

**Collage of Informatics**

**Midocean University**

**UAE**



**Research Title**

Speech Recognition System Using Deep Learning Techniques for Arabic  
Sign Language

Prepared By

Eng. Sultan Ali Senan

ID No.: 253002563

Supervisor Name: Dr. Amjad Qatish

Academic year:

2026

**1. ....**

## 2. Chapter 2 Related Work

### 2.1 Assamese Sign Language Recognition

Unquestionably, communication is difficult for those who are deaf and mute. They use sign language to interact with others and with each other. However, a sizable portion of the population that communicates verbally and in writing is not familiar with sign language. Thus, the development of technical instruments for sign language interpretation is obviously necessary. For most of the world's languages, numerous studies have been done to use technology to recognize sign language. However, there is still need for the development of methods and instruments tailored to local dialects of sign language. In addition to over 19,000 languages spoken locally as mother tongues, India is home to 22 modern languages. The goal of this project is to provide a technical method for identifying Assamese Sign Language, one of India's 22 contemporary languages. This effort aimed at developing a system for recognizing hand movements in Assamese Sign Language using machine learning techniques. The dataset was prepared using both two-dimensional and three-dimensional photographs of Assamese gestures. To find features in the photos, the Media Pipe framework was used. Nine static motions with vowel and consonant sounds were included in a dataset of 2094 data points for Assamese Sign Language (অ, ই, ঈ, উ, এ, ও, য়, ঙ, ¹). A feedback forward neural network was trained using this dataset, and the model's accuracy was 99%. The findings show that the approach used in this study is successful in identifying letters and other Assamese Sign Language motions. The recognition of signs and gestures in numerous additional Indian regional languages can also be evaluated and implemented using this technique (Bora, 2023).

According to (Chetia, 2025) a sophisticated deep learning approach—specifically, Convolutional Neural Networks (CNNs) for Assamese Sign Language (ASL) recognition is presented. The three primary stages of the suggested framework are real-time gesture recognition, model construction, and data preparation. The OpenCV frame capture library and Media Pipe hand feature recognition were used to create a dataset of ten static motions that represented various Assamese alphabet letters. The hand gesture photos were separated into training and test sets after the data was preprocessed. The CNN model was then trained to categorize the motions, attaining remarkable accuracy, recall, and F1 scores in all ten categories in addition to 100% flawless correctness. The

outcomes show how well the CNN model extracts and learns the essential characteristics of Assamese Sign Language motions, allowing the system to precisely identify and translate signs into text in real time. Communication between the hearing and deaf people in Assam could be greatly improved by this real-time recognition capabilities. Future research will concentrate on enhancing the system for real-time deployment in mobile applications and broadening the dataset to incorporate dynamic gestures.

## 2.2 Deep Learning Techniques for Sign Language Recognition

Communicating with hearing people is extremely difficult for deaf people. The shortage of qualified interpreters makes hiring a human interpreter—someone who converts sign language into a language that the hearing person can understand—an ineffective solution. Consequently, the best and most effective solution is a Sign Language Recognition System (SLRS), which lowers the communication barrier between hearing and deaf people by automatically translating sign language into text or speech without the need for an interpreter. The use of computer vision technologies for Indian Sign Language word recognition is covered in this research. Excessive preprocessing, such as collecting characteristics from a wide sequence of frames, causes inefficiencies in the current computer vision-based methods for ISL word recognition. To address the issue of excessive preprocessing, present the CNN-BiLSTM Hybrid Language Recognition System (HCBSLR), a novel computer vision-based ISL recognition system. The suggested model eliminates superfluous or redundant frames using a keyframe extraction technique based on histogram difference (HD), which increases system accuracy and efficiency. The HCBSLR system extracts spatial features using the VGG-19 algorithm and temporal features using a Bidirectional Long Term Memory Network (BiLSTM). Comparing the suggested system to some current SLRS systems, the average accuracy was 87.67%. The suggested HCBSLR system outperforms the current SLRS systems in terms of accuracy and efficiency, according to experimental results (Das, 2023).

To facilitate communication for those who are deaf or have hearing difficulties, sign language recognition is essential. Deep learning approaches can improve the work of sign language identification, which entails identifying signs and converting them into natural language. Three-dimensional convolutional neural networks (3D CNNs) are used in the suggested method to extract

features. This technique can enhance real-time sign language recognition performance and accuracy. This pilot study highlighted the benefits and drawbacks of using a 3D CNN architecture with LSTM for feature extraction in sign language recognition systems in comparison to 3D CNN, 2D CNN, and 3D CNN-integrated models. Accurate sign identification depends on the 3D CNN-LSTM model's ability to efficiently grasp the spatial and temporal characteristics of sign language expressions, which is superior to typical 3D CNN architectures. Additionally, this article discusses some methods for improving the 3D CNN-LSTM architecture's performance. Lastly, a few unresolved issues and potential avenues for further study in this area are noted. The findings of the analysis show that the 3D CNN-LSTM design has a great deal of promise to improve the accessibility of SLR systems and make communication easier for those who use sign language (Paul, 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. This produced 4,000 images of Arabic Sign Language (ArSL): 500 video clips of 10 dynamic gesture words and 10 static gesture words. The convolutional neural network (CNN) and long-range neural network (LSTM) classifiers in their suggested hybrid model show encouraging results, with accuracy rates of 94.40% and 82.70%, respectively. These findings suggest that their strategy can greatly increase the accessibility of communication for Saudi Arabia's hearing-impaired population. As a result, this study is a significant step in advancing inclusion and raising the standard of living for those who are hard of hearing (Noor, 2024).

The deaf community in India uses Indian Sign Language , a visual-gestural communication technique. It is a powerful communication tool that makes use of body language, eye movements,

and gestures. The grammatical rules of ISL are different from those of spoken language. In comparison to the general population, this study provides a way to overcome the communication gap that deaf and mute communities who use sign language experience. This study suggests a real-time program that can translate sign language, specifically ISL, into text. To help non-native English speakers, the final text will be provided in Hindi and Kannada. Advanced machine learning and computer vision technologies are used in this application. To extract features, pre-input video clips were processed using the Media pipe platform. To find the optimal model for the problem at hand, three models were used: a convolutional neural network (CNN), a recurrent neural network (LSTM), and Transformers encoders. The LSTM and Transformers models were integrated to produce a single model as the findings showed that their accuracy levels were comparable, ranging from 95 to 97%. This text is subsequently transformed into cohesive Hindi/Kannada phrases by passing the expected word sequence through a translation mechanism. All things considered, this study provides a quicker and less costly means of closing the current communication gap (Attili, 2024).

### **2.3 Speech Recognition and Speech Synthesis Systems**

The best communication method for those who are hard of hearing is sign language. There is still a communication gap between the hearing and hearing communities in spite of technological improvements. By creating an automated system that uses avatars to translate speech into Hindi Sign Language, this research seeks to close this gap (SISLA). There are three stages to the system's operation: (a) Phase 1: In a speaker-independent setting, speech recognition of single words in Punjabi, Hindi, and English. (b) Phase 2: Hindi Sign Language translation from the original language. (c) Using a 3D avatar built on the HamNoSys technology to represent motions in Hindi Sign Language. Requirements analysis, data collecting, technological development, and evaluation are the four primary SISLA system implementation modules. The efficiency of the system is enhanced by its multilingual capacity. The speech recognition models were trained and tested using speech sample files in Punjabi (12600, 4193), Hindi (12610, 4211), and English (12660, 4218). The suggested trained models obtained accuracy of at least 91%, 89%, and 89% for English, Punjabi, and Hindi, respectively, according to experimental machine translation findings. In order to assess the sign error rate through feedback, sign language specialists were

also consulted. It was suggested that the system be improved in the future by combining sentence-level translation with non-manual sign language elements. Based on survey findings, usability tests verify that the suggested SISLA system is appropriate for communication and education for individuals with hearing impairments (Dhanjal, 2022).

A basic human need is the ability to express oneself. Speech is usually used by people to communicate their ideas. Deaf and hard-to-hearing people, on the other hand, use sign language to communicate their emotions. They make gestures, and understanding sign language is necessary to interpret these signals. One of the biggest barriers to communicating is that not everyone is conversant with sign language. (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) modified two pre-trained models, Vgg16 and MobileNetV2, to speed up training. The gtts library is used by the text-to-speech module. With a verification accuracy of 99.44% for the MobileNetV2 model and 99.76% for the Vgg16 model, the system performs exceptionally well. Additionally, VGG16 and MobileNetv2 had validation losses of 0.9471 and 1.9921, respectively (Ingoley, 2025).

Computer-aided services have long relied heavily on speech recognition from audio signals. To extract information from signals, automatic speech recognition has used a variety of techniques, such as various classification and speech analysis methods. In this study, (Devi, 2023) introduced a new approach to automatically recognize speakers and their reaction tones from continuous

speech. (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 Gaussian Mixture Model mean vectors were used to build sets of equations that yielded preliminary estimations of voice power and noise spectra. This initial step involved determining the type of noise source and assessing the signal-to-noise ratio input. The first estimate of the noise power spectrum was then used to build a chosen noise model. To reduce background noise and enhance the quality of distorted speech, a Wiener filter was then used to the improved estimate. (Devi, 2023) assessed the categorization of the augmented speech data and the synthetic data under the assumption that the uncertainty parameters were known. To determine three frequent human emotional states, they last performed a data aggregation method employing a Gaussian distribution to identify three groups of high, medium, and low pitch voices (Devi, 2023).

An important topic of study in the age of multilingual, voice-driven human-computer interaction is automatic speech recognition (ASR). Many voice applications in multilingual situations operate better thanks to ASR. Most people in India, a densely populated nation with a wide variety of cultures and languages, must communicate verbally with machines using their local tongues. To the best of their knowledge, there haven't been many attempts to provide a thorough evaluation of the substantial advancements in this field of study. (Spandan Dey, 2021) made one of the first attempts to present a thorough overview of the topic of Hindi ASR research in this article. To illustrate the difficulties of scarce resources and the interaction of variables impeding the advancement of speech recognition systems in the Indian setting, a thorough analysis is provided. A thorough description of the available voice datasets, significant research contributions ranging from earlier attempts based on statistical modeling to contemporary approaches based on various neural network architectures, and future research directions are just a few of the important aspects of voice identification research in India that are covered in this study. For any researcher actively working in or interested in relevant subjects, this study will assist in evaluating the present status of speech identification research in India (Spandan Dey, 2021).

The emotional substance of a speaker's viewpoint is communicated through speech. By examining the voice signal in each sentence segment, emotions can be retrieved. Important markers of emotion and tension in a spoken communication are intonation and emphasis. It's crucial to remember that the variety of speech signals makes it difficult to extract emotions. Stress plays a

major role in tonal languages like Assamese. Even a small shift in tone can give the same expression a different meaning. Stress is used by people to encode and decipher signals. (Bharadwaj, 2022) can monitor important emotional differences by deriving correlations between stress and stress factors, such as the fundamental frequency (F0), rate, duration, and severity. fluctuations in glottal closure moments are also caused by fluctuations in stress, which in turn reflect changes in emotions. The crucial excitation moments of the vocal apparatus with the abrupt closure of the glottis are referred to as GCI, or epoch. Both the arousal level of the speech signal and the tone of voice are influenced by emotions. The purpose of this essay is to comprehend how emotions affect tone, the relationships between their traits, and how this affects arousal strength (epoch). To do this, speech analysis software like PRAAT, Python, and MATLAB are used to calculate tone characteristics. ANOVA is then used to statistically analyze these tone features to ascertain their correlations and to detect tone shifts based on various emotions. GCI signals are finally taken out of the stimulus's source, and MATLAB is used to examine how they change the same phrase in various circumstances. Human-machine interaction frameworks for developing emotion recognition in speech, especially in the Assamese language, for real-time applications like mental health treatment will be improved by conducting a thorough investigation of the emotional content of speech through tone characteristics and the source of the stimulus (Bharadwaj, 2022).

## **2.4 Real-Time Sign Language Translation Systems**

An advanced program called the Interactive Sign Language Learning System was created to make learning sign language easier. Among the many important characteristics of this all-inclusive system are built-in audio output, text-to-gesture conversion, multilingual support, and the ability to recognize letters and words in sign language. To precisely interpret hand gestures and movements, the system makes use of sophisticated sign language recognition algorithms as well as methods like image processing and machine learning. A variety of computer vision methods, such as recurrent neural networks (RNNs) or convolutional neural networks (CNNs), can be used to evaluate and categorize input images or videos into the proper sign language words or symbols for letter and word identification. In text-to-gesture conversion, natural language processing (NLP) algorithms are used to analyze the meaning of the input text and produce appropriate sign language

representations by correlating it with relevant gestures or movements. The efficiency of the algorithms employed, and the caliber of the training data determine the system's accuracy, which is determined by how many sign language symbols or words are successfully identified. The system seeks to attain high levels of accuracy in text-to-gesture conversion and sign language recognition tasks through meticulous development and training (Ravikumar, 2024).

The main form of communication for India's deaf and hard-of-hearing population is Indian Sign Language . However, because ISL is not well known, communication hurdles still exist. To facilitate communication between ISL users and non-users, this work introduces a real-time gesture recognition system that converts ISL motions into text and speech. Convolutional neural networks (CNNs) are used by the system to extract spatial features from sequential gesture data, while long-term memory networks (LSTMs) are used to record temporal patterns. The accuracy of the model is improved by a varied dataset that includes differences in hand gestures, face expressions, and body position. Continuous performance optimization is made possible by the incorporation of deep learning algorithms. By addressing the communication gap in public areas, educational institutions, and workplaces, this method fosters inclusivity. For the deaf and hard-of-hearing community, the suggested method provides a scalable and affordable means of enhancing accessibility and social inclusion (Gadagkar, 2025).

Human-computer interaction (HCI) has become an essential component of navigating and communicating in the digital world because of technological advancements. As a natural and user-friendly communication interface, hand gesture recognition has attracted a lot of interest. The design and development of a real-time hand gesture detection system to enable HCI is described in this work, with an emphasis on helping the hearing and speech challenged. The suggested method uses a real-time video camera to record hand movements and creates a reliable, personalized dataset that takes obfuscation, posture, and human differences into account. For every hand gesture, 21 important features are extracted using a convolutional neural network (CNN). A random forest approach is then used to classify these features, yielding an overall accuracy of 94.58% over several cases. Effective and comfortable communication is made possible by the translation of the identified motions into text and speech. This technique makes it possible to combine several gestures to create whole sentences, which are commonly utilized in social situations. High resilience to illumination changes was demonstrated by performance tests

conducted under various lighting circumstances, which revealed a signal-to-noise ratio ranging from 3 to 4.27 dB. Easy two-way communication is made possible by the graphical user interface with feedback, which improves accessibility and usability (Chandrashekar, 2025).

## **2.5 Machine Learning and Computer Vision Approaches for Sign Language Recognition**

Deaf and hard-of-hearing people utilize sign language, which is a visual-kinetic language. It uses a variety of body postures, face expressions, hand gestures, and movements. Because of its distinct grammatical and morphological norms, sign language is an essential mode of communication. One of the biggest challenges is the communication gap between sign language users and non-users. In India, there are over 63 million persons who are deaf or hard of hearing. Therefore, research in sign language recognition demonstrates enormous potential for improving the quality of life of individuals with hearing impairments and promoting communication and social inclusion. Due to its critical role in human-machine interaction, accessibility, interpretation, educational tools, and assistive communication, sign language recognition has attracted a lot of interest in recent years. This paper examines the most recent advancements in Indian sign language recognition systems, going into detail about standard datasets, popular methods, and system performance features. Lastly, it draws attention to the difficulties and potential of these new technologies (Hussain, 2023).

(Kumar, 2023) described a real-time system that uses cutting-edge computer vision and machine learning technology to recognize movements in American Sign Language (ASL). The suggested approach uses a convolutional neural network (CNN) to classify ASL motions using the Media pipe library to extract features. According to test results, the suggested method can identify every ASL letter with up to 99.95% accuracy, suggesting that it might be used in communication devices for those who are hard of hearing. This technique can also be used with other sign languages that use comparable hand movements, which will help those who are hard of hearing live better lives. Overall, this study makes a significant contribution to the fields of computer vision and machine learning by demonstrating the efficacy of utilizing Media pipe and CNN for real-time sign language detection.

Recently, there has been a lot of interest in the topic of automatic sign language recognition. The main use of machine learning techniques is the recognition of hand movements. Most of the current research uses a sign language unique to a particular nation, such American Sign Language, to train machine learning models. The Multilingual Sign Language Interpreter (MSLI) System is a machine learning-based multilingual sign language system that (Najib, 2025) presented in this research. MSLI uses hand gestures in several languages to train a machine learning model. The supplied sign language and its classifications can be recognized by the system. The suggested system offers a two-stage recognition method when test signals from the same language are input. It initially detects only the first sign language and then tests the other signals in accordance with the recognized language. Additionally, MSLI offers distinct categories for each language's signs. Eleven datasets in various languages were used for the experiments. The input data was classified both separately and in combination. The accuracy of the suggested system is shown by the experimental results. On the majority of the various sign language datasets that were used, the suggested system's training accuracy ranged from roughly 90% to 100%. For the individual datasets, most of the test data had categorization accuracy levels above 90%. Training and testing accuracy for the suggested MSLI integrated classification system were 95.87% and 92.33%, respectively (Najib, 2025).

A major advancement in assistive technology, the American Sign Language (ASL) Glove provides a way to improve communication for people with speech, hearing, or vision problems. The creation and assessment of the ASL Glove, a real-time sign-to-speech tool intended to promote smooth communication between users and recipients, are described in this research study. The gadget converts ASL movements into text and subsequently into synthesized speech using flexible sensors and an Arduino Uno microprocessor built into a specially made glove. With its text display interface for visually interpreted signs and audio output for auditory communication, the ASL Glove ensures inclusivity for people with a range of communication needs. The device's efficacy as an assistive technology tool was confirmed by extensive testing and user evaluation sessions, which revealed great accuracy in gesture recognition and real-time interpretation. The ASL Glove enables people with disabilities to engage in social interactions and everyday activities more successfully by removing communication barriers. Further advancements in communication access and increased social inclusion are anticipated as the gadget continues to be developed and improved (Savale, 2024).

For deaf and hard-of-hearing people to effectively engage with the hearing population, creative solutions to their communication challenges are needed. Using Media Pipe and random forest technology, this study suggests a revolutionary method for translating and recognizing sign language. Media Pipe makes it easier to track hand gestures in real time and extract important hand characteristics for deciphering sign language. The signals represented by the tracked gestures are then precisely identified using a random forest classifier. The system improves communication accessibility for the deaf and hard-of-hearing community by smoothly translating sign language into written words and text after successful detection. Combining cutting-edge machine learning methods with real-time tracking technology presents encouraging opportunities to reduce barriers to communication and enhance inclusivity (Sharma, 2025).

## **2.6 Assamese Language Processing and Related Studies**

Using speech and signal processing techniques, this study investigates a novel field of automated emotion recognition in Assamese, a widely spoken but little-studied dialect of Northeast India. Sophisticated supervised machine learning methods are utilized to classify a wide range of basic and derived emotions from spoken Assamese using a special combination of language-independent data. For speech-based emotion recognition, a novel combination of impersonal and personal features is suggested. The Fuzzy C-means clustering technique is used to first split high-dimensional features. Emotion recognition and classification are thus made easier using several random forest classifiers. With nested transformations, special attention is given to several emotions that usually present difficulties for classification systems. K-fold cross-validation is used for performance analysis, and the findings are verified by comparing them to simple random forest techniques and validating several important criteria. Interestingly, the suggested approach outperformed baseline models, increasing accuracy by roughly 4.26% to 4.50%, demonstrating effective classification without the use of language data. Additionally, this research advances the creation of a useful, relevant emotion identification technology. Medical professionals may find use for this instrument, which would allow them to effectively assess a patient's mental condition and help create more potent treatment plans. This research has broad ramifications, especially for understudied languages, and highlights the need to broaden the scope of emotion recognition studies to include a greater variety of languages (Choudhury, 2023).

One of the most difficult aspects of natural language processing (NLP) is speech particle classification (POS). The goal of this study is to create a POS classifier for Assamese. Assamese lacks high-performance POS classifiers because digital linguistic resources are scarce. To close this gap, the proposed study investigates the development of a POS classifier using an Assamese POS dataset by utilizing Long Term Memory (LSTM) and Bi-LSTM (Bi-LSTM). It should be mentioned that, as was to be expected, this experiment had trouble comprehending and handling natural language for computational linguistics purposes. About 50,000 words make up the Assamese dataset used in this study. During the first stage, the classifiers produced satisfactory results while analyzing the first datasets, which had roughly 20,000 words. These results led to the expansion of the Assamese dataset to 50,000 words, and a notable improvement in accuracy, precision, recall, and the F1 metric was noted. Consequently, the accuracy of Bi-LSTM was 91.72% while the accuracy of LSTM was 91.20%. Considering the significance of natural language processing research for the Assamese language, a comparison between the proposed study and current Assamese speech classifiers was provided (Phukan, 2024).

A vital component of all human connections is emotion. Body language, facial expressions, conversation, and written text can all convey emotions. In many areas, such as business, management, healthcare, and others, emotion recognition is a crucial component of natural language processing and a key technology in human-computer interaction. Massive volumes of unstructured data have been produced on social media platforms in recent years due to the growing usage of digital marketing and social media. As a result, it is difficult to identify emotions in text. Researchers have used a variety of machine learning and deep learning approaches to address these issues. LSTM (Long Term Memory) and other deep learning techniques are quite good at identifying emotions in text. In this work, (Chutia, 2024) used an 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).

## **2.7 Research Gap**

The material covered in this chapter exhibits considerable breakthroughs in sign language recognition, voice recognition, and deep learning applications. Several researchers have

successfully utilized Convolutional Neural Networks (CNNs), Long Short-Term Memory Networks (LSTMs), Bidirectional LSTMs, Transformer models, Random Forest classifiers, and Media Pipe-based feature extraction techniques to improve recognition accuracy and real-time performance. Most studies revealed excellent accuracy rates surpassing 90%, indicating the usefulness of these techniques in sign language identification tasks.

Despite these advancements, the bulk of extant research concentrate on commonly used sign languages such as American Sign Language (ASL) and Indian Sign Language (ISL). Comparatively, Arabic Sign Language has received limited scholarly attention, with relatively few studies especially addressing its real-time identification and translation. Furthermore, the availability of large-scale, well-organized Arabic Sign Language datasets remains restricted, providing obstacles for constructing robust and universal recognition algorithms.

In addition, several contemporary systems specialize in either gesture recognition or voice creation individually. Only a small amount of research offers a full end-to-end framework that blends Arabic Sign Language detection, deep learning-based categorization, and real-time voice synthesis into a single system. Challenges relating to dynamic gesture detection, environmental fluctuations, user diversity, and real-time deployment all remain poorly addressed in the available research.

Therefore, there is a need for a specific deep learning-based voice recognition system for Arabic Sign Language that integrates computer vision, deep learning classification, and speech synthesis in an integrated real-time framework. The proposed study intends to solve this gap by building a system capable of reliably identifying Arabic sign motions and translating them directly into speech, hence boosting communication accessibility for the deaf and hard-of-hearing population.