



Title

Speech Recognition System Using Deep Learning Techniques for Assamese Sign Language

Prepared By

Name: Sultan Ali Senan

ID No.: 253002563

At

Collage of Informatics

Midocean University

UAE

Supervisor Name: Dr. Amjad Qatish

Collage of Informatics, Midocean University, Comoros

Acknowledgements

Special thanks to Dr. Amjad Qatish.

Abstract

Despite the significant abilities of the deaf and hard of hearing to communicate, many of them suffer from isolation due to inaccessibility issues. Technical complexities that arise in linking programming languages to speaking languages remain to be one of the major challenges that make the implementation of assistive technologies difficult. This study aims at looking into the linguistic environment in India in terms of language diversity focusing on the Assamese language. Though important in Indian culture, there is little information regarding Assamese and how it can be incorporated in AI.

This problem can be addressed through the development of an innovative architectural design based on learning principles. To achieve this, we developed a segregation method specifically for Assamese language whose deviations from other languages prove to be satisfactory. This leads to the creation of a custom dataset from scratch after which we implement a practical survival model in order to attain productivity.

Table of Contents

List of Abbreviations	vi
List of Figures.....	vii
List of Tables	viii

Chapter 1: Introduction and

Overview.....

1.1: Overview

1.2 Problem Definition

1.3 Thesis Objectives

1.4 Thesis Contribution

1.5 Thesis Organization.....

Chapter 2: Related Work

Chapter 3: Proposed Approach

4.1 Introduction

Chapter 4: Experimental Results and Discussion

Chapter 5: Conclusion and Future Work	
7.1 Important Features	
7.3 Future Work	
References	

List of Abbreviations

- **Acc.:** Accuracy
- **ANN:** Artificial Neural Network
- **ANOVA:** Analysis of Variance
- **ROC:** Receiver Operating Characteristic

List of Figures

2.1 Educational Data Mining Process	7
2.2 Applications of Educational Data Mining	8
2.3 Distribution of Papers that used different attributes for PSP problem	19
2.4 Student Performance Data Set	21
2.5 Educational Process Mining (EPM): A Learning Analytics Data Set ...	23

List of Tables

2.1	Student Performance Data Set Description	21
2.2	Students' Academic Performance Dataset Description	23

Chapter 1 Introduction

Introduction

Despite the fact that communication is a basic human need, millions of deaf and hard-of-hearing people worldwide still struggle with accessibility issues in their day-to-day lives. Sign language is an essential form of communication within deaf communities, even though machine translation between sign language and spoken language is still a major technical challenge. There is a critical need for intelligent systems created especially for these linguistic contexts because local sign languages in India, like Assamese Sign Language, receive little research and development attention.

Recent developments in deep learning have drastically changed computer vision, speech recognition, and natural language processing. Convolutional neural networks, recurrent neural networks, transformer-based designs, and long-term memory networks have all shown remarkable performance in recognizing intricate visual and sequential patterns. These models are ideal for sign language recognition, which necessitates the precise capture and interpretation of hand gestures, facial expressions, and motion dynamics.

There is a need for the development of an Assamese Sign Language recognition system using deep learning for speech recognition to overcome communication challenges through conversion of signs into spoken language. An Assamese Sign Language recognition system comprises of speech synthesis, sequence modeling, gesture detection, feature extraction, and video capturing. By utilizing deep learning, an Assamese Sign Language recognition system can be developed that can automatically obtain temporal and geographical information of gestures with little manual effort.

This innovation in the field of technology can be considered a breakthrough in multiple ways, since it will benefit assistive communication technologies, promote linguistic diversity, facilitate the preservation of sign languages through digitization, and contribute towards future research efforts in developing language technologies for resource-limited communities (Kagalkar, 2025).

In this study, we will focus on the development of an advanced voice recognition system based on deep learning and designed to recognize Assamese Sign Language with high precision and accuracy..

1.1 Problem Definition

Even though significant progress has been made in technologies that can recognize sign languages, most existing systems cater only to popular languages like ASL or BSL. The Assamese Sign Language still stands at a preliminary stage without any technical assistance. The lack of automatic translation tools for the Assamese Sign Language makes it difficult for the deaf community in Assam to communicate, hence making them inaccessible to the general services and opportunities available to people outside their community. Therefore, it becomes imperative to design a fast and efficient tool that converts the gestures in the Assamese Sign Language into voice using deep learning algorithms.

1.2 Thesis Objectives

Primary Objective

The development of a speech recognition system based on deep learning for the Assamese Sign Language.

Secondary Objectives

1. Acquisition and preprocessing of data related to Assamese sign gestures.
2. Development of deep learning models for the detection of sign gestures.
3. Implementation of sequence modeling for the analysis of dynamic signs.
4. Incorporation of a speech synthesis unit for the generation of real-time audio outputs.
5. Performance evaluation of the proposed system using standard parameters such as accuracy, precision, recall, and F1-score.

1.3 Research Questions

- How can deep learning methods improve the accuracy of Assamese Sign Language identification?
- Which neural network design works best for recognizing dynamic gestures?
- Is it possible for the system to translate in real time?
- What difficulties arise when using local sign language with limited resources?

1.4 Thesis Contribution

The thesis is important from both technological and scholarly standpoints, as well as making valuable social contributions to assistive technology and recognition of regional sign languages, which suffer from lack of resources. Some of the major contributions are:

- Creation of a New Dataset for Assamese Sign Language: The key contribution of this thesis involves creation of an indexed and well-structured dataset for Assamese Sign Language. As there is very little data freely available for this language, it serves as the cornerstone of all further researches in the field.

- Framework for Deep Learning-Based Sign Recognition System: This thesis introduces and applies a framework for creating a machine learning-based system, combining CNNs, used for extracting spatial features, with LSTMs, employed to process temporal information in the form of gestures' motion dynamics.

- Sign-to-Speech Conversion Model: The author develops a working model that is capable of converting signs of ASL into English speech in real time. In addition to a theoretical discussion, it provides a practical tool in the sphere of sign language recognition and assistive technologies.

- Using Technology to Address Inequity: The problem of resource disparity is addressed here, as many systems of sign language recognition focus on languages that are most commonly used. In contrast, this work promotes the concept of inclusive artificial intelligence in terms of supporting resource-limited sign languages.

- Evaluation and Comparison of Model Configurations: The paper offers a detailed analysis and performance evaluation of the suggested models. Additionally, the thesis includes several comparison cases that determine the optimal model configuration for ASL recognition.

- Accessibility and Social Impact: In addition to its technical accomplishments, this project promotes social inclusion by providing a means of communication for Assam's hearing and deaf communities. The system encourages accessibility in work, education, and public services.

1.6 Thesis Organization

The study challenge, research methodology, application, and findings of the suggested deep learning-based speech recognition system for Assamese Sign Language are presented in many chapters of this thesis.

Chapter 2 Related Work

For those who are deaf and mute, communication is undoubtedly challenging. They communicate with each other and with others using sign language. Nonetheless, a significant percentage of people who can communicate both orally and in writing do not know sign language. Therefore, it is clear that technical tools for sign language interpretation must be developed. Numerous studies have been conducted to use technology to recognize sign language for the majority of the world's languages. Nonetheless, techniques and tools specific to regional sign language dialects still need to be developed. India is home to 22 contemporary languages in addition to more than 19,000 mother tongues. The goal of this project is to provide a technical method for identifying Assamese Sign Language, one of India's 22 contemporary languages. The goal of this project was to use machine learning techniques to create a system for identifying hand movements in Assamese Sign Language. Photographs of Assamese gestures in two and three dimensions were used to create the dataset. The process of feature detection on the images was done through the use of the MediaPipe framework. A dataset with 2094 data points for the Assamese sign language involved nine motions that included vowels and consonants (অ, ই, ঈ, উ, এ, ও, ' , ঢ, '). Using the dataset, the feedback forward neural network was trained, leading to 99% accuracy. This shows that the process used in this research is efficient in the recognition of letters and other motions in the Assamese sign language. This process can be used to evaluate other Indian regional languages (Bora, 2023).

The deaf people experience great difficulty communicating with people who have normal hearing capabilities. Employing the human interpreter, one who converts the sign language to

understandable language for the hearing people, cannot be effective. A Sign Language Recognition System (SLRS), which helps to decrease the gap between the hearing community and the deaf community through automatic translation of signs to text or speech without requiring the help of an interpreter, is considered the most suitable and efficient approach. The purpose of this research paper is to use computer vision techniques for word recognition in ISL. The current computer vision-based methods for ISL word recognition are inefficient due to excessive preprocessing, such as gathering features from a large sequence of frames. In order to counteract the issue of over-processing, (Das, A deep sign language recognition system for Indian sign language, 2023) propose a novel Computer Vision-based ISL recognition system known as the CNN-BiLSTM Hybrid Classifier Based Language Recognition System (HCBSLR). The proposed model improves system accuracy and efficiency by removing unnecessary or redundant frames using a keyframe extraction method based on histogram difference (HD). The HCBSLR system uses a Bidirectional Long Term Memory Network (BiLSTM) to extract temporal features and the VGG-19 algorithm to extract spatial features. The average accuracy of the proposed system was 87.67% when compared to some existing SLRS systems. Experimental results show that the proposed HCBSLR system performs more accurately and efficiently than the existing SLRS systems (Das, 2023).

A sign language can be described as a visual-kinetic form of communication that uses various movements, hand signs, facial expression, and body posture in conveying information. This language plays an important role in facilitating human communication because of its grammatical and morphological rules. The gap in communication between users and non-users of this type of language is among the major challenges. There are more than 63 million persons who have a problem with hearing in India. As such, there is a potential for research concerning

sign language recognition and translation. These techniques are critical for improving the quality of life of hearing-impaired persons while promoting social inclusion and communication among persons. Currently, a lot of effort has been made in developing sign language recognition techniques because of their crucial importance in human-computer interactions, interpretation, learning aids, and communication assistance. This study focuses on current trends in Indian sign language recognition techniques. It explores the available data sets, existing approaches, and aspects of performance in this area.

According to (Chetia, 2025), Convolutional Neural Networks (CNNs), an advanced deep learning methodology, are employed for the classification of movements of Assamese Sign Language (ASL). The framework involves three stages: data collection, CNN model creation, and live motion recognition. Data was collected by recording ten different static motions signifying Assamese alphabet letters utilizing the frame capturing capabilities of OpenCV and hand feature extraction from the frames through MediaPipe. Post data processing, the gesture image data was divided into training and test datasets, and the CNN model was trained. Following training, the CNN model classified the gestures with 100% precision and excellent accuracy, recall, and F1 values across all ten gestures. These results reflect the effectiveness of the CNN model in recognizing and learning the essential features of the Assamese Sign Language movements, allowing for real-time recognition of signs and their translation into text. The implementation of the developed system can have profound implications for communication between the hearing and deaf communities of Assam. The future work will involve collecting more data related to dynamic hand movements for live motion recognition within a mobile application setting.

Sign language acquisition would become easy due to the existence of an interactive Sign Language Learning System, which is one of the most advanced systems ever invented to facilitate the process of sign language learning. Some of the features of this innovative system include built-in audio output, text-to-gesture conversion, multilingual support, and sign language letter and word recognition. To accomplish the above task, advanced algorithms of sign language recognition, image processing, and machine learning are utilized to identify hand gestures and movements. One of the computer vision algorithms, recurrent neural networks (RNN), and convolutional neural networks (CNN), among others, are some of the methods through which input images and videos could be analyzed and classified to determine appropriate sign language words or symbols for letters and words identification. On the other hand, natural language processing (NLP) algorithms are applied during text-to-gesture conversion. During this stage, input texts' meanings are evaluated before converting them into appropriate gestures and movements. In turn, the level of accuracy of the system, which refers to the extent to which sign language symbols or words could be recognized, will depend on the efficiency of the algorithms and training data used.

Sign language is the best mode of communication for deaf individuals. Although there have been advances made in technology, there is still a communication divide between those who can hear and those who cannot. This study seeks to bring about harmony between these two parties by creating a system which translates voice to Hindi Sign Language via avatars. The system operates in three stages: (a) Phase 1: Punjabi, Hindi, and English single word speech recognition in a speaker-independent environment. (b) Phase 2: The original language is translated into Hindi Sign Language. (c) Representing motions in Hindi Sign Language with a 3D avatar based on HamNoSys technology. The four main SISLA system implementation

modules are requirements analysis, data collection, technological development, and evaluation. The system's multilingual capability increases its efficiency. Speech sample files in Punjabi (12600, 4193), Hindi (12610, 4211), and English (12660, 4218) were used to train and test the speech recognition models. Experimental machine translation results showed that the recommended trained models achieved accuracy of at least 91%, 89%, and 89% for English, Punjabi, and Hindi, respectively. Sign language experts were also consulted in order to evaluate the sign error rate through feedback. It was proposed that sentence-level translation be combined with non-manual sign language components to enhance the system going forward. Usability tests confirm that the proposed SISLA system is suitable for communication and instruction for people with hearing impairments based on survey results (Dhanjal, 2022). Recognizing sign language is crucial to helping people who are deaf or have hearing issues communicate. The task of recognizing signs and translating them into natural language can be enhanced by deep learning techniques. The proposed method extracts features using three-dimensional convolutional neural networks (3D CNNs). This method can improve the accuracy and performance of real-time sign language recognition. In contrast to 3D CNN, 2D CNN, and 3D CNN-integrated models, this pilot study demonstrated the advantages and disadvantages of utilizing a 3D CNN architecture with LSTM for feature extraction in sign language recognition systems. The 3D CNN-LSTM model outperforms standard 3D CNN architectures in its ability to effectively understand the spatial and temporal features of sign language expressions, which is essential for accurate sign identification. This article also covers some techniques to enhance the performance of the 3D CNN-LSTM architecture. Finally, some unanswered questions and possible directions for future research in this field are mentioned. The analysis's conclusions demonstrate how promising the 3D CNN-LSTM design

is for enhancing SLR systems' accessibility and facilitating sign language users' communication (Paul, 2025).

This study explores a novel area of automated emotion recognition in Assamese, a widely spoken but little-studied dialect of Northeast India, using speech and signal processing techniques. Using a unique combination of language-independent data, advanced supervised machine learning techniques are used to classify a broad range of fundamental and derived emotions from spoken Assamese. A novel combination of impersonal and personal features is proposed for speech-based emotion recognition. High-dimensional features are first divided using the fuzzy C-means clustering technique. Thus, multiple random forest classifiers are used to facilitate the recognition and classification of emotions. Numerous emotions that typically pose challenges for classification systems are given particular attention with nested transformations. Performance analysis is done using K-fold cross-validation, and the results are confirmed by comparing them to basic random forest methods and validating a number of significant criteria. The fact that the method was able to outperform the baseline models through an increase in accuracy by 4.26% to 4.50% is quite intriguing because it means that successful classification can be achieved without the use of language information. This tool could be useful to medical professionals, enabling them to evaluate a patient's mental health and develop more effective treatment strategies. This study emphasizes the need to expand the scope of emotion recognition research to include a wider range of languages and has wide implications, particularly for understudied languages (Choudhury, 2023).

Speech conveys the emotional content of a speaker's position. Emotions can be extracted by looking at the voice signal in each sentence segment. Tone and emphasis are key indicators of tension and emotion in spoken communication. It's important to keep in mind that extracting

emotions is challenging due to the diversity of speech signals. Assamese and other tonal languages heavily rely on stress. The same expression can have a different meaning with even a slight change in tone. People use stress to encode and decode signals. (Bharadwaj, 2022) can identify correlations between stress and stress factors, including the fundamental frequency (F0), rate, duration, and severity, to track significant emotional differences. Variations in stress, which in turn reflect shifts in emotions, also contribute to variations in glottal closure moments. GCI, or epoch, refers to the critical excitation moments of the vocal apparatus with the sudden closure of the glottis. Emotions affect both the tone of voice and the arousal level of the speech signal. Understanding how emotions impact tone, the connections between their characteristics, and how this influences arousal strength (epoch) is the aim of this essay. Tone characteristics are calculated using speech analysis software such as MATLAB, PRAAT, and Python. These tone characteristics are then statistically analyzed using ANOVA to determine their correlations and identify tone shifts based on different emotions. Finally, GCI signals are extracted from the source of the stimulus, and MATLAB is used to analyze how they alter the same phrase under different conditions. By thoroughly examining the emotional content of speech through tone characteristics and the source of the stimulus, human-machine interaction frameworks for developing emotion recognition in speech, particularly in the Assamese language, for real-time applications like mental health treatment will be improved (Bharadwaj, 2022).

The ability to express oneself is a fundamental human need. People typically use speech to convey their ideas. In contrast, people who are hard of hearing or deaf express their feelings through sign language. They make gestures, and to decipher these signals, one must be familiar with sign language. The fact that not everyone is familiar with sign language is one of the main

obstacles to communication. (Ingoley, 2025) suggest utilizing deep learning methods to convert Hindi Sign Language into speech and text to remedy this. Through speech conversion, this makes it possible for deaf and hard-of-hearing people to communicate with illiterate or blind people. (Ingoley, 2025) developed a prototype that can decipher Hindi Sign Language numbers. These figures are helpful for representing time and distance in a variety of contexts, including financial organizations like banks and ATMs. The difficulties presented by the following circumstances are not sufficiently addressed by the sign language recognition systems in use today: 1) using the left or right hand, or both, to use sign language; 2) pointing the sign at the camera on a smartphone while taking a selfie; 3) wearing gloves to help communicate in cold or mountainous environments; and 4) the fact that a sign has multiple gestures, all of which need to be considered. (Ingoley, 2025) developed a diversified dataset called ISL_Num14_AS3N24 to overcome these problems. With their suggested methodology, which makes use of important hand points in the preprocessing step, (Ingoley, 2025) offered a system that makes use of deep learning technology. (Ingoley, 2025) Vgg16 and MobileNetV2, two pre-trained models, were altered to expedite training. The text-to-speech module makes use of the gtts library. The system performs remarkably well, with verification accuracy of 99.44% for the MobileNetV2 model and 99.76% for the Vgg16 model. Furthermore, the validation losses for VGG16 and MobileNetv2 were 0.9471 and 1.9921, respectively (Ingoley, 2025).

Indian Sign Language (ISL) is a visual-gestural communication method used by the deaf community in India. It uses gestures, eye movements, and body language as a potent communication tool. Spoken language and ISL have different grammatical rules. This study offers a means of bridging the communication gap experienced by deaf and mute communities

using sign language when compared to the general population. A real-time program that can translate sign language—more especially, ISL—into text is proposed in this study. The final text will be available in Hindi and Kannada to assist non-native English speakers. This application makes use of cutting-edge computer vision and machine learning technologies. The Media pipe platform was used to process pre-input video clips in order to extract features. Three models convolutional neural network (CNN), a recurrent neural network (LSTM), and Transformers encoders—were employed to determine the best model for the given problem. Since the results indicated that the accuracy levels of the LSTM and Transformers models were similar, ranging from 95 to 97%, they were combined to create a single model. The expected word sequence is then passed through a translation mechanism to turn this text into coherent Hindi/Kannada phrases. All things considered, this study offers a quicker and less expensive way to bridge the existing communication gap (Attili, 2024).

Speech recognition from audio signals has long been a major component of computer-aided services. Automatic speech recognition has employed several strategies, including different classification and speech analysis techniques, to extract information from signals. In this investigation, (Devi, 2023) introduced a new approach to automatically recognize speakers and their reaction tones from continuous speech. The voice dataset used in the trials was created using the widely spoken Assamese language, which has been combined with other Northeast Indian cultures. (Devi, 2023) provided a method based on a Gaussian Mixture Model (GMM) to address the crucial problem of controlling speaker background noise in contact centers. The estimates of voice power and the noise spectra were done by formulating equation sets using the Gaussian Mixture Models' mean vectors. Determination of the type of noise generation source and the signal-to-noise ratio input were among the first processes. A noise

model was then formulated by the use of the preliminary noise power spectrum estimate. The Wiener filter was used to reduce background noise through an improvement of the distorted speech estimate. Devi (2023) conducted the analysis of the synthetic and augmented speech data by assuming the known parameters of the uncertainty. The estimates of voice power and the noise spectra were done by formulating equation sets using the Gaussian Mixture Models' mean vectors. Determination of the type of noise generation source and the signal-to-noise ratio input were among the first processes. A noise model was then formulated by the use of the preliminary noise power spectrum estimate. The Wiener filter was used to reduce background noise through an improvement of the distorted speech estimate. Devi (2023) conducted the analysis of the synthetic and augmented speech data by assuming the known parameters of the uncertainty. The estimates of voice power and the noise spectra were done by formulating equation sets using the Gaussian Mixture Models' mean vectors. Determination of the type of noise generation source and the signal-to-noise ratio input were among the first processes. A noise model was then formulated by the use of the preliminary noise power spectrum estimate. The Wiener filter was used to reduce background noise through an improvement of the distorted speech estimate. Devi (2023) conducted the analysis of the synthetic and augmented speech data by assuming the known parameters of the uncertainty. The estimates of voice power and the noise spectra were done by formulating equation sets using the Gaussian Mixture Models' mean vectors. Determination of the type of noise generation source and the signal-to-noise ratio input were among the first processes. A noise model was then formulated by the use of the preliminary noise power spectrum estimate. The Wiener filter was used to reduce background noise through an improvement of the distorted speech estimate. Devi (2023) conducted the analysis of the synthetic and augmented speech data by assuming the known parameters of the uncertainty.

data by assuming the known parameters of the uncertainty. They last used a data aggregation technique using a Gaussian distribution to identify three groups of high, medium, and low pitch voices in order to identify three common human emotional states (Devi, 2023).

Part-of-speech tagging or POS is considered to be one of the most challenging aspects of natural language processing. The development of the Assamese POS classifier is the objective of the presented research. Due to the deficiency of digitized linguistic resources, the Assamese language has no effective POS classifiers. The purpose of the current paper is to fill this gap by using the Assamese POS dataset and Long Term Memory (LSTM) and Bi-LTM (Bi-LSTM) in order to create a POS classifier. It is important to note that the experiment could not handle natural language properly for computational linguistics as it could be expected. This paper has employed an Assamese dataset containing approximately 50,000 words. The classifiers examined the initial dataset including around 20,000 words and showed acceptable results in the first step. As a result, due to the good results obtained from the experiments, the Assamese dataset was increased up to 50,000 words. Accordingly, there was a significant rise in the quality of the F1 metric, precision, recall, and accuracy. Consequently, the accuracy for Bi-LSTM is 91.72% and 91.20% for LSTM. Comparison between the current study and available Assamese speech classifiers is provided below (Phukan, 2024).

Recent times have seen a lot of attention paid to the subject of automatic sign language recognition. Machine learning methods find their primary application in recognizing the movements of hands. At present, the majority of studies utilize a type of sign language that is specific to a certain country like American Sign Language. The Multilingual Sign Language Interpreter (MSLI) System is a machine learning-based multilingual sign language system that (Najib, 2025) presented in this research. MSLI trains a machine learning model using hand

gestures in multiple languages. The system is able to identify the provided sign language and its classifications. When test signals from the same language are input, the proposed system provides a two-stage recognition method. Only the first sign language is initially detected, after which the other signals are tested in accordance with the identified language. Furthermore, MSLI provides unique classifications for the signs of each language. For the experiments, eleven datasets in different languages were used. Both individual and combined classifications were applied to the input data. The experimental results demonstrate the accuracy of the proposed system. The proposed system's training accuracy varied between approximately 90% and 100% on most of the different sign language datasets that were used. The majority of the test data for each dataset had classification accuracy levels higher than 90%. The proposed MSLI integrated classification system had training and testing accuracy of 95.87% and 92.33%, respectively (Najib, 2025).

For those who are hard of hearing, sign language is an essential form of communication. But in some languages, like Saudi Arabia, there is a severe lack of sign language interpreters. Due to this lack, many deaf and hard-of-hearing people are unable to access services, particularly in public settings. By using technology to create systems that can recognize Arabic Sign Language through deep learning techniques, this paper seeks to close this accessibility gap. To represent the spatial and temporal components of sign language (i.e., letters and words), (Noor, 2024) suggested a hybrid model. The hybrid model uses a Long-Term Memory Classifier (LSTM) to extract the spatial and temporal characteristics for handling sequential data (i.e., hand gestures) and a Convolutional Neural Network (CNN) classifier to extract spatial features from sign language data. (Noor, 2024) generated a dataset of 20 distinct words to show the viability of a suggested hybrid model. It resulted in 4,000 images of augmented sign language

(ArSL): 500 video clips of 10 dynamic gesture words and 10 static gesture words. Their CNN and LSTM classifiers have promising results since they had an accuracy rate of 94.40% and 82.70%, respectively. This implies that their approach may significantly enhance the communicative capabilities of deaf Saudi Arabians. As such, this study makes a notable contribution towards achieving inclusivity and enhancing the quality of life of individuals suffering from hearing impairment (Noor, 2024).

Automatic Speech Recognition (ASR) is a key domain of studies during the age of human-computer interaction in a multilingual environment. ASR enhances the functionality of voice applications in multilingual environments. In a country like India, which is populated by individuals belonging to various linguistic and cultural backgrounds, people are compelled to use their native languages while interacting with machines. Thus, the creation of effective Hindi ASR systems is necessary to incorporate smart technologies into every aspect of Indian culture. The field of Hindi ASR has been expanding since the beginning of the twenty-first century, primarily due to the development of several standardized multilingual voice datasets for Indian languages. To the best of their knowledge, few attempts have been made to offer a comprehensive assessment of the significant developments in this field of study. (Spandan Dey, 2021) made one of the first attempts in this article to provide a comprehensive summary of the subject of Hindi ASR research. A detailed analysis is presented to highlight the challenges of limited resources and the interplay of factors preventing the development of speech recognition systems in the Indian context. This study covers important aspects of voice identification research in India, including a detailed description of the available voice datasets, important research contributions ranging from earlier attempts based on statistical modeling to contemporary approaches based on various neural network architectures, and future research

directions. This study will help assess the current state of speech identification research in India for any researcher working in or interested in related fields (Spandan Dey, 2021).

Emotion is an essential part of all human relationships. Emotions can be expressed through written text, conversation, body language, and facial expressions. Emotion recognition is an essential part of natural language processing and a key technology in human-computer interaction in many fields, including business, management, healthcare, and others. Due to the increasing use of social media and digital marketing, enormous amounts of unstructured data have been generated on social media platforms in recent years. This makes it challenging to discern emotions in text. To solve these problems, researchers have employed a range of deep learning and machine learning techniques. LSTM (Long Term Memory) and other deep learning techniques are quite good at identifying emotions in text. In this work, (Chutia, 2024) used a LSTM on an Assamese text to increase prediction and accuracy. Recent studies indicate that some Assamese emotion analysis research has been done. The goal of this work is to identify emotions in Assamese literature because research on text-based emotion recognition is still in its infancy (Chutia, 2024).

Indian Sign Language (ISL) is the primary means of communication for the deaf and hard-of-hearing in India. However, communication barriers persist due to ISL's lack of popularity. This work presents a real-time gesture recognition system that translates ISL motions into text and speech to enable communication between ISL users and non-users. The system uses long-term memory networks (LSTMs) to capture temporal patterns and convolutional neural networks (CNNs) to extract spatial features from sequential gesture data. A diverse dataset that incorporates variations in hand gestures, facial expressions, and body posture enhances the model's accuracy. Deep learning algorithms are incorporated to enable continuous

performance optimization. This approach promotes inclusivity by addressing the communication gap in public spaces, educational institutions, and workplaces. The suggested technique is an efficient solution that allows enhancing the accessibility and social inclusion of people with disabilities such as hearing impairments (Gadagkar, 2025).

(Kumar, 2023) provided information about the proposed real-time technique that recognizes ASL gestures based on current computer vision and machine learning approaches. Features of ASL signs are extracted through the CNN algorithm and Mediapipe framework. Results demonstrate that this technique can recognize all ASL letters with 99.95% accuracy. Therefore, it can be used as an integral part of hard-of-hearing communication devices. The suggested technique will positively affect the life of the hard-of-hearing individuals as it may be applied to other sign languages where similar hand signs are used. In general, the work provides a clear example of effective application of Mediapipe and CNN algorithms in detecting real-time sign language.

The American Sign Language (ASL) Glove is a remarkable piece of assistive technology, which enhances the communication capacity of individuals with speech, hearing, and vision problems. This research paper explains the process of designing and testing of the device, which is called the ASL Glove. It is a tool that translates sign gestures of deaf-mutes and provides the recipients of information with speech messages in real time. Flexible sensors and Arduino Uno processor are included into a special glove in order to interpret sign language and transfer it into text and speech format. The ASL Glove ensures complete inclusiveness by using a text-based display to show visual signs and the audio feature to allow auditory communication. Many tests and evaluations have proved a high level of gesture recognition and translation accuracy, therefore demonstrating the efficiency of this device as a piece of

technology that can help communicate easily. The ASL Glove eliminates any communication problems, which allows deaf people to take part in socialization and everyday life. Future improvements of this remarkable device will only contribute to increased social inclusiveness (Savale, 2024).

New ways of dealing with their problems of communication should be put into use if these deaf or hard-of-hearing people want to communicate effectively with the rest of the world. A new way of translating sign languages using the MediaPipe framework and the random forest technique is proposed in this research. Sign language recognition and the acquisition of key features about the gestures are more easily done using the MediaPipe framework. Using a random forest classifier will make it easier to distinguish between the signals made by these gestures. Using the MediaPipe framework and the random forest classifier, the communication process will become easy and efficient because it will involve the conversion of sign languages into written texts (Sharma, 2025).

The human computer interaction has been made possible by the technological revolution. Hand gesture recognition has gained prominence as a communication interface that is intuitive and friendly. In this study, a hand gesture recognition framework has been proposed for real-time detection to facilitate HCI. The technique will be instrumental in helping people with speech and hearing impairment. The hand gesture recognition system makes use of real-time cameras to collect hand movement information and develop a reliable customized dataset based on individual variability, concealment, and posture. A convolutional neural network has been utilized to capture the 21 most relevant features of each gesture detected. A random forest approach has been employed to classify the extracted features. An average of 94.58% classification accuracy has been achieved with various data. Conversion of the detected

gestures into voice and texts allows effective and comfortable communication. Gestures can be combined to formulate entire sentences, which are common in most social settings. Experiments conducted to test performance at different light levels have returned a signal-to-noise ratio between 3 to 4.27 decibels, implying high resistance to changes in illuminance. The graphical user interface with feedback helps to achieve efficient HCI by enabling easy two-way communication (Chandrashekar, 2025).

<u>Ref (author) year</u>	<u>paper title</u>	<u>description</u>	<u>dataset</u>	<u>methods</u>	<u>results</u>	<u>advantage</u>	<u>disadvantage</u>
Ravikumar (2024)	Interactive Sign Language Learning System using Deep Learning	An all-inclusive system for learning sign language with audio and text-to-gesture features.	Images and videos of sign language words and letters.	Image processing, CNN, RNN, and NLP for text-to-gesture conversion.	High levels of accuracy in text-to-gesture tasks.	Supports multiple languages and provides integrated audio output.	Accuracy is highly dependent on the quality of training data.
Dhanjal (2022)	An automatic machine translation system for multi-lingual speech to ISL	A system (SISLA) that translates spoken Punjabi, Hindi, and English into ISL.	Speech samples in Punjabi, Hindi, and English.	3D avatar based on HamNoSys technology for gesture representation.	Accuracy of 91% for English and 89% for Punjabi/Hindi.	Multilingual capacity enhances system efficiency and education.	Currently lacks non-manual sign language elements like facial expressions.
Paul (2025)	Sign language recognition using improved 3D CNNs	Pilot study comparing different CNN architectures for feature extraction.	Sign language expressions for spatial/temporal analysis.	Hybrid 3D CNN-LSTM model.	3D CNN-LSTM performed superior to typical 3D CNN architectures.	Efficiently grasps both spatial and temporal characteristics.	Being a pilot study, it notes several unresolved issues for further study.
Choudhury (2023)	Enhanced Emotion Recognition from Spoken Assamese Dialect	Classifying emotions from spoken Assamese using language-independent features. Investigating	Language-independent speech data from Assamese dialect.	Fuzzy C-means clustering with Random Forest classifiers.	Increased accuracy by 4.26% to 4.50% over baseline models.	Effective classification without needing specific language data.	Focused on emotion recognition rather than sign language translation.
Bharadwaj (2022)	Exploring human voice prosodic features for Assamese speech	how emotions affect tone and stress in the Assamese language.	Assamese speech segments and voice signals.	PRAAT, Python, and MATLAB for tone analysis and ANOVA for statistics.	Identified correlations between stress factors (F0, duration) and emotions.	Improves human-machine interaction for mental health applications.	Highly technical focus on prosody, less on automated recognition.

Attili (2024)	Real-time Sign Language Translation using Computer Vision	Translating ISL into text and then into Hindi and Kannada.	Pre-input video clips processed via MediaPipe.	Integrated model combining LSTM and Transformers.	Accuracy levels ranging from 95% to 97%.	Provides a quicker and less costly means of closing communication gaps.	Requires a translation mechanism to produce cohesive phrases.
Devi (2023)	Speech Recognition Via Machine Learning in Recording Studio	Automatic recognition of speakers and reaction tones in continuous Assamese speech.	Continuous speech dataset in Assamese language.	Gaussian Mixture Model (GMM) and Wiener filter for noise reduction.	Successful identification of high, medium, and low pitch voice groups.	Effectively controls background noise in recording environments.	Primary application is for contact centers and studios, not assistive sign tech.
Phukan (2024)	Parts-of-Speech Tagger in Assamese Using LSTM and Bi-LSTM	Creating a POS classifier for the Assamese language.	Assamese POS dataset containing 50,000 words.	Long Short-Term Memory (LSTM) and Bidirectional LSTM (Bi-LSTM).	Bi-LSTM accuracy was 91.72% vs 91.20% for LSTM.	Significant improvement in precision and recall with larger datasets.	Faced difficulties handling natural language for computational linguistics.
Najib (2025)	A multi-lingual sign language recognition system using machine learning	A system that recognizes signs across multiple different languages.	11 datasets in various languages.	Two-stage recognition: language detection followed by signal testing.	Training accuracy of 95.87% and testing accuracy of 92.33%.	Handles multiple languages in a single integrated system.	Performance varies across different individual sign language datasets.
Chutia (2024)	Emotion Recognition in Assamese Text Using LSTM	Identifying emotions within Assamese literature and text.	Unstructured text data from Assamese literature/social media.	Long Short-Term Memory (LSTM) networks.	Improved prediction and accuracy for text-based emotion analysis.	Addresses the scarcity of text-based emotion research in Assamese.	Still in its infancy regarding text-based recognition.
Kumar (2023)	Mediapipe and CNNs for Real-Time ASL Gesture Recognition	Real-time system for recognizing American Sign Language (ASL) movements.	ASL gesture dataset.	CNN for classification and MediaPipe for feature extraction.	Achieved up to 99.95% accuracy for ASL letters.	Extremely high accuracy and potential for use in communication devices.	Results are specific to ASL and may need adaptation for other languages.

Chapter 3

Thesis Completion Project Plan (12–14 Weeks)

Phase 1: Methodology & System Architecture (Weeks 1–3)

Goal: Complete Chapter 3 (Proposed Approach).

- **Week 1: Data Finalization.** Finalize the collection of the Assamese Sign Language dataset (the 2,094 data points mentioned in your review). Perform data augmentation (rotation, flipping, brightness) to increase model robustness.
- **Week 2: Preprocessing Pipeline.** Implement the **MediaPipe** landmark extraction scripts. Convert raw video/images into coordinate-based sequences (21 hand landmarks).
- **Week 3: Model Design.** Define the hybrid **CNN-LSTM** architecture.
 - Draft Chapter 3: Explain the mathematical flow, the layers of the CNN (for spatial features), and the LSTM (for temporal movement).

Phase 2: Implementation & Training (Weeks 4-7)

Objective: Constructing the Engine for Chapter 4.

- **Week 4: Environment Setup.** Install Python, TensorFlow/PyTorch, and configure GPU Support.
- **Week 5: Training.** Train the neural network on the Assamese dataset. Tune the learning rate, dropout layers, and batch size.
- **Week 6: Speech Synthesis Integration.** Incorporate the Text-to-Speech (TTS) module (gTTS/pyttsx3) such that detected gestures will activate speech output.
- **Week 7: Prototype Development.** Implement a basic GUI or OpenCV application that acquires camera input and outputs the detected Assamese sign and audio.

Phase 3: Testing & Results Analysis (Weeks 8-10)

Objective: Finalization of Chapter 4 (Experimental Results).

- **Week 8: Evaluation Metrics.** Compute Accuracy, Precision, Recall, and F1 Score. Plot confusion matrices to determine which Assamese gestures confuse the AI.
- **Week 9: Comparative Study.** Compare CNN-LSTM results against baseline models (CNN and Random Forest). This will allow you to validate your methodology.
- **Week 10: Chapter 4 Writing.** Document experimental results with graphs (Accuracy/Loss Curves) and screenshots of the live application.

Phase 4: Finalizing and Submissions (Weeks 11–12)

Goal: To finalize and submit the remaining chapters (Chapter 5), references, and format the document accordingly.

• **Week 11: Chapter 5 - Discussion of constraints (poor lighting, background noise), etc., and future goals (implementing the model in a mobile app application).**

• **Week 12: Finalize the Paper.**

- Verify whether the reference style is in order (based on the references you have provided).
- Ensure List of Figures and List of Tables are accurately paginated.
- Submit the paper for review by your advisor.

Requirements for the Term:

1. A Functioning Model: either the .h5 or .pth file capable of recognizing gestures/signs in the Assamese language.
2. The Speech Module: A working script which translates recognized sign into voice.
3. Chapters 3, 4, and 5: Transform your thesis from an overview into a completed one.
4. Thesis Paper: ready for submission to the College of Informatics, Midocean University.

Chapter 4 Proposed Approach (Methodology)

Chapter 5 :Experimental Results and Discussion

Chapter 6 Conclusion and Future Work

References

- Attili, V. S. (2024). Real-time Sign Language Translation using Computer Vision and Machine Learning. Retrieved from <https://ieeexplore.ieee.org/abstract/document/10533962>
- Bharadwaj, S. (2022). Exploring human voice prosodic features and the interaction between the excitation signal and vocal tract for Assamese speech. 26, 77–93. Retrieved from <https://link.springer.com/article/10.1007/s10772-021-09946-5>
- Bora, J. (2023). Real-time Assamese Sign Language Recognition using MediaPipe and Deep Learning. 1384-1393. Retrieved from <https://www.sciencedirect.com/science/article/pii/S1877050923001175>
- Chandrashekar, L. (2025). Towards Adaptive Real-Time Hand Sign Language Interfaces Addressing User Independence, Pose and Occlusion Variability with Machine Learning. Retrieved from https://www.researchgate.net/profile/Leena-Chandrashekar-3/publication/398822273_Towards_Adaptive_Real-Time_Hand_Sign_Language_Interfaces_Addressing_User_Independence_Pose_and_Occlusion_Variability_with_Machine_Learning/links/69442be40c98040d481eb7db/Towa
- Chetia, H. (2025). Advancing Communication: A CNN-Powered Framework for Assamese Sign Language Recognition. 5(2). Retrieved from <https://journals.dbuniversity.ac.in/ojs/index.php/AJEEE/article/view/4354>
- Choudhury, N. (2023). Enhanced Emotion Recognition from Spoken Assamese Dialect: A Machine Learning Approach with Language-Independent Features. 40(5), 2147-2160. Retrieved from <https://pdfs.semanticscholar.org/5cc5/e912815c6a930e9888d02c9ffd17754e83e0.pdf>
- Chutia, T. (2024). Emotion Recognition in Assamese Text Using LSTM. Retrieved from <https://ieeexplore.ieee.org/abstract/document/10724457>
- Das, S. (2023). A deep sign language recognition system for Indian sign language. 1469–1481. Retrieved from <https://link.springer.com/article/10.1007/s00521-022-07840-y>

- Devi, M. (2023). Speech Recognition Via Machine Learning in Recording Studio. 67–88. Retrieved from https://link.springer.com/chapter/10.1007/978-981-99-1699-3_4
- Dhanjal, A. S. (2022). An automatic machine translation system for multi-lingual speech to Indian sign language. Retrieved from <https://link.springer.com/article/10.1007/s11042-021-11706-1>
- Gadagkar, A. V. (2025). Bridging the Communication Gap: An Improved Real-Time Indian Sign Language Translation with Deep Learning. Retrieved from <https://ieeexplore.ieee.org/abstract/document/10987376>
- Hussain, A. (2023). Advancements in Indian Sign Language Recognition Systems: Enhancing Communication and Accessibility for the Deaf and Hearing Impaired. 12(2). Retrieved from <https://www.ajesjournal.org/index.php/ajes/article/view/4132>
- Ingoley, S. (2025). Interpretation of Indian Sign Language to Text and Speech to Communicate with Speech and Hearing Impaired Community. 258, 1980-1992. Retrieved from <https://www.sciencedirect.com/science/article/pii/S1877050925015522>
- Kagalkar, R. M. (2025). Enhancing Communication for the Deaf: A Machine Learning Approach to Real-Time Kannada Sign Language Translation. Retrieved from <https://link.springer.com/article/10.1007/s40031-025-01279-0>
- Kumar, R. (2023). Mediapipe and CNNs for Real-Time ASL Gesture Recognition. Retrieved from <https://arxiv.org/abs/2305.05296>
- Najib, F. M. (2025). A multi-lingual sign language recognition system using machine learning. 27987–28011. Retrieved from <https://link.springer.com/article/10.1007/s11042-024-20165-3>
- Noor, T. H. (2024). Real-Time Arabic Sign Language Recognition Using a Hybrid Deep Learning Model. 24(11). Retrieved from <https://www.mdpi.com/1424-8220/24/11/3683>
- Paul, H. (2025). Sign language recognition using improved 3D convolutional neural networks. 310-334. Retrieved from <https://www.inderscienceonline.com/doi/abs/10.1504/IJSCC.2025.149364>
- Phukan, R. (2024). Parts-of-Speech Tagger in Assamese Using LSTM and Bi-LSTM. 19–31. Retrieved from https://link.springer.com/chapter/10.1007/978-981-99-9524-0_3

- Ravikumar, S. (2024). Interactive Sign Language Learning System using Deep Learning. Retrieved from <https://ieeexplore.ieee.org/abstract/document/10760668>
- Savale, V. V. (2024). ASL Glove: A Real Time Sign-to-Speech Device. Retrieved from <https://ieeexplore.ieee.org/abstract/document/10882312>
- Sharma, G. (2025). Real-Time Sign Language Recognition and Translation Using MediaPipe and Random Forests for Inclusive Communication. Retrieved from <https://ieeexplore.ieee.org/abstract/document/10932602>
- Spandan Dey. (2021). An Overview of Indian Spoken Language Recognition from Machine Learning Perspective. 21(6), 1 - 45. Retrieved from <https://dl.acm.org/doi/full/10.1145/3523179>