Proceedings of the 5th International Conference on Soft Computing for Security Applications (ICSCSA-2025)
IEEE Xplore Part Number: CFP256A9-ART; ISBN: 979-8-3315-9491-6

2025 5th International Conference on Soft Computing for Security Applications (ICSCSA) | 979-8-3315-9491-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICSCSA66339.2025.11170908

Speech-to-Sign Language Translation Framework for Telugu Language using Residual Bidirectional LSTM with Attention Network

[1*]B.K.V.P.S. Mahalakshmi
[1]*Department of Computer Science and Engineering*
GITAM School of Technology
Gitam Deemed To Be University
Visakhapatnam, Andhra Pradesh 530045, India
mbolired@gitam.in

[2]S. Anuradha
[1]*Department of Computer Science and Engineering*
GITAM School of Technology
Gitam Deemed To Be University
Visakhapatnam, Andhra Pradesh 530045, India
asesetti@gitam.edu

***Abstract*- The Speech-to-Sign Language Translation (S2SLT) system for the Telugu language is developed to improve communication between hearing individuals and the hearing-impaired community. Telugu, a widely spoken language in India, currently lacks adequate support in assistive technologies, especially for translating speech into Indian Sign Language (ISL). This work aims to build an intelligent system that captures spoken Telugu, processes it using Automatic Speech Recognition (ASR), and translates the recognized text into ISL using Natural Language Processing (NLP) techniques. Special attention is given to preserving the meaning and context during translation, as Telugu and ISL differ significantly in grammar and sentence structure. The paper proposes the design and implementation of a model that translates live voice or audio recordings of a native Indian regional language (Telugu) into Indian Sign Language (ISL) animations. In the initial phase, essential audio signals are collected from benchmark resources. From the collected audio, dual features such as spectral features and Mel-Frequency Cepstral Coefficients (MFCC) are extracted to represent the speech data effectively. These extracted features are passed into the speech-to-sign translation module. Here, a Residual Bidirectional Long Short-Term Memory with Attention Mechanism (Res-BiLSTM-AM) is employed to detect the corresponding sign text. The predicted sign text is then matched with a pre-defined Graphics Interchange Format (GIF) dataset containing ISL animation files. If a match is found, the corresponding GIF is generated as the final output. In cases where the predicted text does not match any entry in the GIF dataset, a Graphical User Interface (GUI) is used to display the ISL text output alphabetically, ensuring continued accessibility.**

Keywords-Speech-to-Sign Language Translation; Telugu Language; Residual Bidirectional Long-Short-Term Memory with Attention Mechanism; Graphics Interchange Format; Indian Sign Language

I. INTRODUCTION

Background of the study

People with exceptional needs have always benefited from the use of technology [1]. There are 466 billion visually impaired persons worldwide, or more than 5% of the entire population, and it is predicted that by 2050, there will be 900 million of them [2]. Sign language is an exclusive set of movements used by deaf people to interact with one another [3]. A gesture is a type of body language that includes motions, hand forms, position, and some non-manual characteristics [4]. The majority of the non-manual characteristics are signs of emotion, such as smiling, blinking, raising eyebrows and motions of the tongue and lips [5]. Deaf people cannot read and comprehend material produced in natural languages because sign language phrases vary from a language with written phrases. This makes it difficult for deaf people to profit from any spoken content due to mental constraints.

Every nation has a unique language of signs, and there is no universal language for signs [6]. Furthermore, within a nation, there are several regional sign languages. By creating new and enhancing current assistive technology for those with impairments, especially the deaf population, wealthy countries have been attempting to build inclusive societies [7]. The creation of communication aids and techniques for American Sign Language (ASL), British Sign Language (BSL), and multiple additional languages have been the subject of much research. But it's crucial to keep in mind that nearly all of deaf people reside in nations that are developing [8]. Recently, some developing nations have also started to concentrate on enhancing accessibility to help their deaf citizens. Excellent studies regarding Indian Sign Language (ISL), Arabic Sign Language (ArSL), and some additional sign languages have been produced.

Translating the voice of a speaker or sounds into visual motions is known as S2SLT, and it helps close up the divide between the hearing handicapped and the general public. To guarantee visual clarity, S2SLT often requires an internal video GIF library. Because movements constitute a distinct universal language that has a structure that is different from any speech, translating gesture-based interaction requires a lot of work on the part of the translation [9]. Applications such as emotional voice transformation, which creates video GIFs with an emotive voice teaching sign language to regular people, may be built on top of S2SLT [10]. There aren't many feasible algorithms to translate text to ISL, even though sign communication is used all over the world to help the hearing impaired communicate in daily life. There is a dearth of suitable and efficient audio-visual support for spoken communication. ISL computerization has received barely any attention, despite notable advancements in computer recognition of international language signs. Therefore, all of the difficulties mentioned previously demonstrate how urgent it is to create a structure for translating speech into sign communication.

Contribution over existing works

Authorized licensed use limited to: Vidya Vardhaka College Of Engineering Mysore. Downloaded on November 11,2025 at 06:39:55 UTC from IEEE Xplore. Restrictions apply.

For improving the performance in speech-to-sign language translation models, several studies have employed the Natural Language Processing (NLP) model for enabling contextual understanding and effectively handle the variations [8]. However, the NLP model faces difficulties, while handling the large and diverse languages. Data scarcity is considered as one of the major issues due to insufficient data in the training/testing process [11]. It happens in the imbalanced datasets, whereas some classes are biased towards the majority class. Thus, it limits the generalization performance of the unseen data points, making the model ineffective to get reliable outcomes. In addition to this, implementing the machine learning and deep learning techniques in recent times has become a hot research topic, whereas it is well-suited in various real-time applications [12]. However, it is limited in the quality of translation in the underlying models and provides highly labour-intensive. Validating with a larger annotated dataset is a tedious process and is affected by capturing the relevant features in various domains. Thus, a novel deep learning model is implemented to provide reliable communication of deaf people.

Proposed Work: Thus, this work developed an innovative S2SLT model using deep learning, and the significant contributions of the developed model are shown below.

- To propose an innovative end-to-end S2SLT model that leverages deep learning techniques to convert spoken Telugu, a native Indian regional language, into ISL animations. The model effectively processes audio signals, extracting crucial speech features such as spectral features and MFCC, which serve as representations of the speech data. By using these features, the system translates audio input into text and then matches it with corresponding ISL animations. This development is significant as it bridges the communication gap between hearing and hearing-impaired individuals, specifically in the context of regional languages like Telugu, which are often underrepresented in existing language models.
- To develop Res-BiLSTM-AM to detect the corresponding sign text from the input speech. The Res-BiLSTM-AM model combines the power of Bidirectional LSTMs, which allow the network to capture both past and future contexts in the speech sequence, with residual connections that help the model train deeper architectures without suffering from the vanishing gradient problem. The Attention Mechanism further improves the model's focus on significant features in the input speech sequence, making the translation process more accurate. This hybrid approach results in improved translation quality, better handling of complex sentences, and more accurate detection of sign language text, even for noisy or ambiguous inputs.
- To prove its effectiveness, once the corresponding sign text is detected, the system matches it with a pre-defined GIF dataset that contains ISL animation files. If a match is found, the system generates the appropriate ISL animation and displays it as the final output. This integration of a GIF-based animation dataset enhances the system's ability to provide dynamic and accurate visual translations. In scenarios where the predicted text does not match any entry in the GIF dataset, the system employs a GUI that displays the ISL alphabetically, ensuring that users still have access to the translation in text form. This fallback mechanism guarantees accessibility for users even in cases of missing or incomplete dataset coverage, making the system more versatile and inclusive.

This section provides the model's organization. A literature survey is provided in Section II, the importance of S2SLT to develop an effective model using the audio signal is given in Section III, an overview of the developed S2SLT model with the assistance of the feature extraction process is shown in Section IV, deep learning model with attention mechanism for text to sign language translation by matching the animation file is offered in Section V, results and discussion is illustrated in Section VI, and the conclusion is provided in Section VII.

II. Literature Survey

Based on prior works, the literature survey is elucidated in this section by analyzing deep learning, machine learning and other techniques. Moreover, the section describes the analysis of research gaps in prior works, which is presented in the below sub-sections. This analysis helps the researchers can implement the novel framework based on speech-to-sign language translation model.

A. Related Works

In 2022, Reddy et al. [13] suggested and deployed a model that converted real-time audio or narration of Telugu, a dialect of Hindi spoken in India, into text and then compared it to sign language cartoons using a pre-made GIF database. The Google API was used to turn the voice into text. It produced the appropriate GIF if the content and the phrases in the GIF collection agreed. With an average mistake rate of 4.1% and recall and accuracy scores of 94% each, Google's API-based voice-to-sign language transformation system has shown the most promise and consistency when compared to the other algorithms.

In 2020, Khan *et al.* [14] developed a revolutionary automated translation system that converts English statements into similar Pakistan Sign Language (PSL) utterances, utilizing NLP approaches to assist the deaf community. To convert English phrases into similar PSL phrases, a grammar-based automated translation method was proposed. A highly encouraging Bilingual Evaluation Understudy (BLEU) rating of 0.78 was found in the quantitative data. Subjective assessments showed that the tool's output represented as an easily comprehensible avatar, made up for the users' mental difficulties with hearing.

In 2021, Kapoor *et al.* [15] have developed a multitasking transformer system that has been taught to produce signers from voice patterns. Using an extra cross-modal distinction and speech-to-text as an extra task, the model was able to train to produce continuous sign posture sequences from beginning to finish. The method's efficacy was shown by several tests and compared against different precedents. To examine the impact of various network modules, they also carry out further ablation tests.

In 2024, Joy *et al.* [16] have introduced a cutting-edge means of communication system intended for those with hearing loss. In the first stage, the speech-to-text algorithms

Authorized licensed use limited to: Vidya Vardhaka College Of Engineering Mysore. Downloaded on November 11,2025 at 06:39:55 UTC from IEEE Xplore. Restrictions apply.

were used to precisely translate conversations into text. This strategy sought to continuously enhance the system's usability and functionality. In the end, this study advances assistive technology, removes obstacles to interaction, and promotes inclusion for those with hearing loss.

In 2023, Liang *et al.* [17] have proposed SinDiff, a transformer-based dispersion paradigm that investigated all of the benefits of the transformer (such as broad context and variable attention) for spoken-driven SLG. They achieved this by creating a linear biased cross-modal multi-head attention mechanism to align both spoken and sign language paradigms and a linear prejudicial multi-head self-attention with regular location coding to extrapolate longer-term gestures. According to test-bed testing findings, the SinDiff's phrase reliability for the back-translated stated translation with the actual words uttered reached 21.02%, a 2.8% increase above the standard SLG method's (18.22%) reliability on the PHOENIX14T test set.

B. Problem statement

A speech-to-sign language translation system faces several challenges despite its potential benefits. These include issues with accuracy and contextual understanding due to the complexity of both spoken and sign languages. Variations in sign languages across cultures, the difficulty in replicating non-verbal cues, and the need for real-time processing also pose significant obstacles. The features and challenges of existing techniques are provided in Table I. The research gaps of the traditional speech-to-sign language translation methods are given below,

- Traditional methods struggle to capture the contextual meanings and ambiguities present in spoken language. Deep learning models can be used to understand the contextual relationships between words in a sentence. These models can be trained on large datasets of paired spoken and signed language, enabling the system to learn context-sensitive translations.
- In certain cases, the effectiveness of the existing models is minimized due to the dimensionality of the data and they struggle to process huge datasets. Therefore, employing a feature extraction technique assists in enhancing the performance of the model by selecting the most significant features from the input data.
- Traditional models face certain challenges to understand the context as well as the effectiveness of the model is minimized due to the occurrence of vanishing gradients. These issues can be addressed by employing attention mechanisms.

TABLE I. FEATURES AND CHALLENGES OF EXISTING SPEECH-TO-SIGN LANGUAGE TRANSLATION TECHNIQUES

Author [citation]	Methodology	Features	Challenges
Reddy *et al.* [13]	Google API	• This method is capable of converting live speech or recorded audio into sign language animations in real time. • This technique remains functional even if the word isn't present in the predefined dataset.	• The accuracy and performance of the system are limited by the size and completeness of the predefined GIF dataset. • If the dataset doesn't include the necessary sign gestures, the system might not be able to provide accurate sign language translations.
Khan *et al.* [14]	Machine Translation Model	• This method ensures that the translations follow PSL's unique linguistic structure, leading to more accurate and contextually appropriate translations. • Better translation quality.	• This method struggles with complex and compound-complex sentences. • This technique might not generalize well to a wide range of linguistic variations.
Kapoor *et al.* [15]	Multi-Tasking Transformer Network	• This system eliminates the need for intermediate text-based transcriptions, directly generating sign language pose sequences from speech. • It allows the technique to effectively learn the relationship between speech and sign language, improving the quality of the generated sign language sequences.	• This method may struggle with noisy or ambiguous speech, especially when the input speech is unclear or heavily accented. • This method does not address the broader aspects of sign language, which are crucial for fully conveying meaning in sign language.
Joy *et al.* [16]	Google API	• This method facilitates better communication between individuals with hearing and speech impairments. • This method allows real-time translation from speech to sign language.	• This method may struggle to convey all aspects of spoken language, including sarcasm, idiomatic expressions, and context-specific terms. • This method has concerns about privacy and the collection of sensitive data, especially when dealing with real-time audio and video inputs.
Liang *et al.* [17]	SinDiff	• This method helps capture long-term dependencies in sign language and the connection between spoken language and	• This technique is not suitable for real-time applications due to its high time complexity.

Authorized licensed use limited to: Vidya Vardhaka College Of Engineering Mysore. Downloaded on November 11,2025 at 06:39:55 UTC from IEEE Xplore. Restrictions apply.

		sign language. • This model can capture global speech semantic features.	

III. Importance of Speech-to-Sign Language Translation to Develop an Effective Model using the Audio Signal

The section briefly explores the need for a speech-to-sign translation model is briefly discussed. Also, the effective performance is attained by focusing on the standard Kaggle dataset. Utilizing this dataset has the efficiency to strengthen the model's outcome, which makes reliable communication of deaf people.

A. Importance of Speech-to-Sign Language Translation

S2SLT plays a vital role in bridging the communication gap between the hearing and hearing-impaired communities. In a diverse and multilingual country like India, where access to inclusive communication tools is limited, S2SLT systems provide an essential solution for improving accessibility. These systems convert spoken language into visual sign language, enabling deaf and hard-of-hearing individuals to understand real-time conversations, announcements, educational content, and public services. This technology is especially important in regions where sign language interpreters are scarce or unavailable. By supporting regional languages such as Telugu, S2SLT promotes linguistic inclusivity and ensures that native language speakers are not left behind in digital communication advancements. Furthermore, the integration of speech recognition, natural language processing, and gesture animation technologies helps in developing intelligent, real-time translation tools that are both scalable and user-friendly. The development of S2SLT systems not only enhances social inclusion but also empowers individuals with hearing disabilities by enabling more independent interaction with the world around them.

B. Data Collection

The developed model used the following dataset to collect the signals.

Free Spoken Digit Dataset (FSDD): The signals are taken using the link of https://www.kaggle.com/datasets/joserzapata/free-spoken-digit-dataset-fsdd. Access date: 2025-04-26. This dataset has 6 speakers and 3k recordings (50 of every numeral per speaker). It has a total of 3012 files and is 21.15MB in size.

Moreover, the GIFs needed to perform the matching process are collected manually, which has 50 classes and 500 counts of GIFs.

The collected audio signals are indicated by the terms D_B^{Audio}, and Fig. 1 offers the collected audio signals for the Telugu text.

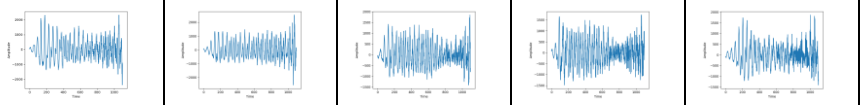
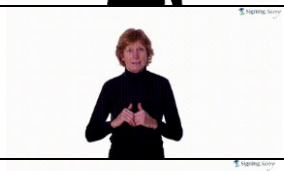
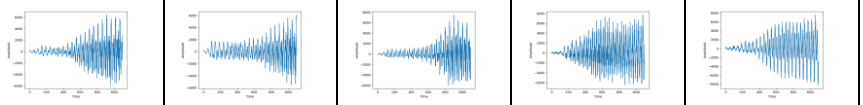
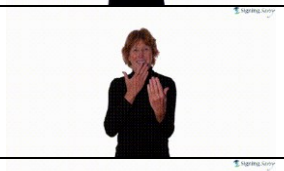
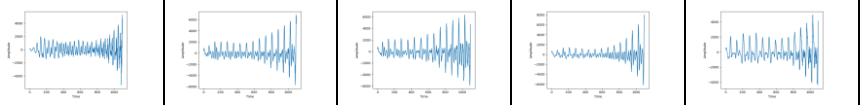
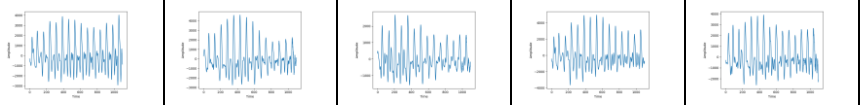
Telugu text	English Text	ISL for the Text	Audio Signals
హాయ్	Hai		
మీరు ఎలా ఉన్నారు	How are you		
నాకు సహాయం కావాలి	I need help		
క్షమించండి	Sorry		

Fig. 1. Collected Audio Signals

IV. Overview of the Developed S2SLT Model with the Assistance of Feature Extraction Process

An overview architectural details of the proposed model is discussed. Here, the process of developed model in a speech-to-sign language translation framework is done collecting the audio signals. Then, the flow process of the developed model is discussed in sub-section A. In order to get the accurate outcomes, the relevant features are required so that, the feature extraction process of the developed model also discussed in sub-section B.

Authorized licensed use limited to: Vidya Vardhaka College Of Engineering Mysore. Downloaded on November 11,2025 at 06:39:55 UTC from IEEE Xplore. Restrictions apply.

A. *Overview of the Developed S2SLT Model*

The paper presents the development and implementation of a S2SLT model aimed at translating spoken Telugu, a widely used Indian regional language, into ISL animations. The motivation behind the model is to bridge the communication gap between hearing individuals and the hearing-impaired community, particularly in linguistically diverse regions where regional languages are underrepresented in assistive technologies. The system begins by collecting live voice or pre-recorded audio inputs from benchmark resources. These audio signals undergo a feature extraction process where both spectral features and MFCC are derived. These features are critical in accurately representing the characteristics of the speech signal and serve as the input to the translation module. In the core translation module, a Res-BiLSTM-AM is utilized. This deep learning architecture effectively captures temporal dependencies in both directions of the input sequence, while the attention mechanism enhances the model's ability to focus on relevant parts of the audio data during translation. The output from the Res-BiLSTM-AM is the predicted sign-equivalent text corresponding to the input speech. Once the sign text is detected, it is mapped against a pre-defined dataset of GIF files, each representing a specific ISL sign animation. If a match is found, the system outputs the appropriate animated GIF, offering a visual and accessible representation of the spoken Telugu input. This enables hearing-impaired users to receive the spoken message in a familiar and understandable visual format. In scenarios where the recognized text does not exist in the GIF dataset, the system maintains accessibility by triggering a GUI. This GUI displays the ISL translation of the given input in alphabetical form, ensuring that users still receive a meaningful output even in the absence of direct GIF matches. This modular design ensures the system is both scalable and adaptable for future expansion, including support for additional regional languages and more extensive sign datasets. By integrating audio processing, deep learning, and animated visualization, the developed S2SLT model provides a practical and inclusive communication tool for Telugu-speaking individuals with hearing impairments, contributing significantly to the field of accessible technology. The architectural view of the developed S2SLT model is given in Fig. 2.

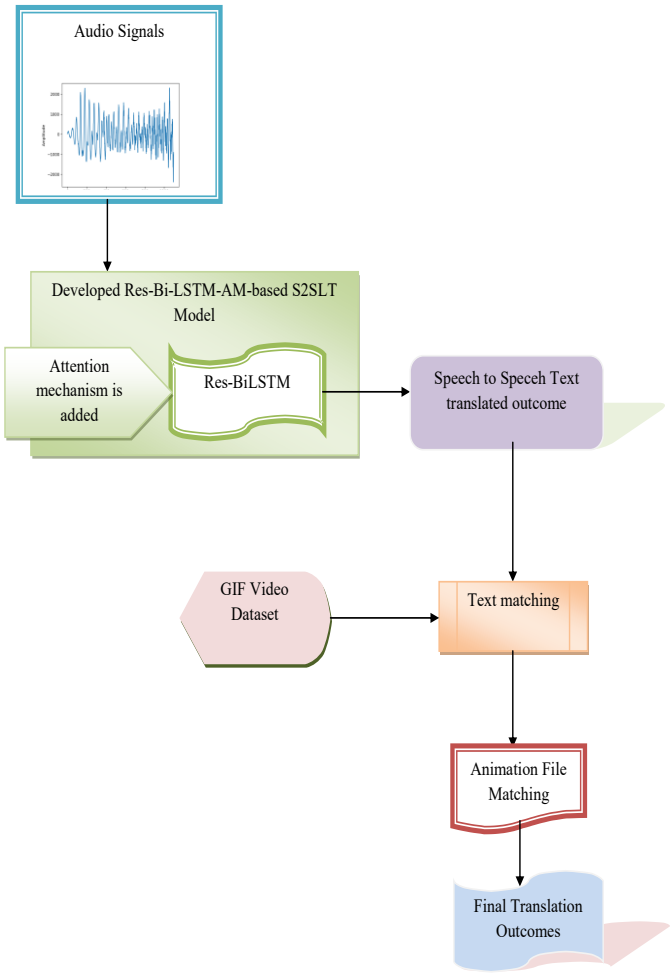


Fig. 2. Architectural View of the Developed S2SLT Model

B. *Feature Extraction Process*

The feature extraction process is a critical component in the S2SLT system, as it directly influences the accuracy and efficiency of speech recognition and translation. In the proposed model, feature extraction is applied to the raw audio input collected from live voice recordings or benchmark audio datasets in the Telugu language. Since raw audio data contains a wide range of information, it is essential to extract meaningful characteristics that best represent the phonetic and temporal properties of speech. To achieve this, the system employs a combination of spectral features and MFCC.

Initially, the collected audio signals D_B^{Audio} are given for extracting the spectral features. Spectral features [18] capture the distribution of energy across various frequency components in the audio signal. These features are useful for distinguishing between different speech sounds, as they reflect the tonal and harmonic structure of spoken language.

Here, the resultant spectral features are given as input for further extracting the MFCC. In parallel, MFCC [19] features are extracted because they closely mimic the human auditory perception system. MFCCs are derived by applying a series of transformations to the audio signal: pre-emphasis, framing, windowing, and Mel filter banks. This process converts the time-domain signal into a set of coefficients that represent the short-term power spectrum of the speech,

Authorized licensed use limited to: Vidya Vardhaka College Of Engineering Mysore. Downloaded on November 11,2025 at 06:39:55 UTC from IEEE Xplore. Restrictions apply.

allowing for robust and compact representation. The final extracted features are indicated using the term Fe_N^{Ext}.

The combination of spectral and MFCC features enhances the system's ability to handle variations in pronunciation, pitch, and accent factors, particularly important in a linguistically rich language like Telugu. These dual features are then normalized and structured into sequences suitable for input into deep learning models. By focusing on both spectral and perceptual aspects of the audio input, the feature extraction process lays a strong foundation for accurate speech-to-sign translation. It ensures that the model receives the most relevant and noise-resilient information, thereby improving the overall performance and reliability of the system. This phase not only supports robust recognition in clean conditions but also contributes to maintaining translation quality in real-world environments with background noise and speaker variability.

V. Deep Learning Model with Attention Mechanism for Text to Sign Language Translation by Matching the Animation File

For getting the final translation outcome, the Res-BiLSTM-AM model is suggested by combining the advantages of Bidirectional LSTMs, residual connections, and attention mechanisms to empower the system performance in the speech-to-sign language translation model. This developed model can handle the complex relationships and data patterns and improve the model's efficiency. Moreover, the detailed discussion for getting a reliable outcome is shown in sections below.

A. Residual Bi-LSTM

The study builds on the benefits of this model by presenting a residual Bi-LSTM [20] system. It is possible to use normalization techniques in the Bi-LSTM system. For RNN, Layer Normalization (LN) is particularly advantageous when compared to Batch Normalization (BN). LN functions in the same way as BN and may be expressed using Eq. (1).

$$\overline{C^N} = \frac{C^{(N)} - J\left(C^{(N)}\right)}{\sqrt{X\left(C^{(N)}\right)}} \tag{1}$$

Here, the input matrix is $C^{(N)}$, and it has N numbers of degree and CN's result is $\overline{C^N}$.

Residual Bi-LSTM is a novel combination of residual architecture and LN in the Bi-LSTM system. Eq. (2) to Eq. (4) are used to define the recurrence data on qualities A.

$$C_Y^{K(N+1)} = P_N\left(C_Y^{K(N)} + Q\left(C_Y^{K(N)}, A_N\right)\right) \tag{2}$$

$$C_Y^{G(N+1)} = P_N\left(C_Y^{G(N)} + Q\left(C_Y^{G(N)}, B_N\right)\right) \tag{3}$$

$$D^Y = E_C\left(C_Y^K + C_Y^G\right) \tag{4}$$

The term P_N is LN, Q refers to the input that the LSTM network is analyzing. The subscript Y in $C_Y^{K(N+1)}$ indicates the Y^{th} time step during the sequence. The forward state, backward state, and the number of processed sections are all represented by K, G, and $(N+1)$. The data being encoded D^X at the current time step Y is integrated using both the forward and backward claims, combining information from past and future inputs to enhance the network's understanding of the sequence.

B. Development of Res-BiLSTM-AM

The development of the Res-BiLSTM-AM plays a pivotal role in enhancing the accuracy and efficiency of S2SLT in the proposed model. This deep learning architecture is designed to handle the complexities inherent in the translation of spoken Telugu into ISL. The Res-BiLSTM-AM model is specifically developed to address the challenges posed by sequential and contextual dependencies in speech data, ensuring that the system generates accurate sign language outputs.

Novelty: Here, the extracted features Fe_N^{Ext} are given as input into the BiLSTM of the developed Res-BiLSTM-AM. The BiLSTM network, which can process sequential input both forward and backwards, is the central component of this approach. BiLSTMs are superior to regular LSTM models for voice recognition tasks where context is essential because they can capture temporal dependencies from both past and future inputs. Understanding the complete word or phoneme sequence is essential for producing contextually accurate translations when translating speech to sign language. However, because of problems like vanishing gradients, BiLSTMs might occasionally have trouble learning long-range dependencies, even while they are good at learning temporal relationships. The residual connection is presented as a solution to this constraint. By enabling gradients to flow directly across layers, residual connections aid in reducing the degradation issue and promote deeper and more effective learning. The model is strengthened by adding residual connections, which enhances its ability to handle a variety of input variables and complex language patterns. Additionally, the model may concentrate on those segments of the input sequence that are most pertinent to the translation task because to the integration of the Attention Mechanism. By dynamically weighing various input components, the attention mechanism makes sure the model focuses more on critical elements like certain words or phonemes that are essential for producing the right sign language. The Res-BiLSTM-AM can better manage ambiguous or noisy inputs thanks to its selective focus, which improves translation accuracy.

The model produces a series of predicted sign text that matches the spoken words or phrases after the input voice has been analyzed by the Res-BiLSTM-AM network. The related ISL signs are then represented by a pre-defined dataset of GIF animations that are mapped to this projected sign text. The relevant ISL animation is shown if a match is discovered, enabling visual communication. In cases where the predicted sign text does not match any entry in the GIF dataset, a GUI is employed to display the ISL text output alphabetically, ensuring that users can still understand the translation. This fallback mechanism ensures continued accessibility, even when the exact translation cannot be visualized through existing animations. Thus, the Res-BiLSTM-AM model leverages the strengths of Bidirectional LSTMs, residual connections, and attention mechanisms to create an advanced system for speech-to-sign language translation. By combining these powerful techniques, the model is able to process speech data effectively, handle

Authorized licensed use limited to: Vidya Vardhaka College Of Engineering Mysore. Downloaded on November 11,2025 at 06:39:55 UTC from IEEE Xplore. Restrictions apply.

temporal dependencies, reduce training difficulties, and focus on the most important aspects of the input signal. The resulting system is robust, efficient, and capable of delivering accurate translations of spoken Telugu into ISL animations, significantly improving communication accessibility for the hearing-impaired community. The structural illustration of the developed Res-BiLSTM-AM-based S2SLT model is given in Fig. 3.

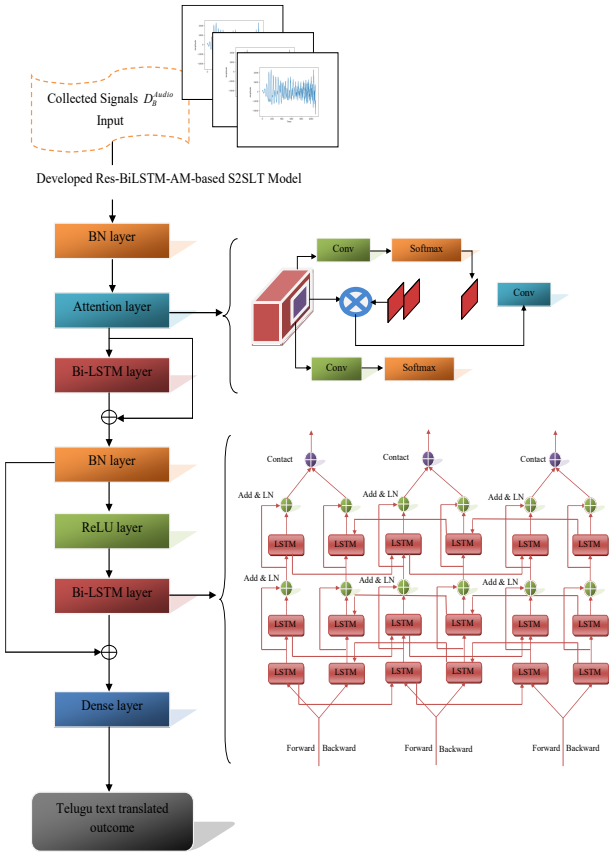


Fig. 3. Structural Illustration of the Developed Res-BiLSTM-AM-based S2SLT Model

C. *Text-to-Sign Language Translation Steps with Text Matching*

The S2SLT system is designed to convert spoken or written text into a visual representation in sign language. The process involves several crucial stages, each focused on transforming text into an easily understandable form for the hearing-impaired community. Below is a detailed breakdown of the steps involved in the S2SLT pipeline, with special attention to the matching of the animation file and final translation outcomes:

1. Sign Language Translation (Text to Sign)

The next step is the translation of the processed text into ISL. This is achieved by using a specialized deep learning model, such as the Res-BiLSTM-AM, which predicts the corresponding sign text for the inputted text. The model interprets the text and identifies the signs or gestures associated with each word, phrase, or sentence.

2. Animation File Matching

Once the sign text is generated, the system proceeds to match it with pre-defined animation files. These animation files are typically stored as GIF files, each corresponding to a particular ISL sign gesture. The system maintains a comprehensive database of these GIFs, each representing a specific gesture, hand movement, or sequence in Indian Sign Language.

The matching process is performed by comparing the predicted sign text with entries in the GIF dataset. Each word or phrase in the text is cross-referenced with the dataset to find an appropriate animation that accurately represents the corresponding ISL sign. This process relies on key features such as:

- Linguistic Matching: Matching the predicted sign with its exact representation in the dataset.
- Contextual Matching: In the case of complex sentences or phrases, matching the context of the translation with the correct sequence of gestures.
- Synonym Matching: In some cases, multiple signs might correspond to the same word or phrase, and the system selects the most appropriate sign based on context.

If a match is found, the corresponding GIF animation is retrieved, showcasing the sign gesture visually.

3. Final Translation Outcomes

Once the correct animation file is matched, the final translation outcome is generated. The animation (GIF) is then displayed to the user, serving as the visual translation of the spoken or written text into sign language. This animation file represents the gesture or sign in ISL and is typically shown in real time as part of the user interface.

For example, if the input text is "hai," the system identifies the corresponding sign in ISL, finds the relevant animation from the GIF dataset, and displays it to the user. If the translation is more complex, such as a sentence, the system may generate a sequence of animations that correspond to each word or phrase in the sentence.

4. Final User Interface

The final translation output, whether in the form of a GIF animation or a fallback text representation, is displayed through a GUI. The GUI allows for easy interaction, enabling users to see the translated sign language animation or the alphabetic output. For users unfamiliar with specific signs or gestures, the interface may also provide additional explanations or instructions to ensure the translation is understood clearly.

VI. Results and Discussion

In this experimental analysis, the developed model shows significant outcomes by validating with various performance measures. This accurate performance validation enables the developed model to work in real-time applications.

A. *Simulation Setup*

Python was used for the execution process. The effectiveness of the model was compared with existing models such as LSTM [21], GRU [22], Bi-GRU [23], and Res-BiLSTM [24], and proved its superiority over others. In order to provide superior outcomes, the collected data is split into a training and testing process. Here, 75% of the data is given into training and then the remaining 25% of data given in the testing phase. Moreover, the software

Authorized licensed use limited to: Vidya Vardhaka College Of Engineering Mysore. Downloaded on November 11,2025 at 06:39:55 UTC from IEEE Xplore. Restrictions apply.

requirement contains software - pycharam, version 3.11 and anaconda, version 3 and the hardware requirements involve machine-windows, version 11, processor-i3, RAM-8GB, ROM-500GB. Also, the hardware configuration involves GPU: NVIDIA RTX 3090 (24GB VRAM), CUDA 11.8, cuDNN 8, PyTorch 2.0, and Python 3.10.

B. Performance Measures

This section offers the performed measures applied in the designed system.

(a) Character Sequence Error Distance (CSED):

$$CS = \frac{1}{A} * \Sigma\left(EdiDis\left(R_m, V_m\right)\right) \tag{5}$$

Here, the term A is the total number of character sequences, R_m is the predicted sequence for the m^{th} sample, and V_m is the ground truth sequence for the m^{th} sample.

(b) Character Sequence Information Index (CSII):

$$CSI = \Sigma\left(\frac{\left(length\left(R_m\right) \cap length\left(V_m\right)\right)}{\Sigma length\left(V_m\right)}\right) \tag{6}$$

Here, the term $length\left(R_m\right)$ is the length of the predicted sequence and $length\left(V_m\right)$ is the length of the ground truth sequence.

(c) Extended Short-Time Objective Intelligibility (ESTOI):

$$EST = \Sigma y_p * ST\left(y_p, \overline{y_p}\right) \tag{7}$$

Here, p is the number of frames, and y_p is the weight for each frame.

(d) MAE (Mean Absolute Error):

$$MAE = \frac{1}{A} * \Sigma\left|R_m - V_m\right| \tag{8}$$

(e) Pearson Correlation Coefficient (PCC):

$$PCC = \frac{\left(\Sigma\left(R_m - \overline{R}\right)\left(V_m - \overline{V}\right)\right)}{\sqrt{\left(\Sigma\left(R_m - \overline{R}\right)^2 * \Sigma\left(V_m - \overline{V}\right)^2\right)}} \tag{9}$$

(f) Perceptual Evaluation of Speech Quality (PESQ):

$$PES = g\left(R_m, V_m\right) \tag{10}$$

Here, $g\left(R_m, V_m\right)$ is a perceptual model that compares the predicted signal to the ground truth signal based on human auditory perception.

(g) Peak Signal-to-Noise Ratio (PSNR):

$$PSNR = 10 * \log 10\left(\frac{\max^2}{MSE}\right) \tag{11}$$

Here, the term $\max$ is the maximum possible pixel value (for images or videos, this could be 255), and MSE is the MSE between the predicted and ground truth signals.

(h) Short-Time Objective Intelligibility (STOI):

$$ESOI = \frac{1}{A}\Sigma ST\left(y_p, \overline{y_p}\right) \tag{12}$$

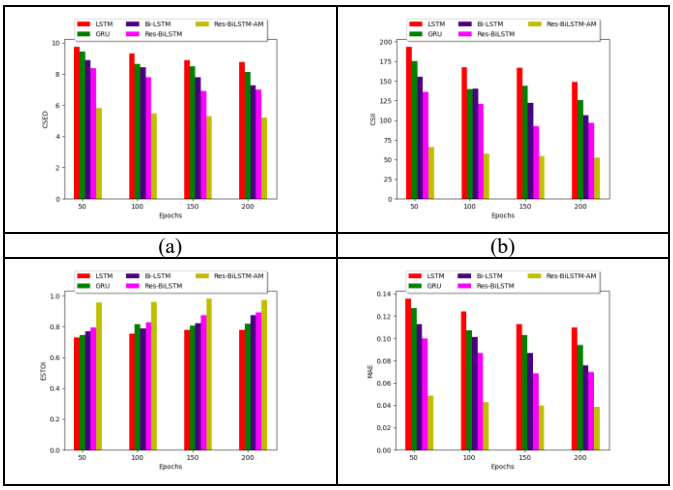
(i) WER (Word Error Rate):

$$WER = \frac{\left(c + j + x\right)}{A} \tag{13}$$

Here, the terms c, j and x is the number of substitutions, deletions, and insertions.

C. Performance Analysis of the Developed S2SLT Model

Performance analysis of the developed S2SLT model is given in Fig. 4. The performance evaluation of the developed S2SLT model focused on assessing its ability to accurately and efficiently translate Telugu speech. The evaluation was conducted through a series of experiments using a labeled dataset of Telugu audio inputs and corresponding ISL signs. The proposed model, built on a Res-BiLSTM-AM architecture, was compared against traditional deep learning models including LSTM, BiLSTM, GRU, and Res-BiLSTM. The comparison was based on several key performance indicators such as translation CSED, CSII, and ESTOI and so on. Here, the STOI of the developed Res-BiLSTM-Am model is 90% at the 50th epoch. Results showed that the Res-BiLSTM-AM model consistently delivered superior performance across all metrics. It achieved the highest translation STOI of over 94%, while LSTM and GRU lagged further behind, indicating a lower ability to capture both long-term dependencies and bidirectional context. Moreover, the attention mechanism in Res-BiLSTM-AM allowed the model to selectively focus on critical parts of the input sequence, improving its precision in sign text prediction. The residual connections contributed to faster convergence and reduced training time, allowing for deeper network design without suffering from vanishing gradients. The performance evaluation confirms that the Res-BiLSTM-AM architecture significantly outperforms traditional models such as LSTM, BiLSTM, GRU, and Res-BiLSTM in the context of Telugu speech-to-sign language translation. With higher accuracy, faster processing, and better sequence modeling capabilities, the Res-BiLSTM-AM provides a more effective solution for real-time and context-aware translation tasks. Its ability to capture both forward and backward dependencies, enhanced with attention and stabilized through residual learning, makes it the most robust and high-performing model for S2SLT applications.



Authorized licensed use limited to: Vidya Vardhaka College Of Engineering Mysore. Downloaded on November 11,2025 at 06:39:55 UTC from IEEE Xplore. Restrictions apply.

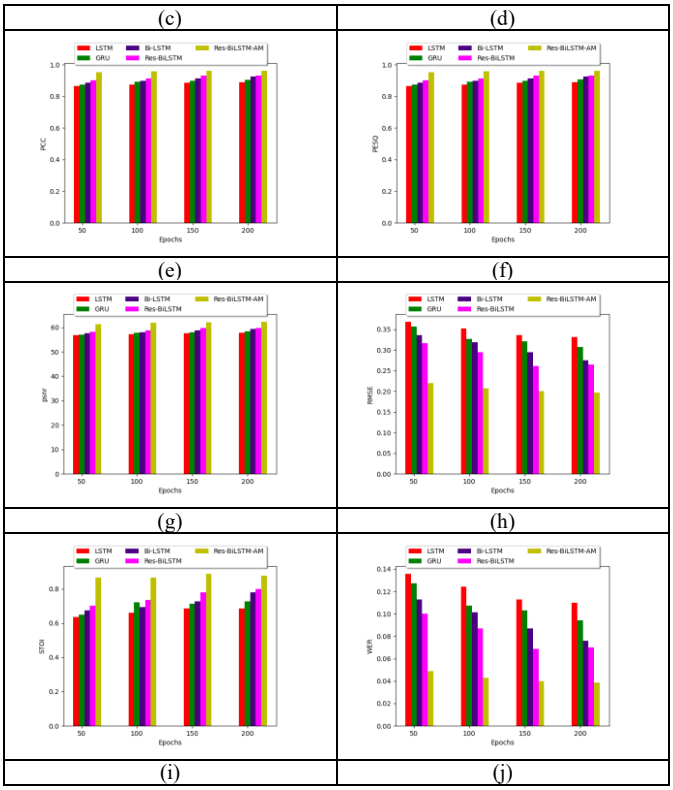


Fig. 4. Performance Analysis of the Developed S2SLT Model based on (a) CSED, (b) CSII, (c) ESTOI, (d) MAE, (e) PCC, (f) PESQ, (g) PSNR, (I) STOI, and (j) WER

D. Comparison Evaluation of the Proposed System

A comparison evaluation of the proposed system is given in Table II. A comprehensive comparison evaluation was conducted to analyze the performance of the proposed Res-BiLSTM-AM model in the S2SLT framework against traditional deep learning models such as LSTM, BiLSTM, GRU, and Res-BiLSTM. The results clearly highlighted the superiority of the Res-BiLSTM-AM model across all evaluation parameters. It achieved the highest PCC (96.1%), outperforming BiLSTM (92.4%), Res-BiLSTM (93%), LSTM (89%), and GRU (90.6%). Unlike the other models, Res-BiLSTM-AM efficiently handled longer speech sequences without degradation in performance, thanks to its residual connections that prevent vanishing gradients during deep training. Moreover, the attention mechanism allowed the model to dynamically focus on the most relevant parts of the input, improving context awareness and translation quality. From the comparison evaluation, it is evident that the Res-BiLSTM-AM model significantly outperforms LSTM, BiLSTM, GRU, and Res-BiLSTM in translating Telugu speech into ISL. Its architecture effectively combines bidirectional context learning, residual connections for deeper training, and an attention mechanism for focused interpretation, making it more accurate, faster to converge, and better suited for real-world S2SLT applications. Therefore, Res-BiLSTM-AM stands as the most advanced and efficient model in this domain.

TABLE II. COMPARISON EVALUATION OF THE PROPOSED SYSTEM

TERMS	LSTM [21]	GRU [22]	Bi-LSTM [23]	Res-BiLSTM [24]	Res-BiLSTM-AM
PCC	0.890	0.906	0.924	0.930	0.961
MAE	0.110	0.094	0.076	0.070	0.039
RMSE	0.332	0.307	0.275	0.265	0.196
PSNR	57.717	58.386	59.339	59.680	62.268
CSED	8.775	8.124	7.280	7.000	5.196
STOI	0.685	0.727	0.779	0.800	0.878
PESQ	0.890	0.906	0.924	0.930	0.961
CSII	149.094	125.706	106.764	97.028	52.866
WER	0.110	0.094	0.076	0.070	0.039
ESTOI	0.779	0.821	0.874	0.895	0.972

VII. CONCLUSION

This paper presented the design and implementation of a S2SLT system for translating spoken Telugu language into ISL animations. The proposed model successfully utilized dual audio features, and they were spectral features and MFCCs to effectively represent speech data. These features were processed using a Res-BiLSTM-AM, which accurately predicted the corresponding ISL sign text. The output text was then matched with a pre-defined GIF animation dataset, and the appropriate ISL animation was displayed. For unmatched inputs, the system displayed the ISL text alphabetically through a GUI, ensuring accessibility. The ESTOI of the proposed Res-BiLSTM-AM model was 97.2%, which was better than LSTM, GRU, Bi-LSTM, and Res-BiLSTM. Thus, it proved that the proposed system achieved high translation accuracy and faster convergence due to the use of residual connections and attention mechanisms, which enhanced learning and focused on contextually important features. The model demonstrated strong performance in handling long or complex speech sequences and showed robustness when dealing with unseen data. However, the system had some limitations. The translation capability was constrained by the size and coverage of the GIF dataset. Additionally, the model was specifically trained on Telugu speech, limiting its direct applicability to other Indian languages without retraining. For future scope, the system can be extended by incorporating a larger and more diverse GIF dataset to cover a wider range of ISL vocabulary. Further, multilingual support can be integrated by training the model in additional Indian languages. The research work adopts an effective deep learning model to enhance the performance in speech-to-sign language. It has the credibility to enable better communications for deaf people. In the recent era, 'Nora' application has been implemented in Germany. The main intention of this application enables the deaf people to communicate effectively during emergency situations to make calls. Moreover, this application involves an icon-based menu and text-based chat for communications. However, this app provides limited service and it is problematic because of uneven social engagement. In order to provide better social engagement, the complementary digital applications, like smartphones, are introduced. It can promote equality with the help of Artificial Intelligence (AI), machine learning and deep learning models. The mobile-based application helps to support the deaf people with digital environments.

References

[1] Cassim, Muhammed Rashaad, Jason Parry, Adam Pantanowitz, and David M. Rubin, "Design and construction of a cost-effective, portable sign language to speech translator," Informatics in Medicine Unlocked, Vol 30, 2022.

[2] Paneru, Biplov, Bishwash Paneru, and Khem Narayan Poudyal, "Advancing Human-Computer Interaction: AI-Driven Translation of American Sign Language to Nepali Using Convolutional Neural

Authorized licensed use limited to: Vidya Vardhaka College Of Engineering Mysore. Downloaded on November 11,2025 at 06:39:55 UTC from IEEE Xplore. Restrictions apply.

Networks and Text-to-Speech Conversion Application," Systems and Soft Computing, 2024.

[3] Ojha, Ankit, Ayush Pandey, Shubham Maurya, Abhishek Thakur, and P. Dayananda, "Sign language to text and speech translation in real time using convolutional neural network," IJERT, Vol. 8, no. 15, Vol. 191-196, 2020.

[4] Yin, Kayo, "Sign language translation with transformers," arXiv preprint arXiv:2004.00588, Vol. 2, 2020.

[5] Sharma, Purushottam, Devesh Tulsian, Chaman Verma, Pratibha Sharma, and Nancy Nancy, "Translating speech to indian sign language using natural language processing," Future Internet, Vol. 14, no. 9, 2022.

[6] Yin, Kayo, and Jesse Read, "Better sign language translation with STMC-transformer," arXiv preprint arXiv:2004.00588, 2020.

[7] Camgoz, Necati Cihan, Oscar Koller, Simon Hadfield, and Richard Bowden, "Multi-channel transformers for multi-articulatory sign language translation," Computer Vision–ECCV 2020 Workshops, pp. 301-319, 2020.

[8] Amin, Mohamed, Hesahm Hefny, and Mohammed Ammar, "Sign language gloss translation using deep learning models," International Journal of Advanced Computer Science and Applications, Vol. 12, no. 11, 2021.

[9] Majumdar, Shoumik Sovan, Shubhangi Jain, Isidora Chara Tourni, Arsenii Mustafin, Diala Lteif, Stan Sclaroff, Kate Saenko, and Sarah Adel Bargal, "Ani-GIFs: A benchmark dataset for domain generalization of action recognition from GIFs," Frontiers in Computer Science, Vol. 4, 2022.

[10] Mujtaba, Ghulam, Sunder Ali Khowaja, Muhammad Aslam Jarwar, Jaehyuk Choi, and Eun-Seok Ryu, "FRC-GIF: Frame Ranking-Based Personalized Artistic Media Generation Method for Resource Constrained Devices," IEEE Transactions on Big Data, 2023.

[11] Deepak Rai; Niharika Rana; Naman Kotak; Manya Sharma, "Real-Time Speech to Sign Language Translation Using Machine and Deep Learning," 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2024.

[12] Durjay Thalisetty; Anvitha Kopparthi; Snigdha Kambhampati; Kanuri Sathvika; Suresh Kumar Natarajan, "A Novel Deep Learning Approach for Real Time and Accurate Sign Language Translation," 2024 2nd International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT), 2024.

[13] Reddy, Bandi Rupendra, Darukumalli Sai Tharun Reddy, Sandeep Preetham, and Susmitha Vekkot, "Creation of GIF dataset and implementation of a speech-to-sign language translator in Telugu," IEEE North Karnataka Subsection Flagship International Conference, pp. 1-7, 2022.

[14] Khan, Nabeel Sabir, Adnan Abid, and Kamran Abid, "A novel natural language processing (NLP)–based machine translation model for English to Pakistan sign language translation," Cognitive Computation, Vol. 12, pp. 748-765, 2020.

[15] Kapoor, Parul, Rudrabha Mukhopadhyay, Sindhu B. Hegde, Vinay Namboodiri, and C. V. Jawahar, "Towards automatic speech to sign language generation," arXiv preprint arXiv:2106.12790, 2021.

[16] Joy, Linu, Midhuna Eldho, Meera Ajith, and Chinnu Mariya Varghese, "Speech to Sign Language Converter," International Journal of Scientific Research & Engineering Trends, Vol. 10, Issue. 4, 2024.

[17] Liang, Wuyan, and Xiaolong Xu, "Sindiff: Spoken-to-Sign Language Generation Based Transformer Diffusion Model," Jiangsu Key Laboratory of Big Data Security & Intelligent Processing, 2023.

[18] Rao, K. Sreenivasa, V. Ramu Reddy, and Sudhamay Maity, Language identification using spectral and prosodic features, Springer, 2015.

[19] Vijaya, J., Chahat Mittal, Chinmay Singh, and M. A. Lekhana, "An Efficient System for Audio-Based Sign Language Translator Through MFCC Feature Extraction," In 2023 International Conference on Sustainable Communication Networks and Application (ICSCNA), pp. 1157-1164. IEEE, 2023.

[20] J. Zhang, Y. Liu and H. Yuan, "Attention-Based Residual BiLSTM Networks for Human Activity Recognition, "IEEE Access, vol. 11, pp. 94173-94187, 2023.

[21] Kumaravel, Dr S. "Generating Sign Language by Utilizing the LSTM Algorithm to Process and Convert Spoken Language Inputs into Corresponding ISL." (2024).

[22] Ilham, Amil Ahmad, and Ingrid Nurtanio. "Applying LSTM and GRU Methods to Recognize and Interpret Hand Gestures, Poses, and Face-Based Sign Language in Real Time," Journal of Advanced Computational Intelligence and Intelligent Informatics, vol. 28, no. 2, pp. 265-272, 2024.

[23] Pradhan, Anima, Manas Ranjan Senapati, and Pradip Kumar Sahu, "A multichannel embedding and arithmetic optimized stacked Bi-GRU model with semantic attention to detect emotion over text data," Applied Intelligence, vol. 53, no. 7, pp. 7647-7664, 2023.

[24] Zhao, Yantao, Yao Wang, Shanshan Zhang, Xin Wang, and Hongnian Yu, "Res-BiLSTMs model based on multi-task attention for real-time measurement of the free calcium oxide content," Measurement Science and Technology, vol. 35, no. 9, 2024.

Authorized licensed use limited to: Vidya Vardhaka College Of Engineering Mysore. Downloaded on November 11,2025 at 06:39:55 UTC from IEEE Xplore. Restrictions apply.