Speech to Sign Language Translation for Indian Languages

Jashwanth Peguda
Department of Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
jashwanthpegudapj@gmail.com

V Sai Sriharsha Santosh
Department of Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
harshasantoosh2000@gmail.com

Y Vijayalata
Department of Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
vijaya@griet.ac.in

Ashlin Deepa R N
Department of Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
rndeepa.pradeep@gmail.com

Vaddi Mounish
Department of Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
mounish789@gmail.com

2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS) | 978-1-6654-0816-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICACCS54159.2022.9784996

***Abstract*-- Hearing-impaired people and mute people face a lot of difficulty in communication while interacting with others in society. It may reduce their self-confidence and might make them feel isolated from others. Sign language acts as a communication medium between deaf people and ordinary people. Many technologies are used to convert the text to American Sign Language. There is a limited amount of research done on Indian Sign Language and is widely used by deaf people in India. This research aims at conversion of speech to Indian sign language for six Indian regional languages Telugu, Hindi, Malayalam, Marathi, Kannada and Tamil. The proposed model takes speech as the input and displays a sequence of corresponding gestures as the output. It involves speech recognition using Wavelet-based MFCC with GMM, text translation using LSTM and mapping the text with the sign language.**

Keywords— Sign Language, Speech Recognition, Text Translation, Gaussian Mixture Model, Expectation-Maximization, Long Short Term Memory, Indian Sign Language, Speech to Sign language.

I. INTRODUCTION

Sign Language involves visual gestures and signs, which deaf people and mute people use. It involves manual and non-manual signals, where manual signs involve fingers, hands, arms, and non-manual signs involve the face, head, eyes, and body. There are 18 million hearing-impaired in India; four in every 1000 children suffer from severe to profound hearing loss. Many firms are constantly searching for skilled and talented individuals, but the people who cannot talk and hear happen to lose many job opportunities. Deaf and mute people feel ostracized as they cannot communicate with ordinary people. It is challenging for ordinary people to communicate with deaf and mute people as they are unfamiliar with sign language. There are many sign languages in the world where each country has its sign language, such as American Sign Language (ASL) [1], Japanese Sign Language (JSL) [2], Indian Sign Language (ISL), Arabic Sign Language [3], Etc. American Sign Language uses one hand, whereas Indian Sign Language involves using both hands. Furthermore, Japanese sign language considers mouthing along with hand signs, while Arabic sign language is still developing. But, in India, ISL is more widely used than any other sign language. Many systems are built on ASL, but only a few are developed using ISL. Some ISL systems convert the sign language to speech, but no system converts the regional speech to Sign Language [4].

Much research has been conducted in the field of continuous Speech Recognition of Indian languages such as Telugu [5], [6], Tamil [7], [8], Kannada [9], Marathi [10], Malayalam [11], Hindi [12], Etc. Along with speech recognition, text translation has also been a field of research that has been active for an extended period. Many papers are present on Text Translation for various languages, such as Telugu [13], Marathi [14], Malayalam [15], Hindi [16], Etc. to English text. Some systems translate regional text into Indian sign language using LSTM models, while others convert regional speech to text by MFCC with HMM, Naïve Bayes, etc. However, no system directly translates regional speech to Indian Sign Language. This work builds a system that can convert the speech to ISL for six Indian regional languages such as Telugu, Hindi, Tamil, Malayalam, Kannada, and Marathi. It is implemented using wavelet-based Mel-Frequency Cepstral Coefficients(MFCC) with Gaussian Mixture Model(GMM) for Speech Recognition, Encoder-Decoder based Long Short Term Memory(LSTM) for Text Translation, and Indian Sign Language (ISL) generation. Research shows that Gaussian models outperform in recognition applications [17], [18].

The paper is divided into six sections. Section I gives a brief introduction to Indian Sign Language. Section II describes the related works in Speech Recognition and Text Translation. Section III gives a brief description of the proposed approach. Section IV shows the evaluation and

978-1-6654-0816-5/22/$31.00 ©2022 IEEE

results achieved by the system. Finally, Section V gives the conclusion and future scope of the system.

II. RELATED WORKS

Yogeshwar I. Rokade et al. [19] proposed a model that uses a publicly available dataset containing signs corresponding to all English alphabets. These signs are captured from the images, and the model is built, which provides the significance of ISL. Another model is developed using vision-based hand gesture recognition using a web camera [20]. Shashi Pal Singh et al. proposed a model using deep neural networks in machine translation [21] by performing word alignment, rule selection and reordering, language modelling, and joint translation. Finally, a model has been proposed in [22], which gives an idea of the working of various algorithms with Mel-Frequency Cepstral Coefficients (MFCC) like Gaussian Mixture Model (GMM), kullback-Leibler divergence, DT-CWT with RVM, and weighted VQ.

In research conducted for Telugu language speech, a speech synthesizer [5] is developed based on feature extraction, acoustic model generation, and language model generation for the input speech. Also, the sphinx-3 algorithm, which has two phases: training and decoding, converts Telugu speech to text [6]. A rule-based language morphological analyzer is used for speech processing, machine translation, and information extraction with the help of an unsupervised stemmer, word segmentation algorithm, and clustering stems and suffixes [13]. Various MFCC models are developed for Speech Recognition of Tamil and Kannada languages, but no such research is done to produce text to ISL models [7], [8], [9]. Similarly, Hidden Markov Model Toolkit (HTK) [10] is used to train a speech analyzer that takes 20 phonetically rich Marathi words spoken by ten native speakers. A rule-based translator [14] translates Marathi phrases and Signing Gesture Mark-up Language (SiGML) to interpret the ISL signs. In another research conducted in the Malayalam language, the Kaldi toolkit [11] is used for Speech Recognition, and SiGML generates HamNoSys [15] text notation. Likewise, with the help of the GoldWave Software tool [12], Hindi, Manipuri, and Urdu speech signals are captured and synthesized. Furthermore, a bi-directional method is proposed to convert text to gesture (T2G) [16].

In the field of Audio classification and segmentation, various MFCCs with GMM models are developed by adjusting their parameters [23], [24]. Additionally, the output is improved with the parameters having a maximum log-likelihood score [25]. A gesture recognition glove-based model [26] uses LSTM for 30 epochs to train 40 samples of each character present in ISL. Moreover, sentences that consist of complex words use the RTR approach [27] alongside LSTM. Due to the availability of many regional languages, besides the script's complexity and limited resources such as corpus, no work has been done to implement a system that utilizes more than one language at a time to generate ISL.

III. PROPOSED APPROACH

The speech corpus for the languages Tamil, Telugu, Hindi, Marathi, Malayalam, and Kannada are collected from CommonVoice by Mozilla and MultiIndicMT by Kyoto University to train the model. Similarly, the text corpus for the same above languages is collected from ManyThings.org Anki.

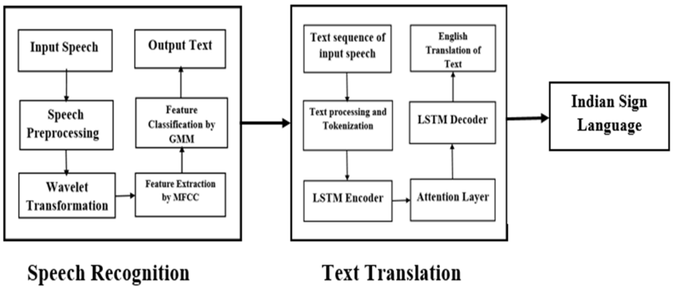


Fig.1. Model Architecture

A. Speech Recognition:

Feature description and classification are two important stages of all machine learning applications. This model starts with a Speech Recognition phase considering Wavelet-based MFCC with GMM. Then, Discrete Wavelet Transformation (DWT) is performed on input speech to decompose them into sub-signals using orthonormal wavelet function. Finally, the sequence of sub-signals with reduced noise is obtained after DWT is given as input to MFCC.

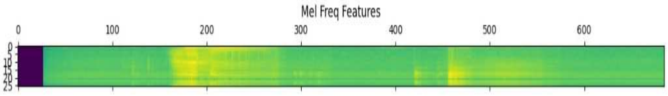


Fig.2. MFCC features of a sample speech

MFCC is used for feature extraction in the speech signal as it is more resistant to handling noisy speech and provides the essential features present in the audio signal. Next, GMM takes the extracted features as input and performs classification. The parameters of GMM are identified using Expectation-Maximization.

First, the input speech is stored in a .wav file using a 16-bit pulse core modulator at a sampling rate of 8 kHz. Then, the wave files are decomposed into approximation coefficients and detailed coefficients using DWT. The sub-signals generated are given to the MFCC for feature extraction. It involves windowing, pre-emphasis of the speech, filtering based on the Mel Frequency scale, Applying Discrete Cosine Transformations for speech compressions, and using Fast Fourier transformations with logarithmic functions to give an output of reduced speech for a smooth recognition process. A frame of 256 speech cases is considered, and 128 samples that are overlapped are kept in an adjacent frame. Finally, the cepstral coefficients are calculated from each frame by performing feature extraction. The number of coefficients, Frame length, Frame stride: The step between the successive frame in seconds, the Filter number, FFT length, Low frequency, High frequency, Window size are the parameters present in MFCC.

$$w = \alpha - \beta \cos\left(\frac{2\pi n}{N-1}\right) \quad (1)$$

(1) denotes hamming window coefficients where α, β are constants, N is the length of the filter and n = 0, 1… N – 1.

$$c(k) = \sum_{m=0}^{M-1} S(m) \cos\left(\frac{\pi k\left(m+\frac{1}{2}\right)}{M}\right), \quad 0 \leq k < K \qquad (2)$$

(2) denote Discrete Cosine Transformation where k is the coefficients of each frame.

$$mean(\mu_i) = \frac{\sum_{t=1}^{T} p(i|x_t)}{\sum_{t=1}^{T} p(i|x)} \qquad (3)$$

$$variance = \frac{\sum_{t=1}^{T} p(i|x_t,\lambda)(x_t)}{\sum_{t=1}^{T} p(i|x_t,\lambda)} \qquad (4)$$

$$Weights\ (w_i) = \frac{1}{T}\sum_{t=1}^{T} P(i) \qquad (5)$$

(3), (4), and (5) denotes the parameters considered in GMM to classify the features obtained from MFCC.

Finally, the DCT is applied to obtain the feature vector containing the cepstral coefficients. GMM (Gaussian Mixture Model) is a probability density function with specific parameters such as mean, variance, and weights of the Gaussian distribution.

$$p\left(\frac{x}{t}\right) = \prod_{t=1}^{T} p\left(\frac{x_t}{\lambda}\right) \qquad (6)$$

(6) denotes maximum-likelihood in expectation-maximization where T is number of training vectors.

The parameters of GMM is calculated by using maximum-likelihood estimation method. Iterative EM is used to find the model parameters in order to maximize the likelihood of GMM through training and matching. GMM provides the best selection of random values for missing data, and it classifies the features extracted using MFCC. The Gaussian mixture density is composed of N component densities of the distribution. GMM considers the weighted sum of Gaussian densities, and each part of it is computed by its variance, mean, and weight of the respective Gaussian distribution. The probability density function used in GMM consists of continuous measurements of vocal-related features present in input speech. It combines the probability distribution of various classes, calculates the probability for a single class, and then checks for the specific word present in a dictionary, and the output consists of the text of the speech given.

$$p\left(\frac{x}{\lambda}\right) = \sum_{i=0}^{M} p_i b_i(x) \qquad (7)$$

$$b_i(x) = \frac{1}{2\pi^{D/2}|\sum i|^{1/2}} e^{-1/2(x^2-\mu_{i)'}\Sigma_i^{-1}(x^2-\mu)} \qquad (8)$$

In (7) and (8), $b_i(x)$ denotes component densities where i is the mean vector.

B. Text Translation:

Text Translation can be performed using Long Short Term Memory (LSTM). LSTM is an updated version of Recurrent Neural Networks used to solve sequence estimation problems by utilizing learning order dependence. It is capable of maintaining long-term and short-term dependencies. The semantic relationship between the words is not lost during the translation phase, unlike using RNN. LSTM uses an optimized algorithm similar to gradient descent along with Back Propagation. LSTM operates on vectors that are generated from the text. These vectors are of two types. The one is a previous LSTM unit output that provides information about the previous unit inputs, and another one is the current input. The previous unit outputs are multiplied to obtain the current output and judge the influence of an LSTM unit. If the vector returned consists of minimal values, the state is removed.

Translation involves pre-processing, modelling, prediction, and iteration. Pre-processing tokenizes, filters, and performs padding on the input text. Next, the model is built on pre-processed text, and translations from one language to another are carried through using prediction. While iteration is used to re-evaluate the model's efficiency and to assess the model's compatibility with other build designs. LSTM is used with encoder-decoder for text translation. The encoder part of the model takes the input as one word at a time in a sequential manner. The internal vectors are updated after every word is consumed. The decoder part of the model takes the contents present in the encoder block and produces a translated token for every word arrived at the decoder phase. The encoder and decoder part of the LSTM only work effectively with shorter sequences.

$$h_t = LSTM(x_t, h_{t-1}) \qquad (9)$$

In (9), h_t denotes the hidden state for encoder at time step t and x_t is the input.

$$s_t = LSTM(y_t, c_t, s_{t-1}) \qquad (10)$$

In (10), s_t denotes the hidden state for decoder, y_{t-1} is the output from previous time step, and c_t is a context vector.

Attention Layer can retain longer sequences and transform them to fixed-length vectors, and it, in turn, predicts the next word focussing only on the relevant parts of the sequences. This layer is added between the encoder and decoder part of the LSTM model as it extracts essential information from the encoder and transmits it to the decoder. The expected target tokens and error values are calculated, which helps the model update itself for the translation of every word. The model works with fixed-length sentences, and padding is provided if the maximum length of a sentence is violated. Priority is given more to the internal state for both the encoding and decoding phase to create a probability distribution of all translated words that have the vocabulary appropriate to the target language. It helps the model to pick the word which has a higher probability. Then, an inference model is built upon the encoder and decoder phase that utilizes the training. The encoder part of the inference model takes the training layer and produces an output fed to the decoder. For every word being translated, previous internal states of a decoder are supplied to gain insight into a current word to predict the following sequence. Thus, it maintains the semantic balance. An Adam optimizer is used to achieve maximum training stability and translation accuracy. The Adam optimizer gave better accuracy when compared to other optimizers such as SGD, RMSprop, etc.

C. Sign Language Generation:

Finally, the text generated from LSTM is converted into Indian Sign Language using the Direct Translation method. In this method, the text is mapped character-by-character with corresponding signs of Indian Sign Language using index-based mapping.

IV. RESULTS AND ANALYSIS

The expected system is developed, and the epochs for the encoder-decoder based LSTM model is set at 100 with the learning rate of 0.001.

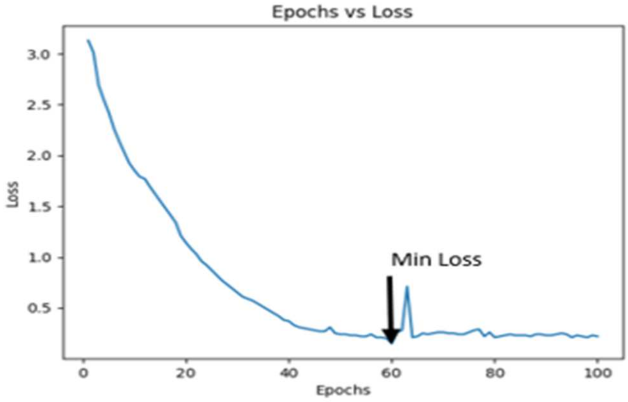


Fig. 3. Training Loss of the model

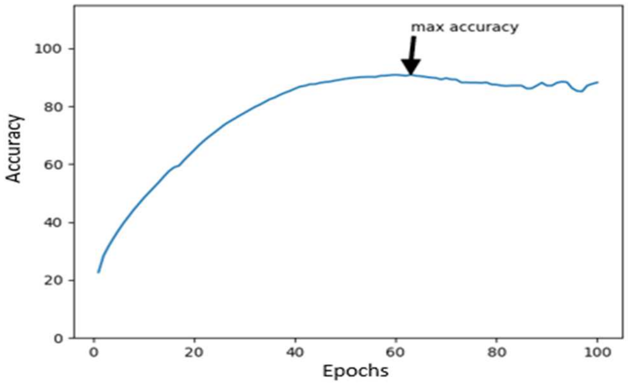


Fig. 4. Training Accuracy of the model

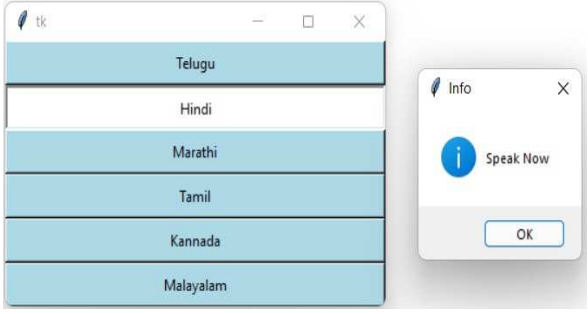
Fig. 3 and 4 illustrate the training loss and accuracy for the experiment conducted on the Hindi language dataset. The maximum accuracy and minimum loss are 91.88% and 0.15% obtained at 62nd and 60th epochs, respectively for the same above experiment conducted.

TABLE I.
ACCURACY FOR SIX INDIAN REGIONAL LANGUAGES

Regional Language	**Number of test cases**		
	100	***300***	***500***
Telugu	90	88.67	86.4
Hindi	91	90.33	88.6
Tamil	89	87.33	85
Malayalam	88	85	81.4
Kannada	89	87.67	85.8
Marathi	90	88.33	86.6

TABLE I. denotes the validation accuracy of speech to sign language for the six Indian languages along with the number of test cases.

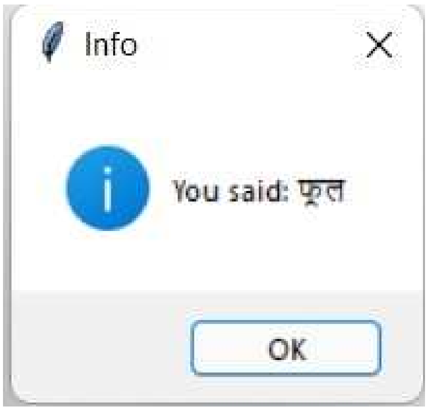
The following images depict the sequence of steps involved in the translation of speech to sign language translation. For the experiment, the language chosen is Hindi.



Speak Now

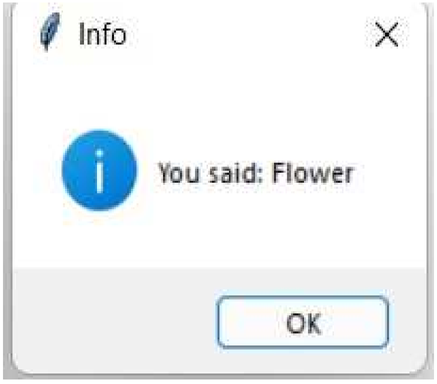
Fig.5. Dashboard to select the regional language

Fig. 5 shows the dashboard consisting of six regional languages. Among which, the user selects the Hindi language for the sample experiment conducted. After language selection, the system requests the user to provide input speech in the selected language.



You said: फूल

Fig.6. Text in regional language



You said: Flower

Fig.7. Text in English

Fig. 6 and 7 depicts the output generated after Speech Recognition and Text Translation phases, respectively.

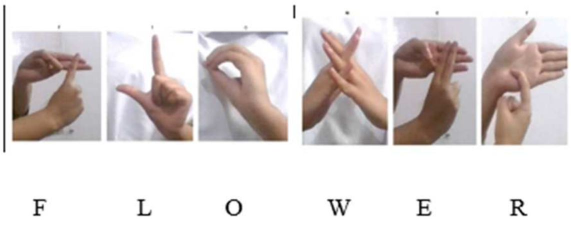


Fig.8. Direct translation of text to Indian Sign Language.

Fig. 8 shows the sequence of signs generated using direct translation method of Sign Language Generation phase.

V. CONCLUSION AND FUTURE SCOPE

This paper focuses on converting the speech in regional language into Indian Sign Language. This model supports up to six Indian Regional languages which are Telugu, Hindi, Tamil, Marathi, Malayalam, and Kannada and it acquires more than 80% accuracy for each regional language. The Speech Recognition is implemented using Wavelet-based MFCC with GMM, Text Translation using LSTM with encoder-decoder and Sign Language Generation using Direct Translation.

The system can be further extended to implement other Indian language dialects. Instead of direct machine translation, a semantic translation along with 2D animated sign generation can be added to the existing system for better user experience.

References

[1] M. Taskiran, M. Killioglu and N. Kahraman, "A Real-Time System for Recognition of American Sign Language by using Deep Learning," 2018 41st International Conference on Telecommunications and Signal Processing (TSP), 2018, pp. 1-5, doi: 10.1109/TSP.2018.8441304.

[2] S. -i. Ito, M. Ito and M. Fukumi, "A Method of Classifying Japanese Sign Language using Gathered Image Generation and Convolutional Neural Networks," 2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), 2019, pp. 868-871, doi: 10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00157.

[3] A. M. Zakariya and R. Jindal, "Arabic Sign Language Recognition System on Smartphone," 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2019, pp. 1-5, doi: 10.1109/ICCCNT45670.2019.8944518.

[4] Anuja V. Nair, Bindu, “A Review on Indian Sign Language Recognition”, International Journal of Computer Applications (0975 – 8887), Jul. 2013, vol. 73, no. 22, pp. 33-38, doi: 10.5120/13037-0260.

[5] G. Ramya and N. S. Naik, "Implementation of telugu speech synthesis system," 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2017, pp. 1151-1154, doi: 10.1109/ICACCI.2017.8125997.

[6] M. R. Reddy et al., "Transcription of Telugu TV news using ASR," 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2015, pp. 1542-1545, doi: 10.1109/ICACCI.2015.7275832.

[7] D. Pubadi et al., "A focus on codemixing and codeswitching in Tamil speech to text," 2020 8th International Conference in Software Engineering Research and Innovation (CONISOFT), 2020, pp. 154-165, doi: 10.1109/CONISOFT50191.2020.00031.

[8] K. R., N. K., P. D. S. and S. T., "Voice and speech recognition in Tamil language," 2017 2nd International Conference on Computing and Communications Technologies (ICCCT), 2017, pp. 288-292, doi: 10.1109/ICCCT2.2017.7972293.

[9] S. C. Sajjan and Vijaya C, "Continuous Speech Recognition of Kannada language using triphone modeling," 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2016, pp. 451-455, doi: 10.1109/WiSPNET.2016.7566174.

[10] S. Sawant and M. Deshpande, "Isolated Spoken Marathi Words Recognition Using HMM," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1-4, doi: 10.1109/ICCUBEA.2018.8697457.

[11] L. B. Babu, A. George, K. R. Sreelakshmi and L. Mary, "Continuous Speech Recognition System for Malayalam Language Using Kaldi," 2018 International Conference on Emerging Trends and Innovations In Engineering And Technological Research (ICETIETR), 2018, pp. 1-4, doi: 10.1109/ICETIETR.2018.8529045.

[12] S. Bansal and S. S. Agrawal, "Development of Text and Speech Corpus for Designing the Multilingual Recognition System," 2018 Oriental COCOSDA - International Conference on Speech Database and Assessments, 2018, pp. 1-8, doi: 10.1109/ICSDA.2018.8693013.

[13] K. V. N. Sunitha and N. Kalyani, "A Novel approach to improve rule based Telugu morphological analyzer," 2009 World Congress on Nature & Biologically Inspired Computing (NaBIC), 2009, pp. 1649-1652, doi: 10.1109/NABIC.2009.5393637.

[14] S. R. Bhagwat, R. P. Bhavsar and B. V. Pawar, "Translation from Simple Marathi sentences to Indian Sign Language Using Phrase-Based Approach," 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), 2021, pp. 367-373, doi: 10.1109/ESCI50559.2021.9396900.

[15] M. S. Nair, A. P. Nimitha and S. M. Idicula, "Conversion of Malayalam text to Indian sign language using synthetic animation," 2016 International Conference on Next Generation Intelligent Systems (ICNGIS), 2016, pp. 1-4, doi: 10.1109/ICNGIS.2016.7854002.

[16] S. Chaman, D. D'souza, B. D'mello, K. Bhavsar and J. D'souza, "Real-Time Hand Gesture Communication System in Hindi for Speech and Hearing Impaired," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), 2018, pp. 1954-1958, doi: 10.1109/ICCONS.2018.8663015.

[17] Babu, Domala Kishore & Yellasiri, Ramadevi & Ramana, K, “RGNBC: Rough Gaussian Naïve Bayes Classifier for Data Stream Classification with Recurring Concept Drift,” Arabian Journal for Science and Engineering, 2016, doi: 42. 10.1007/s13369-016-2317-x.

[18] Babu, D. Kishore; Ramadevi, Y.; Ramana, K.V, “PGNBC: Pearson Gaussian Naïve Bayes classifier for data stream classification with recurring concept drift,” Intelligent Data Analysis, 21(5), pp. 1173–1191,2017, doi: 10.3233/IDA-163020 .

[19] Rokade, Yogeshwar & Jadav, Prashant, “Indian Sign Language Recognition System,” International Journal of Engineering and Technology, vol. 9, pp. 189-196, Jul. 2017, doi: 10.21817/ijet/2017/v9i3/170903S030.

[20] Sadhana Bhimrao Bhagat, Dinesh V. Rojarkar , “Vision based sign language recognition: a survey,” JETIR (ISSN-23495162), 2017, vol. 4, pp. 130-134.

[21] S. P. Singh, A. Kumar, H. Darbari, L. Singh, A. Rastogi and S. Jain, "Machine translation using deep learning: An overview," 2017 International Conference on Computer, Communications and Electronics (Comptelix), 2017, pp. 162-167, doi: 10.1109/COMPTELIX.2017.8003957.

[22] Sunitha, C. & Chandra, Evania, “Speaker Recognition using MFCC and Improved Weighted Vector Quantization Algorithm,” International Journal of Engineering and Technology, 2015, vol. 7, pp. 1685-1692.

[23] Z. Weng, L. Li and D. Guo, "Speaker recognition using weighted dynamic MFCC based on GMM," 2010 International Conference on Anti-Counterfeiting, Security and Identification, 2010, pp. 285-288, doi: 10.1109/ICASID.2010.5551341.

[24] L. Jiqing, D. Yuan, H. Jun, Z. Xianyu and W. Haila, "Sports audio classification based on MFCC and GMM", 2009 2nd IEEE International Conference on Broadband Network & Multimedia Technology, 2009, pp. 482-485, doi: 10.1109/ICBNMT.2009.5348520.

[25] M. V. Unnikrishnan and R. Rajan, "Mimicking voice recognition using MFCC-GMM framework," 2017 International Conference on Trends in Electronics and Informatics (ICEI), 2017, pp. 301-304, doi: 10.1109/ICOEI.2017.8300936.

[26] E. Abraham, A. Nayak and A. Iqbal, "Real-Time Translation of Indian Sign Language using LSTM," 2019 Global Conference for Advancement in Technology (GCAT), 2019, pp. 1-5, doi: 10.1109/GCAT47503.2019.8978343.

[27] X. Huang, H. Tan, G. Lin and Y. Tian, "A LSTM-based bidirectional translation model for optimizing rare words and terminologies," 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD), 2018, pp. 185-189, doi: 10.1109/ICAIBD.2018.8396191.