2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT) | 979-8-3503-7024-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCCNT61001.2024.10725421

Beyond Words: Speech to Sign Language Interpreter

Lalitha S
Department of Electronics and Communication Engineering,
Amrita School of Engineering, Bengaluru,
Amrita Vishwa Vidyapeetham, India
s_lalitha@blr.amrita.edu

Varshita Adavi
Department of Electronics and Communication Engineering,
Amrita School of Engineering, Bengaluru,
Amrita Vishwa Vidyapeetham, India
varshita.adavi@gmail.com

Abstract - The communication between hearing-impaired or mute people and others becomes arduous and may create a gap among the people. The process of learning and understanding sign language may be inconvenient to some, and therefore, this work proposes a solution to overcome this problem by providing an automatic voice-to-sign language interpretation system using Speech processing, HamNoSys, and SiGML techniques. This work aims to convert speech in any language into a 3d animated video displaying the corresponding Indian sign gestures. This work precisely converts real-time speech into 3D sign language videos with at most accuracy, leading to efficacy in facilitating communication for the especially able community.

Keywords - Sign Language, Speech Recognition, Text Translation, Hamburg Notation System, SIGML, Indian Sign Language, Speech to Sign Language.

I. INTRODUCTION

Given the advancements in technology, it's clear that machines can simplify challenging tasks. As each day passes, technology has made human life easier. Approximately, 250 certified sign language interpreters are translating for 1.8 to 7 million hearing-impaired people in India [1]. So, the system introduced here helps the speech and hearing impaired by making the surroundings more interactive and diverse.

According to the World Health Organization (WHO), it is determined that more than 5% of the world's population has some form of disability. Moreover, it is predicted that nearly 2.5 billion people are expected to have some degree of deafness by 2050 [2]. Over one billion young people are at possibility of permanent and preventable deafness due to unsafe listening practices. In developing countries, children with hearing and speech disabilities are often uneducated. Adults with deafness also have a much higher unemployment rate. Also, few people feel socially isolated, lonely, or stigmatized.

This work gives an automated system utilizing sign language recognition, HamNoSys, and the SIGML player to interpret live or recorded audio into animated sign language gestures, fostering inclusive communication for the hearing-impaired. Unlike existing methods [3], it focuses on real-time conversion of audio to sign language, facilitating equal participation and breaking communication barriers. Additionally, the system accommodates multiple languages, enhancing its accessibility and usability.

II. LITERATURE SURVEY

The overall design of the system can be categorized into three parts consisting of speech-to-text conversion, creating a dataset, and Sign Language video generation. Neha Sharma and Shipra Sardana proposed a system to convert speech to text using Bidirectional Kalman Filter in MATLAB [3]. Comparing Kalman filter's dataset to the standard TIDGIT database, resulted in more accuracy of at least 80% than the Hidden Markov Model Speech Recognition System. In the other system proposed by Kanchan, Unique English phrases and accents in varying pitches are used as input speech and it uses HMM and MFCC to evaluate the performance of Speech Recognition they used periodogram power spectral density, and spectrogram was visualized for variation in frequency concerning time and power frequency [4]. Yogitha et al. came up with a method where it converts Speech to text for multiple languages using MFCC for feature extraction, Minimum Distance Classifier, and Support Vector Machine (SVM) for speech classification. Performance metrics were calculated by visualizing the graphs for the Power spectral Density of the audio, the MFCC filter weights, and the MFCC Discrete Cosine Transform Matrix. They also calculated the accuracy of each language [5].

In a review of Sign Language translation systems, Yuvraj et al. use both text and speech as input and use Google

TTS and STT API to convert speech to text, and real-time video is converted into relevant speech which is in turn converted into animated gestures. Generation of SiGML by taking English text as input in some of the papers has an accuracy of 100% ISL dataset and 98.91% ASL dataset by feeding the speech to Long- Short-Term Memory (LSTM) Networks [6]. Megha et al. map the HamNoSys string to the corresponding word. This HamNoSys string is then converted to an XML file called SiGML. It takes the help of the NLTK library in Python to extract the root words. 174 sentences are taken into consideration giving an accuracy of 87% [7]. Khushdeep et al. use a JA SiGML Player to generate the sign language outputs and SiGML words generated in eSIGNEditor which contains the database for American and British Sign Language. It is tested for 250 words and evaluated by sign language experts at an organization [8].

The 3D avatar-based approach proposed by Debashish generates a continuous moment for Sign generation. The text translation results are evaluated based on the Word Error Rate which is about 25.2% and it uses the Sign Error Rate of 10.50 which signifies an accuracy of 89.5% of signs were correctly generated [9]. In the Sandeep et al. system, A total of 100 words are generated and it uses HamNoSys which is then converted into SiGML using JA SiGML URL App [10]. MultiFacet introduces a framework for speech-to-sign language generation, which aids in converting content from audio and semantic context through text. It is developed using Facial Action Unit prediction within a multi-tasking setup [11]. Audio to Sign Language by Nayana J, aims in converting speech into sign language, specifically focusing on Indian Sign Language. With the use of NLP and Artificial Intelligence, it achieved an accuracy of 99%, enabling especially able individuals to communicate easily [12].

The literature survey encompasses multiple approaches to converting speech to text, dataset creation, and sign language video generation. While existing systems demonstrate promising results in accuracy and effectiveness, there are gaps in comprehensive coverage across different languages, real-time performance, and evaluation methodologies, suggesting a need for further refinement and enhancement in sign language translation systems.

III. METHODOLOGY

This work aims to automatically convert live audio into sign language in real-time. This system works with multilingual audio input. The following block diagram Figure 1 depicts the overall system design of this work.

The whole process can be categorized into three parts. They are as follows:

1. Dataset Preparation:

   The dataset created is a dictionary containing 650 words and their corresponding HamNoSys codes facilitating the identification of HamNoSys notation for any given word. The Hamburg Sign Language Notation System (HamNoSys) is a transcription system for all sign languages that has a direct relationship between symbols and gesture elements such as hand location, form, and movement [13]. HamNoSys notations consist of six factors to indicate the gestures: Handshape, Orientation, Location, Movement 1, Movement 2, Two-handed

2. Speech to Indian Sign Language Text Conversion:
   A. Input speech:
      The input is obtained from the user. It can be in either live speech or recorded audio file format in any language. Live speech is recorded using a microphone as an aid.
   B. Speech recognition engine:
      This module includes several models that are used to transcribe the audio into text. Speech Recognition is one of the modules which converts audio signals to text [14].
      The SR modules embody the following techniques:
      i. Hidden Markov Model (HMM):
      The Hidden Markov Model (HMM) is the formal basis for building a probabilistic model for linear sequence labeling problems [15]. HMM is a statistical model used to display the results of unobserved hidden states using the observed outcomes.
      ii. Neural Networks (NN):
      The neural network is a circuit that is implemented as a replication of the human brain. The circuit works with an input layer, an output layer, and many hidden layers. Recurrent Neural Networks are used for handling sequential data like text and speech.
      iii. Mel-Frequency Cepstral Coefficients (MFCC):
      The MFCC is the factor that forms the Mel-Frequency Cepstral together. In MFC, the frequency bands are evenly spaced on the Mel scale, whereas in cepstrum the frequency bands don't need to be in equal intervals. There is a total of 39 MFCC features, but only 12-13 are considered.

C. Text translation:

The text after conversion may be in any regional language. The text is translated into a standard language for further analysis. Here, English is considered as a standard language. All the text in any other language is translated to English using Google API translator [16].

D. Text to ISL text:

The English text is converted into ISL using Natural Language Processing techniques to normalize the text for further use [17].

Steps involved in NLP:

- Tokenization: The process of breaking operating strings into smaller parts.
- Stemming: Refers to normalizing the words into their base or root form.
- Lemmatization: Ensures that the root word has proper linguistic meaning.
- POS Tags: These refer to the parts of speech of the given word.
- Name Entity Recognition (NER): Detects a named entity be it a person, location, etc.
- Chunking: The process of grouping words or tokens into chunks.

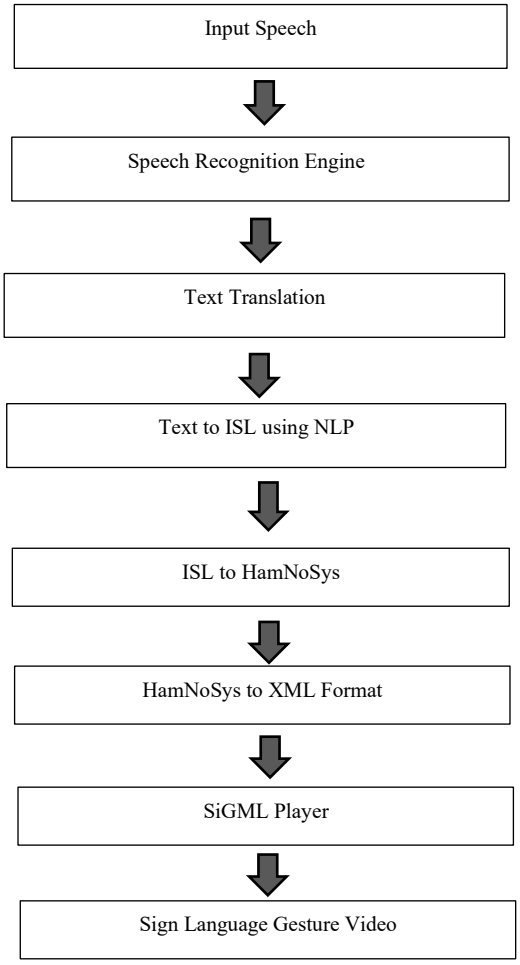

Figure 1 Overall Block Diagram of the proposed work.

3. Sign language Generation:

The following steps depict the process of text-to-video conversion.

i. The ISL text is scanned throughout the dataset dictionary to find the HamNoSys code for the required words of the given sentence.

ii. The HamNoSys code is converted into XML format for the SiGML Player application.

iii. SiGML Player uses XML format to make the application easier for it to display the output.

iv. The corresponding sign language is generated automatically.

SiGML Player:

JASigning is a virtual signing software developed at the University of East Anglia to perform signs in various natural deaf sign languages that use 3D virtual characters. The software uses real-time graphics technology to turn a representation of a sign in SiGML, an XML language based on HamNoSys notation, into animation data for a signing avatar.

IV. RESULTS AND ANALYSIS

The results are categorized into three different stages. The first step is to convert speech to English text using Speech recognition. The second step is to transform the English text to Indian Sign language text meaning it extracts the keywords from the English sentence. The third step is to transmute the Indian Sign language text to the Hamburg Notation System (XML format), which aids SiGML Player to generate 3d animated video output.

1. Speech to English Text:

   The Speech Recognition module utilizes the Hidden Markov Model (HMM) and Mel Frequency Cepstral Coefficients (MFCC) for speech-to-text conversion and the text is translated into English. Accuracy varies by language, with English at 100%, Hindi at 95%, Telugu at 90%, and Tamil and Kannada at 85%. Malayalam poses challenges due to pronunciation nuances. Overall accuracy is approximately 91%, assessed through metrics like Word Error Rate (WER), Character Error Rate (CER), and Accuracy Rate [18].

2. English text to Indian Sign Language text (ISL):

   The English text is translated to ISL text using the Natural Language Processing techniques. In the NLTK library, various methods are involved such as Tokenization, Lemmatization, Post Tags, and Keyword Extraction. Using this a meaningful sentence that only has root words or keywords is created. The accuracy again varies depending on different languages for English, Hindi, and Telugu the accuracy is about 95%, and for Tamil and Kannada, it is about 90%. The total average accuracy, derived from the combination of accuracies across different languages, amounts to approximately 93%.

3. ISL Text to SiGML Player:

   ISL text is mapped to the HamNoSys dataset. If the word exists in the dataset, it maps and gives the HamNoSys text for the word, and the same is repeated for all the words in the sentence in sequential order. Finally, the HamNoSys text of the whole sentence is identified. The HamNoSys text of the whole sentence is changed as a script in an XML format with proper syntax which supports the function of the SiGML Player application using Python. The system is automated so the output will be displayed by just running the program in Python in 20 seconds.

The system has been validated in numerous case studies and couple of the results have been illustrated. The system has undergone validation through multiple case studies, where the resulting output is compared with the original sign language gestures to assess accuracy. The accuracy of the result is determined by comparing the output video with the standard sign language gestures. The system is trained in such a way that it takes input in any language and is loaded with a dataset of 650 words in English. The screenshots are taken during the streaming video display and the number of images illustrated depends on the length of the final animated video.

Analysis Summary:

Accuracy = [(0.5 * ISL_accuracy) + (0.5 * SR_accuracy)] * 100,
where ISL_accuracy = Accuracy of Indian Standard Language text, SR_accuracy = Accuracy of Speech Recognition, and ‘0.5’ stands for the weight of each component as half.

The system achieves speech-to-text accuracies: English (100%), Hindi (95%), Telugu (90%), and Tamil/Kannada (85%), resulting in output in 20 seconds. Moreover, English-ISL translation yields high accuracy (95-100%), ensuring accurate sign gestures for real-time applications.

Parameters:

The parameters provide essential benchmarks for evaluating the system's performance and capabilities. They offer insights into the system's efficiency, flexibility, and accuracy, crucial for ensuring reliable communication and user satisfaction. Understanding these parameters is pivotal for assessing the system's effectiveness in meeting the needs of hearing-impaired individuals. The total number of words in dataset are 650, the average time take to produce the result is 20 seconds, the ISL text accuracy is 85% and the sign language gesture result output is 90%.

Word Error Rate:

Word Error Rate (WER) measures the difference between the recognized text and the ground truth text. It considers substitutions, deletions, and insertions to calculate the error rate. Substitutions refer to the words that are incorrectly recognized and substituted for the correct ones. Deletions are the words in the ground truth text that are missed or deleted in the recognized text. Insertions are words that are incorrectly added or inserted into the recognized text but are not pre ground truth text [18].
Word Error Rate = (Substitutions + Deletions + Insertions)/ (Number of Words Spoken)

Few test cases are:

A. Case Study 1:

Form of input	Live speech
Input language	English
Input in text form	that child is very intelligent, and he always concentrates while reading any book which helps to increase knowledge
ISL format	That Child Very Intelligent He Always Concentrate While Reading Book Which Help To Increase Knowledge.
Output	

B. Case Study 2:

Form of input	Live speech
Input language	Hindi
Input in text form	namsthe hum sabh kal subhah chruch me jaake bible kitaab padenge
ISL format	Hello We Will Go To Church Tomorrow Morning To Read Bible Book.
Output	

Performance Metrics:

For the verification of the proposed work, various parameters such as precision, time taken, and accuracy are validated [19].

Table 1 Performance metrics of the proposed work

LANGUAGE	AVERAGE TIME TAKEN	SPEECH RECOGNITION	ISL TEXT ACCURACY	OUTPUT ACCURACY
Hindi	10-15 sec	95%	95%	95%
English	5-10 sec	100%	95%	97%
Telugu	10-15 sec	90%	95%	92%
Kannada	20 sec	85%	90%	87%
Tamil	20 sec	85%	90%	87%

V. CONCLUSION AND FUTURE SCOPE

In this work, a prototype is proposed for a speech-to-sign language interpreter. This aids people with hearing or speaking disability to understand others effortlessly. Since the procedure works irrespective of the input language, the barrier dissolves and strengthens communication in society. Moreover, this not only helps people with disability but also assists others in learning sign language. This can further be developed by ameliorating various features to enhance the performance and the efficiency of the system depends on the amount of training given to the system with numerous types of inputs. Here, English is considered as the standard language, and since it is a gigantic language the number of words in it is never limited, so the synonyms are to be considered. Furthermore, emotion detection can be added to the process to make the output more sensible in the form of a 3D avatar.

REFERENCES

[1]https://theworld.org/stories/2017-01-04/deaf-community-millions-hearing-india-only-just-beginning-sign
[2]https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss, 2023
[3] Neha Sharma, Shipra Sardana, "A Real Time Speech to Text Conversion System Using Bidirectional Kalman Filter In MATLAB", Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI), 2016
[4] Kanchan Naithani, V.M Thakkar, Ashish Semwal, "English Language Speech Recognition using MFCC and HMM", International Conference on Research in Intelligent and Computing in Engineering(RICE), 2018
[5] Yogitha H. Ghadage, Sushama D. Shelke, "Speech to Text Conversion for Multilingual Languages", International Conference on Communication and Signal Processing, 2016
[6] Yuvraj Grover, Riya Aggarwal, Deepak Sharma, "Sign Language Translation Systems for Hearing/Speech Impaired People: A Review", International Conference on Innovative Practices in Technology and Management (ICIPTM), 2021
[7] Megha Vargheese, Sindhya K. Nambiar, "English to SiGML Conversion for Sign Language Generation", International Conference on Circuit and Systems in Digital Enterprise Technology (ICCSDET), 2018
[8] Khushdeep Kaur, Parteek Kumar, "HamNoSys to SiGML Conversion System for Sign Language Automation", International Multi Conference on Information Processing (IMCIP), 2016
[9] Debashis Das Chakladar, Pradeep Kumar, Shubham Mandal, Partha Pratim Roy, Masakazu Iwamura, Byung-Gyu Kim, "3D Avatar Approach for Continuous Sign Movement Using Speech/Text", 2021
[10] Sandeep Kaur, Mahinder Singh, "Indian Sign Language Animation Generation System", International Conference on Next Generation Computing Technologies, 2015.
[11] Mounika Kanakanti, Shantanu Singh, and Manish Shrivastava. 2023. "MultiFacet: A Multi-Tasking Framework for Speech-to-Sign Language Generation.".
[12] N. J, S. Bhat, R. R. Nair and T. Babu, "Audio to Sign Language conversion using Natural Language Processing," 2022 3rd International Conference on Communication, Computing and Industry 4.0 (C2I4), Bangalore, India.
[13] Alisha Kulkarni, Archith Vinod Kariyal, Dhanush V., Paras Nath Singh, "Speech to Indian Sign Language Translator", International Conference on Integrated Intelligent Computing Communication and Security (ICIIC), 2021.
[14] Lalitha S., Gupta D., "Investigation of Automatic mixed-lingual affective state recognition system for diverse Indian languages", Journal of Intelligent and Fuzzy Systems, 2021, pp. 1-10.
[15] Roshan S. Sharma, Sri Harsha Paladugu, K. Jeeva Priya, Deepa Gupta, "Speech Recognition in Kannada using HTK and Julius: A Comparative Study", International Conference on Communication and Signal Processing(ICCSP), 2019
[16] Deepa Gupta, K. Vani, Charan Kamal Singh, "Using Natural Language Processing techniques and fuzzy-semantic similarity for automatic external plagiarism detection", International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2014
[17] Afnan Zafar, Hira Gull, Uzma Farooq, Mohd Shafry Mohd Rahim, Adnan Abid, "Teaching English and Science to the Deaf Students", International Conference on Innovative Computing (ICIC), 2021
[18] Papadimitriou, E., Patsios, C., Siozios, K., & Tzouvaras, V. Towards Translating Natural Language to Sign Language Glosses using Neural Machine Translation with Monolingual Data. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 16-21), 2023
[19]https://waywithwords.net/resource/speech-recognition-systems-performance/, 2024