International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 12 | Dec - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

AUDIO TO SIGN LANGUAGE CONVERSION

Lisha Kurian[1], Sreelakshmi K Anil[2], Souparnika Abhilash[3] ,Kashmeera Valsan[4] ,

Vinaya Venugopal[5]

[1]*Assistant Professor, Dept of CSE, Sree Narayana Gurukulam College of Engineering, Kochi, India,* lishakurian@sngce.ac.in

[2]Student, Dept of CSE, Sree Narayana Gurukulam College of Engineering, Kochi, India, Sreeluxmi787@gmail.com

[3]Student, Dept of CSE, Sree Narayana Gurukulam College of Engineering, Kochi, India, Souparnikaabhilash29@gmail.com

[4]Student, Dept of CSE, Sree Narayana Gurukulam College of Engineering, Kochi, India, kashmeeravalsan05@gmail.com

[5]Student, Dept of CSE, Sree Narayana Gurukulam College of Engineering, Kochi, India, vinayavenugopal346@gmail.com

---

Abstract - Due to high demand for inclusive technologies and the rapid electrons communication age, there is growing interest in automatic audio to sign language conversion systems. This article has given the relevant way to solve this problem by converting spoken language into sign languages in real time using machine translation that can shape communication path for deaf and dumb community. In our method, we use Natural Language Processing (NLP) to transcribe audio into text and then deep learning for sign language generation. The system recognizes keywords and important phrases in the transcribed text, associates them with corresponding sign language gestures, and presents such signs animations using an avatar-based visualization. The recent accomplishment is different from previous systems that work directly and only translate a sentence to its corresponding sign, rather focusing on the contextual analysis for better conveying of meaning.

***Key Words*:** Audio to Sign Language Translation, Speech- to-Text Conversion, Natural Language Processing (NLP), Machine Translation, Deep Learning, Gesture Recognition, Avatar-based Sign Language Visualization, Accessibility, Multilingual Sign Language

1.INTRODUCTION

Communicational barriers between the hearing people and the deaf or hard-of-hearing people may prove pretty challenging in personal life and professional environments as well. Such people relying more on sign language to communicate lack the availability of interpreters or the efficient tools for real-time interpretation, which hinders a person's ability to convey information fully in spoken words. The major impact of this barrier will be in the socialization and the way information or services are sought. This is where audio to sign language detection technology can act as a bridge in real time by translating speech into the gestures of sign language. The latest advancement in AI, particularly natural language processing and computer vision, will increasingly make direct conversion from spoken language into sign language feasible. This technology provides much hope for inclusive communication and uses audio processing in writing speech into text, then converting this text using gestures through recognition algorithms. Thus, this paper aims to present a framework for translating sign language. through the detection and processing of audio-based input in Python. This system integrates speech-to-text conversion with gesture mapping for accurate real-time communication transcending language barriers.

This paper is divided into the following sections: First, we discuss related work in speech recognition and sign language translation technologies. Then we get into the methodology part of the paper, where audio processing, natural language processing, and gesture recognition techniques are discussed. Next, we present the experimental setup, data collection process, and system evaluation. Finally, we discuss the results, limitations, and potential future enhancements for audio-to-sign language detection systems. This paper contributes to accessible and inclusive technologies for the deaf and hard-of-hearing community.

2. AUDIO TO SIGN

The Audio to Sign Language Conversion facilitates communication between the hearing impaired and the speaking population. Seamless interaction occurs between users when audio input is converted real-time into visual sign language gestures. Its modules include well-defined modules that relate to the processing of the audio input, natural language understanding, sign generation, and visualization of gestures cohesively to achieve an efficient user experience.
The system underlines accessibility as well as inclusivity: features allow for the support of multilingual sign languages, real-time processing, and customizable avatar-based outputs.

Advanced machine learning (ML) and natural language processing (NLP) algorithms analyze audio inputs to provide accurate conversion corresponding to the signs while acknowledging linguistic subtleties.
Every interaction is allowed to happen in real-time with minimal latency, to have live conversations. Regional Sign Language databases are also provided for integration in support of diverse user communities. Central elements of the Audio to Sign Language Conversion system involved speech recognition, linguistic analysis, sign generation, and visual representation.
By combining NLP techniques to interpret context and ML models to map text into signs, the system ensures contextual relevance and fluency. Users experience privacy with local data processing, reducing dependence on external servers, and maintaining data confidentiality. This solution provides an optimal balance of functionality, usability, and inclusiveness, promoting seamless communication across diverse audiences.

3. LITERATURE REVIEW

Globally, over 300 different sign languages are in use today. This diversity causes a major communication challenge in the deaf community, particularly when they interact with other people from different regions. This gap has been used to research technologies that may unify or translate between sign languages, making communication much easier. Early methods employed motion-capture systems like the Swiss Ranger SR-2 sensor, which captures depth and intensity to track gestures in 3D space. Depth cameras, like the Kinect, have pushed these systems forward with their ability to enhance hand and body tracking even in low lighting. Recent developments involve the application of machine learning techniques, especially SVMs and neural networks, in enhancing recognition through gesture trajectories and complex hand shapes and movements classification. The universal sign languages were proposed to work toward a standard that might bridge regional variations. Real-time translation tools that possess cameras and sensors can even capture gestures via text or speech that then allow interactive communication between the deaf. Though much progress has been made, universal solutions will be hard to produce based on the reasons of cultural adaptability, real-time performance, and affordability. Further research must be pursued in the manner of tool development to facilitate inclusive communication among the community of the deaf. [1]

Sign language recognition" has moved from the earliest sensor-based systems which required gloves through to the most current vision-based approaches which use cameras. According to Bowden et al. (2003), early vision or camera-based systems found an improvement in the accessibility of natural signing behavior but were a failure in uncontrolled environments. Deep learning has revolutionized SLR, especially in the application of CNNs and RNNs. CNNs have been most effective in extracting spatial features, while RNNs, especially LSTM networks, have managed the temporal aspects of sign language. Such a combination has been proven to enhance gesture understanding, as in the case of Huang et al. (2015). According to Pigou et al. (2018), hybrid models built using the integration of CNNs and LSTMs have obtained better performances with dynamic gestures. Thorough datasets are required to train efficient SLR models. Valuable public datasets exist, such as the RWTH-PHOENIX-Weather 2014 dataset; however, no large-scale datasets are currently found for Indian Sign Language. Other areas for research are synthetic data generation by researchers such as Pansare et al. 2020, data augmentation, and so on. Non-manual markers - facial expressions and body language are important because most meanings in SL are expressed with these. Real-time capability is a must for its applied use. The optimized models should result in faster inference without any trade-offs on accuracy to improve usability. Increasingly important, therefore, are user-centered design principles, so systems have to be intuitive to use. It is vital that engagement with the deaf and hard-of-hearing community is part of developing culturally relevant SLR applications. Increasingly, data privacy and consent-related ethical considerations and biases in the training of models are commonplace. These issues will ensure effective and accurate use of SLR systems. Though SLR has associated challenges such as dataset limitations and the need for real-time processing, the advancements in deep learning and user experience focus are promising directions for SLR Future research will include diverse populations and be easier for the hearing impaired to gain proficiency in oral communication. In addition, further collaborative endeavors between technologists, linguists, and members of the deaf community will allow for the development of advancements in SLR technology through increased accuracy. Explored new machine learning algorithms will increase efficiency by showing greater recognition accuracy. [2]

The communication gap between deaf individuals and the hearing population presents significant challenges, as traditional sign language remains largely unrecognized by those unfamiliar with it. Many research articles demonstrate the potential that technology may hold in reducing or eradicating such divides. For example, Amit Kumar Shinde refers to the significance of hand gesture recognition systems that will bridge gaps between deaf and hearing individuals without the necessity for interpreters. Neha Poddar et al., in their paper, mentioned the significant prevalence of deafness in India and appealed for portable devices that are capable of translating complex issues, such as mathematics into sign language, for improvement in educational accessibility. Researchers Anbarasi Rajamohan et al. explained glove- based interpreters for demonstrating how flex sensors, along with accelerometers can capture gestures, but, of course, have limited applicability due to financial and technological instability. Neha V. Tavari and others discuss the integration of hand images captured through webcams into the classifications but highlight hardware-related challenges. The development discussed here, under this paper, is the audio- to-sign language translation system, where the issue on spoken language to ISL translation is addressed through the use of Python. It employs speech recognition through Google's API, natural language processing in parsing text, and dictionary of predefined gestures to produce visual representations of words that are spoken. In that regard, it becomes pretty clear that NLP serves as the backbone for communication and understanding, and that is why it is such a promising avenue for enhancing the interactions in different settings - from educational institutions to public services. Further improvement of the system can be achieved by introducing facial expressions to sign language. It will further improve the system, and hence the application of the same can be used in real-time communication for news broadcasting and the like.[3]

The literature on audio-to-sign language conversion emphasizes the contribution of the technological advancements to the betterment of deaf and hard-of-hearing communities in terms of enhancing communication. Sahu et al. (2020) explored deep learning methodology for real-time transformation of sign language, whereby the high accuracy of its recognition was achieved in regards to the spoken language. Zhao et al. (2018) proposed an approach that used deep learning algorithms in developing a translation system between speech and sign language to demonstrate the role of machine learning models to facilitate effective communication. Huang et al. (2019) further added to the research by applying deep learning for audio inputs translated into sign language gestures by emphasizing accuracy in audio recognition. The real-time capability of conversion was highlighted by Parate and Upline (2020), with the support of Python that makes the interaction efficient. The involvement of NLP techniques discussed by several authors shows how context and semantics play an important role in improving the accuracy of translations. Lin and Wei (2021) proposed a hybrid approach that improves the quality of translation with LSTM added to CNN. Advanced designs of neural networks thus fulfill some gaps in effective communication for better translation efficiency based on emphasis on the current contributions of deep learning into further improvement in natural language processing capabilities for the successful establishment of audio-to-sign systems which may encourage societies to become inclusive for helping individuals become members of an ecosystem for persons with an aural limitation.[4]

This paper presents a novel approach to speech-to-Indian Sign Language (ISL) translation using a multi-task transformer architecture. Our method exploits a unique dataset containing audio and sign language annotations, which allows training and evaluation effectively. This work fundamentally advances assistive technologies so that communications between the hard of hearing and the main stream would be possible due to real-time sign-language generation based on speech and provides interaction with the flexibility and intimacy of spoken words, thus capturing the rich nuances involved in communication over voice as well as incorporating intonation and emotion as a medium. The use of a multi-task transformer network in the model is innovative because it accelerates the generation of sign language while bringing deeper meaning to the spoken input. Our dataset is fundamentally important to the future for the appropriate speech-level annotations lacking in existing sign language datasets. Real-time translation applications could be empowered with better communication in such settings as education, health care, public services, etc. The reason behind this is that it takes into account the segmentation of sentences as a serious drawback of previous approaches and concentration on continuous speech rather than segmented sentences, thus reaching the goal of more accurate representation of spontaneous conversation. This cross- modal discriminator further enhances the model's ability to discern relationships between different modalities, and its sign language outputs become more coherent and contextually appropriate. Our results show that generating sign language from speech can enhance understanding and preserve the expressiveness inherent in both languages. Our Indian Sign Language dataset is a landmark step in the direction of recognizing and supporting regional sign languages around the world within the research world. Ultimately, it presents new avenues for developing AI-driven aids in communication that empower the hearing-impaired and facilitate the normalization of daily, unstructured interaction with them.[5]

The literature in sign language recognition has discussed various methodologies that have evolved for communication among the deaf and hard of hearing. The traditional glove-based method, as reported by Wang and Popovic (2009) and Deora and Bajaj (2012), though precise, does not work effectively for practical use because they are quite complex and costly. While the vision-based approaches have gained momentum lately, primarily because they are easier to implement and more flexible, such as the appearance-based method and hand model-based system (3D) that is applied in this method nowadays (Cheng et al., 2016). Many studies were conducted to discuss the feature extraction methods: Discrete Wavelet Transform, Zernike moments, and Histogram of Orientation Gradient, which each achieved different recognition accuracies (Tripathi et al., 2015; Nanivadekar and Kulkarni, 2014). The important contribution regarding co-articulation detection for dynamic gesture recognition was from Bhuyan et al. (2006). Other researchers are Kalpana Sharma and Dutta (2014); they put forward the requirement of capturing hand gestures quite precisely and used Zernike moments. Machine learning algorithms are also important with multi-class SVM, with which enhanced performance of recognition systems are achieved (Kumar et al., 2016). In that direction, present efforts are mainly focused on the achievement of real-time recognition systems which overcome the complications of the background and lighting, all of which have been described in the developed methods from this study. In this context, there is a wide integration of these vision-based techniques along with co-articulation detection combined with novel feature extraction approaches, allowing for more interactive and user-friendly sign language recognition systems which will enhance accessibility and make the inclusion of the hearing- impaired community much easier.[6]

This paper introduces a new approach in audio-based video analysis, specifically focusing on the utilization of audio features for video summarization and categorization. It underscores the importance of audio information in video content analysis, an area that has traditionally focused on visual data. The proposed model uses 40 Mel-Frequency Cepstral Coefficients features extracted from audio to train various machine learning classifiers, including Artificial Neural Networks, Support Vector Machines, K-Nearest Neighbors, and Decision Trees. This research follows the emerging trend of works that consider the capabilities of audio features in video analysis and the limitations of earlier approaches based on the predominance of visual content. The authors have referenced the evolution in methods of audio detection and classification, from traditional machine learning into deep learning methodologies, adopting architectures such as RNNs and CNNs for the purpose of increasing their accuracy and efficiency. Testing on the Urbansound8k dataset shows impressively high accuracy rates with the model, demonstrating a potential use in real- time scenarios from infrastructure monitoring to criminal investigation. It reflects the critical necessity for audio-based approaches in video analysis, which not only throws additional light on contemporary approaches and research lines in improving

audio feature role in the respective domain but also underlines their applications in several diverse domains.[7]

4. PROPOSED METHODOLOGY

This system develops an advanced model to identify the spoken language and then converts the information into sign language representations, addressing the variability in speech and the different hand signs that are used across the sign languages. The methodology has phases including data collection, preprocessing, feature extraction, and model evaluation.

1. Data Collection

Collecting an extensive database of language in speech on audio that is accompanied or not with videos or written notations in sign languages. Speakers will vary because of their different accents or speaking styles. Audio recordings can be sourced from such places as existing speech databases, educational materials, recorded speeches. Data will be collected from video repositories containing sign language interpreters for varied representation in different sign languages. Documentation of metadata such as the speaker demographics, recording conditions, and context of the spoken language will also be considered to add value to the analysis.

2. Preprocessing

Audio recordings will be preprocessed after data collection for analysis preparation. Processing steps involved are: Noise Reduction: Spectral subtraction is used to suppress the background noise and make the spoken language more audible. This will therefore enhance feature extraction. Segmentation: Voice activity detection algorithms will be used in segmentation. This means that only the relevant segments of the audio file will be used in analysis. Normalization: Normalize the audio recordings to the same level of volume for uniform data.

3. Feature Extraction

Once preprocessing is completed, then feature extraction comes into the scene. It includes: Audio Feature Extraction: It includes methods such as Mel- Frequency Cepstral Coefficients (MFCCs) and spectrogram analysis for extracting the most prominent acoustic features representing the spoken language. Such features would be representing frequency and temporal characteristics that need to be precise for recognition purposes. Temporal Encoding: In addition to sequence-based features like phonemes or prosody, the dynamics of spoken language over time would be captured, which are essential for translation into sign language gestures.

4. Data Augmentation

Data augmentation techniques will be applied in order to make the model more robust and counter the class imbalance. Techniques to be used include: Pitch Shifting: The pitch of audio samples will be changed to create variations that increase the diversity of representations of spoken language. Time Stretching: Audio samples stretched to change the duration without changing the pitch so that the model observes different speech rates. Noise Injection: Synthetic noise injected into audio samples to mimic natural environments to increase the strength of the model against various acoustic conditions.

5. Model Building

In classification, a deep learning model shall be developed on architectures of RNNs, LSTMs, or CNNs combined with attention mechanisms to effectively capture the temporal relationships between audio features and their corresponding sign language representations.

6. Model Training and Evaluation

The dataset is divided into three subsets - training, validation, and testing subsets usually in 80:10:10 ratio. Training phase will apply to the training set: we will use it during training of the model with intent to learn the hidden dependence between audio features and the sign language representation. Cross Validation phase: this will utilize validation set to help fine tune the selected set of the hyperparameters and thus keep us from over-fitting a training set. Testing: the testing set, will provide an estimation for the real performance of our model, which pays regards to unseen audio, hence it can give us evaluation over its generalization capability, as a conclusion.

7. Performance metrics

To evaluate the model's performance in transducing spoken language to sign language, the following performance metrics are employed: Accuracy: The percent of correctly predicted sign language gestures. Precision and Recall: Metrics to evaluate the performance of specific sign gesture recognition by the model. F1-Score: Harmonic mean of precision and recall to deliver a balanced measure for performance. Confusion Matrix: A summary of the model's predicted results on various categories of sign language that will help to identify any misclassifications and areas where improvement is required.

8. Error Analysis

After the performance evaluation of the model, an error analysis is done to better understand what the model does well and what it doesn't do well. It involves review of misclassified instances, identification of the patterns of errors, and deciding on the changes that may be required for the accurate model.

5. ARCHITECTURE

The diagram illustrates the workflow of an **Audio to Sign Language Conversion system** that explains the steps in converting spoken language into visual sign language gestures. The process starts with the **Input Audio**, which is spoken language captured via a microphone or audio file. This raw audio data is then processed in the **Speech Recognition** stage, where Automatic Speech Recognition (ASR) technology transcribes the spoken words into text. Once transcribed to text, the system applies **NLP** to analyze the text while considering linguistic patterns in order to be able to translate that to signing language accurately. Finally, following the NLP phase, the **Sign Language Translation** module translates the input text into sign language movements through a database of known movements and machine learning algorithms. The final translation in the **Display Sign** phase is then presented as moving avatars or graphic representations of hand movements. All these phases are set in place to ensure an efficient and real-time method for signing that is both precise and accessible to users of the signing system.

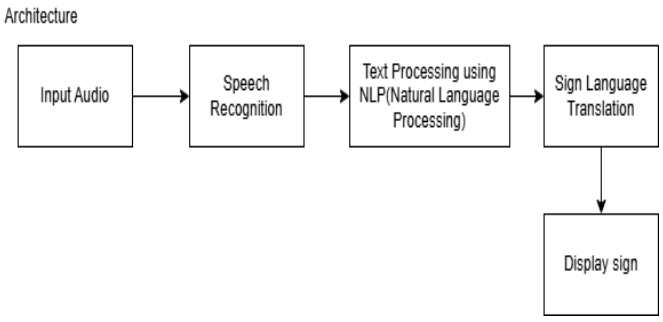

Figure 1: Architecture Diagram

6. CONCLUSIONS

The audio-to-sign language detection system holds much importance in filling a long-gone gap among the deaf and hard-of-hearing population. Using superior deep learning techniques for interpreting audio features, the intention is to translate the speaking language to the corresponding signs in the sign language correctly. This translation is also a medium for furthering accessibility by promoting inclusivity so that people who use sign language can fully interact in numerous settings. Data collection should be made in a planned manner by ultimately capturing divergent datasets that encompass the multiple nuances of accents, dialects, and patterns of speech. Thorough preprocessing, such as noise reduction and segmentation, improves the quality of audio input to ensure that the model accurately identifies spoken words. Powerful feature extraction techniques, including MFCCs and other audio descriptors, provide a solid representation of the audio data for translation to be possible. With this, the machine learning model allows the system to learn from huge amounts of data, therefore enhancing the accuracy and reliability with time. Such real-time processing abilities will enable the users to get instant translations. In this way, these systems will make them more adept to communicate effectively in everyday life. With such a system, when technology gets advanced, communication and mutual understanding between different communities may be facilitated. It all finally leads to a well-connected and fair society with its interactions and relationships where nothing is held back because of language barriers.

7.REFERENCES

[1]. Vinay Kumar K*, R.H.Goudar*, V T Desai** *Dept. Of CNE, Visvesvaraya Technological University, Belagavi-590018, Karnataka **Dept. of MCA, KLE Dr. M S Sheshgiri College of Engineering and Technology, Belagavi

.

[2] Ambreen Sabha1*, Arvind Selwal2 1, 2Department of Computer Science and Information Technology Central University of Jammu, Samba, Jammu, and Kashmir, India-181143

[3]. Amit Kumar Shinde and Ramesh Khagalkar “sign language to text and vice versa recognition using computer vision in Marathi” International journal of computer Application (0975-8887) National conference on advanced on computing (NCAC 2015).

[4]. Huang, Y., Zhang, M., Huang, H., & Zhu, J. (2021). Audio-to-Sign Language Translation Based on Deep Learning. In Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV), 128-133.

[5]. A Video Based Indian Sign Language Recognition System (INSLR) Using Wavelet Transform and Fuzzy Logic by P. I. V. Kishore and P. Rajesh Kumar

[6]. G. Fang and W. Gao, “Large vocabulary continuous sign language recognition based on transition-movement models,” IEEE Transaction on Systems, MAN, and Cybernetics, vol. 37, no. 1, pp. 1-9, January 2007.

[7]. B. O. Olusanya, K. J. Neumann, and J. E. Saunders, “The global burden of disabling hearing impairment: a call to action,” Bull World Health Organ, vol. 92, no. 5, pp. 367–373, May 2014.

[8]. J. Zelinka and J. Kanis, “Neural sign language synthesis: Words are our glosses,” in 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3384–3392

.

[9]. S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Ma lik, “Learning individual styles of conversational gesture,” in IEEE/CVF Conference on Computer Vision and Pattern Recog nition (CVPR), 2019, pp. 3492–3501.