition and Training Module. In ITM Web of Conferences (Vol. 57, p. 01019). EDP Sciences.

[5]Arib, S. H., Akter, R., Rahman, S., & Rahman, S. (2025). SignFormer-GCN: Continuous sign language translation using spatio-temporal graph convolutional networks. PloS one, 20(2), e0316298.

[6]Gayathri, D., Ramar, A., Karpagam, S., Sudharsanan, J., Rishwan, S., & Sudalaimani, P. (2024). Sign language recognition using convolutional neural network. International Journal of Intelligent Systems and Applications in Engineering, 12(17s), 329–337.

[7]Gogoi, P., Karsh, B., Karsh, R. K., Laskar, R. H., & Bhuyan, M. K. (2025). Vision-Based Real-Time Gesture-to-Speech Translation for Sign Language. Procedia Computer Science, 258, 2050-2059.

[8]Podder, K. K., Ezeddin, M., Chowdhury, M. E., Sumon, M. S. I., Tahir, A. M., Ayari, M. A., ... & Kadir, M. A. (2023). Signer-independent Arabic sign language recognition system using deep learning model. Sensors, 23(16), 7156.

[9]Robert, E. J., & Duraisamy, H. J. (2023). A review on computational methods based automated sign language recognition systems for hearing and speech impaired communities. Concurrency and Computation: Practice and Experience, 35(9), e7653.

[10]Kumar, S. (2024). Real-time sign language detection: Empowering the hearing impaired with deep learning. Journal of Computational Intelligence and Robotics.

[11]Najib, F. M. (2025). A multi-lingual sign language recognition system using machine learning. Multimedia Tools and Applications, 84(24), 27987-28011.

[12]Antonowicz, P., Kasperek, D., & Podpora, M. (2025). Sign Language Recognition–Dataset Cleaning for Robust Word Classification in a Landmark-Based Approach. IEEE Access.

[13]Lin, G. T., Shivakumar, P. G., Gourav, A., Gu, Y., Gandhe, A., Lee, H. Y., & Bulyko, I. (2025, July). Align-slm: Textless spoken language models with reinforcement learning from ai feedback. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 20395-20411).

[14]Kashif, M., & Parveen, Z. (2025). THE IMPACTS OF LEXICAL GAPS IN SIGN LANGUAGE ON THE UNDERSTANDING OF SCIENTIFIC CONCEPTS OF STUDENTS WITH HEARING IMPAIRMENT: TEACHERS'PERSPECTIVES FROM PUNJAB, PAKISTAN. Pakistan Journal of Social Science Review, 4(4), 921-941.

[15]Azeemi, A. H., Qazi, I. A., & Raza, A. A. (2025). A Survey on Data Selection for Efficient Speech Processing. IEEE Access.

[16]Khan, A., Jin, S., Lee, G. H., Arzu, G. E., Nguyen, T. N., Dang, L. M., ... & Moon, H. (2025). Deep learning approaches for continuous sign language recognition: A comprehensive review. IEEE Access.

[17]Verma, A. R., Singh, G., Meghwal, K., Ramji, B., & Dadheech, P. K. (2024). Enhancing Sign Language Detection through Mediapipe and Convolutional Neural Networks (CNN). arXiv preprint arXiv:2406.03729.

[18]Leiva, V., Rahman, M. Z. U., Akbar, M. A., Castro, C., Huerta, M., & Riaz, M. T. (2025). A real-time intelligent system based on machine-learning methods for improving communication in sign language. IEEE Access.

[19]Sharshar, A., Khan, L. U., Ullah, W., & Guizani, M. (2025). Vision-language models for edge networks: A comprehensive survey. IEEE Internet of Things Journal.

[20]Zou, J., Li, J., Tang, J., Huang, Y., Ding, S., & Xu, X. (2024, October). Sign Language Recognition and Translation Methods Promote Sign Language Education: A Review. In 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 3479-3484). IEEE..