INTERNATIONAL JOURNAL FOR INNOVATIVE RESEARCH IN MULTIDISCIPLINARY FIELD
ISSN(O): 2455-0620 [Impact Factor: 9.47]
Monthly, Peer-Reviewed, Refereed, Indexed Journal with IC Value : 86.87
Volume - 10, Issue - 5, May - 2024

DOIs:10.2015/IJIRMF/202405004 --:-- Research Paper / Article / Review

Sign & Speak Companion

[1]Satyaprakash Tiwari, [2]Avadhut Mulaye, [3]Nomaan Khan, [4]Deepti V. Chandran
[1]BE Students, Computer Engineering Department, Smt. Indira Gandhi College of Engineering, Mumbai, India
[2]BE Students, Computer Engineering Department, Smt. Indira Gandhi College of Engineering, Mumbai, India
[3]BE Students, Computer Engineering Department, Smt. Indira Gandhi College of Engineering, Mumbai, India
[4] Assistant Professor, Computer Engineering Department, Smt. Indira Gandhi College of Engineering, Mumbai,
Email – [1]satyaprakash4253@gmail.com , [2]avadhutmulaye@gmail.com, [3]nomaank750@gmail.com
[4]dvcnr@yahoo.co.in

Abstract: *Sign & Speak Companion represents a significant contribution to the field of assistive communication technology. This research paper presents a comprehensive overview of the development and implementation of the Sign & Speak Companion project. Most people in India are not aware of Indian Sign Language (ISL), this makes things difficult for these people. This application acts as companion for hearing/speech impaired people, by giving features such as real time speech-to-sign, sign-to-speech/text. Detecting motion gestures in ISL is pretty difficult for a simple ANN, To detect these, we have implemented an FST (Finite State Transducer) to map various gestures and make sense out of it. We have used Google's MediaPipe for hand land marking and Python for model training. Additionally, the integration of JavaScript via the Python EEL framework enhances user experience through an intuitive interface. This paper discusses the technical innovations, methodologies, and outcomes of the project, highlighting its potential to ease the lives of individuals with hearing and speech impairments. The Sign & Speak Companion project aims to break down barriers and promote inclusivity in an increasingly interconnected world.*

Key Words: *Google's MediaPipe, Python, Motion Gesture Recognition, Hand Landmarking, Model Training, Accessibility Technologies, ISL gesture recognition, Neural Network..*

1. INTRODUCTION:

In India, where the diversity of languages and cultures is vast, addressing the needs of those who communicate using Indian Sign Language (ISL) presents a significant challenge. Traditional methods of communication support often fall short in meeting the dynamic and multifaceted requirements of this community.

According to recent statistics from the World Health Organization (WHO), an estimated 63 million people in India suffer from significant hearing loss, with around 7.5% of the population experiencing some form of speech or hearing impairment. Despite the magnitude of this issue, accessibility to adequate communication tools and resources remains limited. To bridge this gap, our research project aims to develop a comprehensive application that facilitates seamless communication for people with speech and hearing impairments. This application uses the power of machine learning and natural language processing to convert speech to sign, sign-to-speech, speech to text, and text to speech, all based on Indian Sign Language. By leveraging state-of-the-art technology, our project endeavors not only to provide a practical tool for daily communication but also to foster a more inclusive learning environment. This research paper delves into the development process, functionalities, and potential impact of our web-based application, offering insights into its implications for education, accessibility, and inclusivity in the Indian context. Through this endeavor, we aim to contribute to the advancement of assistive technology and empower individuals with diverse communication needs to participate fully in society.

2. LITERATURE REVIEW:

A. *M. M. Chunduru, M. Roy, D. R. N. S and R. G. Chittawadigi, "Hand Tracking in 3D Space using MediaPipe and PnP Method for Intuitive Control of Virtual Globe," 2021 IEEE 9th Region 10 Humanitarian Technology Conference (R10-HTC), Bangalore, India, 2021 [1]*

There is a need for new interactive techniques to engage students with academic content. One of the hardest concept students deal with is using the Atlas and maps to understand the Geography, Social Sciences, and History of the world. The current alternative to the Atlas is accessing virtual globes on computers but controlling the globe is done using a mouse or touchscreen. The interaction may turn monotonous with usage. To address this problem, the authors propose a methodology to track a hand using a camera feed of a computer, while further using hand gestures to manipulate a virtual globe. Google's MediaPipe library is employed, along with OpenCV's pose estimation functions that use PnP (Perspective-n-Point) method. Later, an interactive globe environment has been implemented on a web browser using Cesium which can be commanded to perform rotate, pan and zoom on virtual globe and 2D maps as per the user's hand position and gestures. Results regarding the accuracy of the hand tracking system and robustness of control of virtual globe are also reported here.[1]

B. *N. Zhao and H. Yang, "Realizing speech to gesture conversion by keyword spotting," 2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP), Tianjin, China, 2016 [2]*

The paper proposed a method to realize a speech-to-gesture conversion for communication between normal and speech-impaired people. Keyword spotting was employed to recognize the keywords from input speech signals. At the same time, the three-dimensional gesture models of keywords were built by 3D modeling technology according to the 'Chinese sign language'. The speech-to-gesture conversion was finally realised by playing the corresponding 3D gestures with OpenGL from the results of keyword spotting. Tests show that the realized keyword spotting achieves 90.1% of average recognition rate on letters and numbers. The converted gestures obtain 4.4 of the mean opinion score. Therefore, the proposed method can be applied to the communications between normal and speech-impaired people.[2]

C. *C. Navneet Upadhyay, Abhijit Karmakar, Speech Enhancement using Spectral Subtraction-type Algorithms: A Comparison and Simulation Study, Procedia Computer Science, Volume 54,2015 [3]*

The spectral subtraction is historically one of the first algorithms proposed for the enhancement of single channel speech. In this method, the noise spectrum is estimated during speech pauses, and is subtracted from the noisy speech spectrum to estimate the clean speech. This is also achieved by multiplying the noisy speech spectrum with a gain function and later combining it with the phase of the noisy speech. The drawback of this method is the presence of processing distortions, called remnant noise. A number of variations of the method have been developed over the past years to address the drawback. These variants form a family of spectral subtractive-type algorithms. The aim of this paper is to provide a comparison and simulation study of the different forms of subtraction-type algorithms viz. basic spectral subtraction, spectral over-subtraction, multi-band spectral subtraction, Wiener filtering, iterative spectral subtraction, and spectral subtraction based on perceptual properties. To test the performance of the subtractive-type algorithms, the objective measures (SNR and PESQ), spectrograms and informal listening tests are conducted for both stationary and non-stationary noise types at different SNRs levels. It is evident from the results that the modified forms of spectral subtraction method reduce remnant noise significantly and the enhanced speech contains minimal speech distortion.[3]

D. *D. T. Bohra, S. Sompura, K. Parekh and P. Raut, "Real-Time Two Way Communication System for Speech and Hearing Impaired Using Computer Vision and Deep Learning," 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2019 [4]*

Sign Language is the most expressive form of communication for speech and hearing-impaired people to communicate with a normal person, but a normal person cannot understand sign language. So, in order to break this barrier of communication there needs to be a system that can enable conversion of sign language to voice or text and voice or text to sign language and do it in real time. The systems that currently exist are not real time, do not facilitate two-way communication, require static surrounding conditions or have low recognition accuracy. There exist systems that have good accuracy but require external hardware like gloves [3] which increases the cost. Our contribution to solving this problem consists of a Sign Language Communication System. It is a real-time communications system built using the advancements in Image Processing, Deep Learning, and Computer Vision that provides real-time sign language to text and text-to-sign language conversion. The project is software-based which can be installed on any computer with good specifications. It is also a two-way communication system that allows not only speech and hearing impaired to communicate with normal people but also the other way around. The primary goal of our system is to enable hearing and speech impaired people to communicate with people that are not disabled in real time by interpreting alphabets,

numbers, and words in the Indian sign language. The system is able to predict 17600 test images in 14 seconds with an average prediction time of 0.000805 seconds with an accuracy of 99%.[4]

E. *Sanyam Jain, ADDSL: Hand Gesture Detection and Sign Language Recognition on Annotated Danish Sign Language. arXiv:2305.09736. https://doi.org/10.48550/arXiv.2305.09736 [5]*

For a long time, detecting hand gestures and recognizing them as letters or numbers has been a challenging task. This creates communication barriers for individuals with disabilities. This paper introduces a new dataset, the Annotated Dataset for Danish Sign Language (ADDSL). Annotations for the dataset were made using the open-source tool LabelImg in the YOLO format. Using this dataset, a one-stage object detector model (YOLOv5) was trained with the CSP-DarkNet53 backbone and YOLOv3 head to recognize letters (A-Z) and numbers (0-9) using only seven unique images per class (without augmentation). Five models were trained with 350 epochs, resulting in an average inference time of 9.02ms per image and a best accuracy of 92% when compared to previous research. Our results show that modified model is efficient and more accurate than existing work in the same field.[5]

3. SCOPE:

The scope of this research project includes designing and developing an efficient, reliable, and user-friendly system to assist hearing and speech-impaired individuals in their daily interactions. The system will utilize MediaPipe's capabilities for real-time gesture recognition to assist users. An application will be designed and developed compatible with MediaPipe, ensuring it is user-friendly and accessible to individuals with hearing and speech impairments. The application will incorporate features for speech-to-text and text-to-speech conversion, allowing users to communicate through written or spoken language. Additionally, a Python-based model will be developed to interpret the detected landmarks and translate them into corresponding text or speech output for the user. This research project aims to address the challenges faced by hearing and speech impaired individuals by providing them with a reliable and efficient assistive system that can enhance their independence and involvement in their surroundings effectively.

4. DESIGN:

A. Flowchart of the Application

The application can convert spoken speech to sign gestures/text as well as sign to speech. In sign to speech, the sign gestures will be converted into speech. To detect the points on fingers, Google's Mediapipe library is used, after that a neural network classifies these points into 150 gestures on which it was trained. The resulting sign gesture is then outputted as speech. Here we are trying to mimic human speech, which is generated by the sign gestures, solving the problem that not many people know ISL. The next part is speech to sign/text, here we let the user know, what people are talking about. The speech is taken as an input and is converted into text, after that depending on their preference, the user can see the corresponding sign gesture or text. The internal process is shown in the fig. 1.

B. Model Creation

Creating a neural network model for sign language detection is a complex task, the initial intuition is to use object detection to classify/detect all the gestures, but object detection takes a huge and diverse training dataset. For these reasons, we used Mediapipe to detect the hands, face, and finger points. After that we created an ANN model which classifies based on these points. We can say that actual object detection is being carried out by Mediapipe, and based on the output of it, the ANN is trying to classify the gesture. The main advantage we gained due to this was that we do not need huge training dataset for each gesture, 150-200 data points for each gesture is sufficient. The ANN training time is significantly reduced since ANN is not trying to classify from an Image. Refer Fig.2 for model creation flowchart.

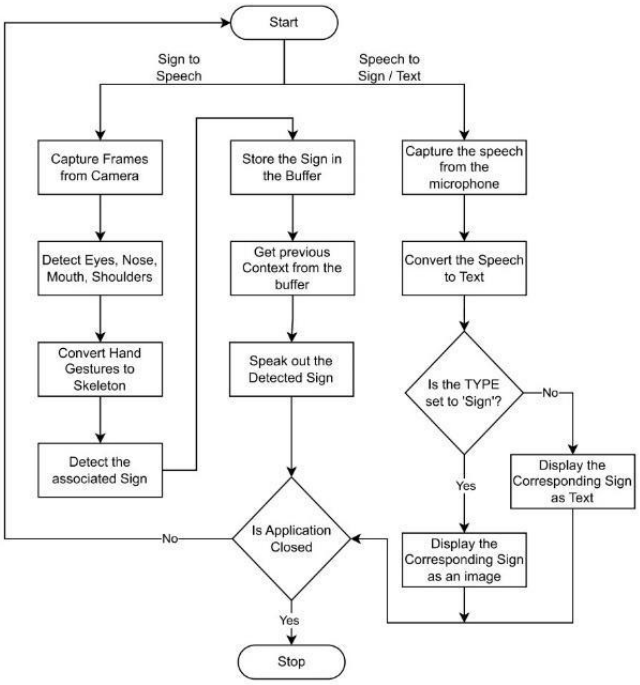

Fig. 1 Flowchart of the Application

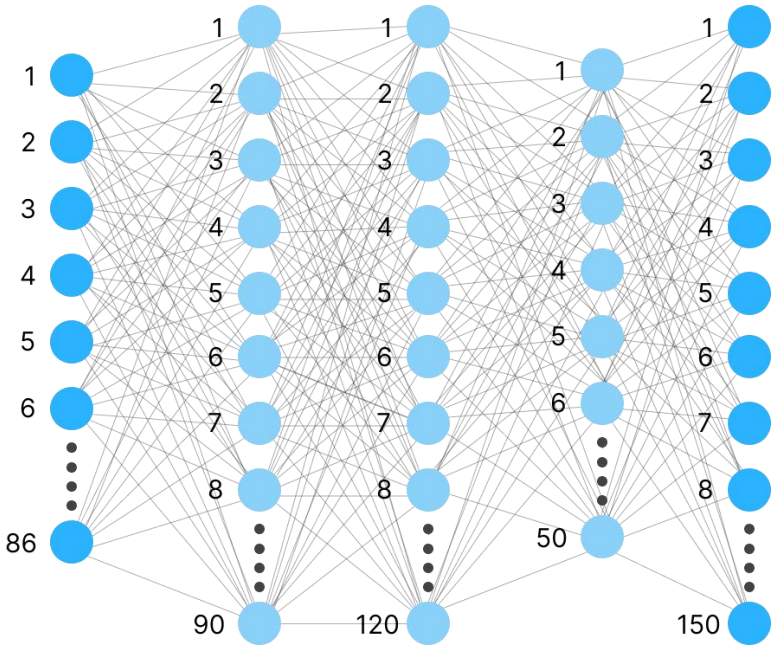

Fig. 2 Neural Network Model Structure

4. IMPLEMENTATION:

A. Sign Language Detection ANN Model

A neural network is a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain. It is a type of machine learning process, called deep learning, that uses interconnected nodes or neurons in a layered structure that resembles the human brain. It creates an adaptive system that computers use to learn from their mistakes and improve continuously. Thus, artificial neural networks attempt to solve complicated problems, like summarizing documents or recognizing faces, with greater accuracy. As stated earlier, Mediapipe is used between the Camera module (open cv) and the ANN model. The task of media pipe is to detect hands, fingers and face and output their (x,y) coordinate values. In total we have 43 points i.e. 21 points on each hand and 1 point on the face. The Input

layer of the ANN has 86 nodes (43 * 2 for x and y coordinated). The output layer consists of 150 nodes (the total number of sign gesture we are classifying). There are 90, 120, 50 nodes in the hidden layers respectively. ANN model is shown in fig. 2

B. Motion gesture detection

Most of the already existing ISL detection models fail to recognize motion gesture (gestures which are not static). Since most of the gestures in ISL are motion gestures, these models are not very useful. We divide the motion gestures into various break point and instead of labelling the entire gesture, we label these individual parts. For example, “accident” is a motion gesture. So, we can break it into 2 parts “accident 1” and “accident 2”. The detected parts are stored in a BUFFER. Now the next task is to read this gesture from the array/BUFFER. We read the BUFFER in reverse order and create a FST (Finite State Transducer) as shown in fig.3. The FST is such that if “accident 2” occurs after “accident 1” the gesture is “accident”. The Alphabet” h” is also a motion gesture. We can say “h” occurred if “h 2” occurs after “h 1” in the BUFFER. The sample FST for 9 motion gestures is shown in fig.3. We do not require FST for static gestures.

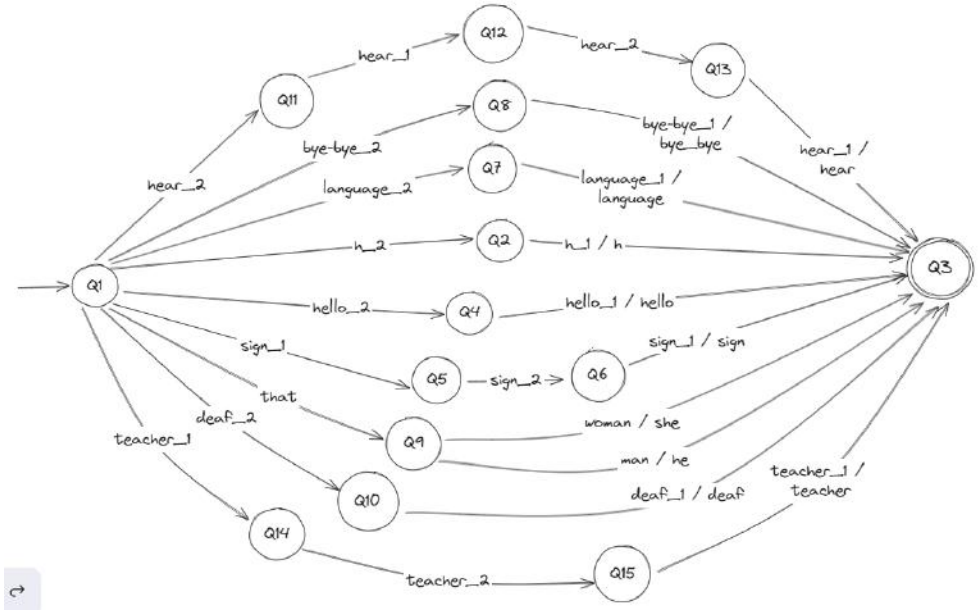

Fig. 3 FST for sample 9 gestures

C. Desktop View

The Desktop application is where the Hearing & Speech Impaired User will interact. The desktop app is created in Python and front-end technologies such as JavaScript, HTML, CSS using eel framework. This framework helps us to create a desktop application where the front-end can communicate with python without fast API. The communication is much faster. Python is used to create model, run Mediapipe over Open-CV2 and detect sign gesture. HTML, CSS, JavaScript is used to create User Interface. In Desktop view the user can create/register an account and use that account to store relevant information/data. All user preferences will be asked at the time of registration. In the Sign to speech, the task of sign to speech is implemented i.e. first detecting the landmarks sing Mediapipe, then detecting the gesture based on these landmarks/points. After that the resulting text will be outputted as speech from a mobile device. The idea is that, when the student is saying something through sign language, the output should be at teacher’s mobile device.

5. DISCUSSION:

The Approach proposed by T. Bohra and Team [1] detect static gestures for 26 alphabets, 10 digits and 4 words, which is not much but the model’s efficiency can be increased by using SVM or YOLO. Angela C. Caliwag, Han-Jeong Hwang and Team [3] proposed the merging approach, the disadvantage of using this approach for ISL is stated above in review of literature. Kumund Tripathi and Team performed key frame extraction, i.e. the detection will be conducted when a gesture change motion is detected and not for all the frames.

While Considering Different Approaches and their drawbacks, the one approach that seemed possible for dynamic gesture classification was dividing the motion gesture into parts and while at the time of classification, if all the parts are in the buffer, then it is considered, that gesture was performed.

The conversion of Speech to sign will include 2 parts, the part one is to convert speech to text and the second part is to check in a table if the word is present and render the associated 3D model. As this application will be for a Hearing and speech impaired student, the possibility for recording the spoken words by other normal people can also be considered. At time of some Disaster, common people can hear the noises and alarm bells but it not easy for the hearing-impaired individual to react to these situations. Therefore, monitoring of background noise and alarms is a crucial component. If the mean frequency is above a threshold value, then the mobile handset can vibrate according to that value, the vibration can be low for low frequencies and high for high frequencies.

6. RESULT:

The Snapshots of the Desktop Application are show below in fig.4, 5 & 6. The result also involves the confusion matrix of the trained ANN Model.

1. The Login and Sign Up page

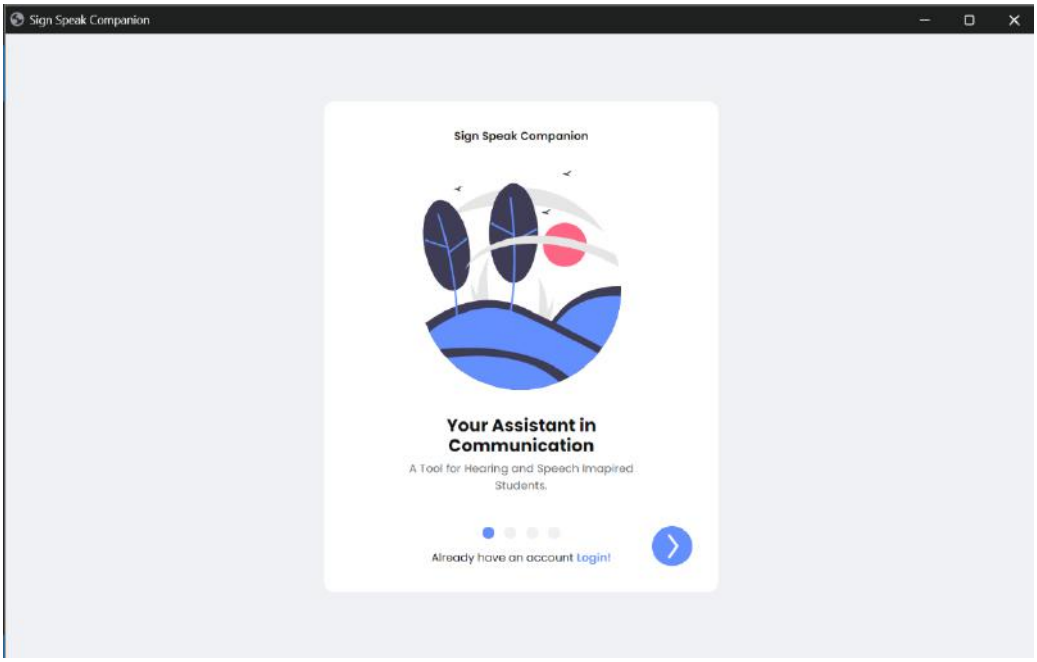

Fig. 4 Desktop Home Page

2. Sign-to-speech Conversion

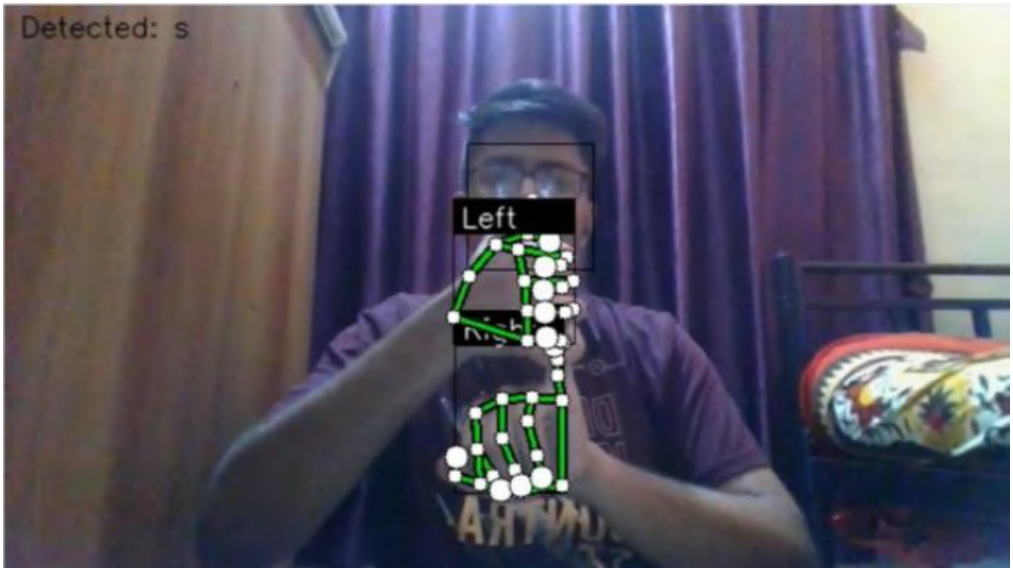

Fig. 5 Sign to Speech (Landmarking)

3. Speech-to-Sign Conversion

Fig. 6 Sign to Speech (Interpretation)

7. FUTURE SCOPE:

The enhancement of the system's language capabilities is imperative to broaden its accessibility and usability. Specifically, expanding beyond the existing support for Indian Sign Language (ISL) to include a more comprehensive range of sign languages can significantly improve the inclusivity of the system. This expansion would necessitate the integration of diverse linguistic datasets, which would be pivotal in capturing the nuances and variations inherent in different sign languages, thereby fostering wider user accessibility.

Additionally, the augmentation of system performance is vital for improving its overall efficacy and reliability. One approach to achieve this is through the incorporation of an extensive and diversified dataset that encompasses a broad spectrum of sign language gestures and speech patterns. Such a dataset would be instrumental in enhancing the system's accuracy and proficiency in recognizing and interpreting various sign language gestures and speech nuances. The development of a platform-agnostic application is also essential to ensure accessibility for a broader user base. By creating an application that is compatible with mobile devices, tablets, and various operating systems, the system can reach a wider audience, thereby increasing its impact and usability. Moreover, the integration of voice diarization functionalities is a promising avenue for facilitating more natural and fluid user interaction. This feature would enable the system to distinguish between multiple speakers within a conversation, thereby enhancing the user experience and making the interaction with the system more intuitive and user-friendly.

8. CONCLUSION :

In summary, the Sign & Speak Companion project represented a pivotal endeavor in the realm of assistive communication systems for individuals with hearing and speech impairments. Throughout this project, we created an application that helps hearing and speech impaired individuals communicate.

At the core of our system lay Google's MediaPipe, renowned for its robust hand landmarking capabilities. Using this technology, we enabled accurate gesture recognition, allowing users to convey thoughts and emotions through gestures in sign language. This aspect not only facilitated expressive communication but also empowered users to interact meaningfully with their communities.

Moreover, our used Python for model training. Through extensive optimization, we crafted a model that interpreted sign language gestures with good accuracy and reliability, meeting users' expectations.

The integration of JavaScript via the Python EEL framework enhanced the user experience. We created an intuitive interface that enabled users to navigate seamlessly through the Sign & Speak Companion ecosystem, ensuring accessibility for diverse abilities.

Reflecting on this project, the Sign & Speak Companion held immense promise in revolutionising the lives of people with hearing and speech impairments. By providing inclusive communication, empowerment, and independence, our solution serves as a foundation for more equitable future. Moving forward, we remained committed to advancing assistive communication technologies, breaking down barriers for individuals to thrive in an interconnected world.

REFERENCES:

Journal Papers:

1. M. M. Chandra, S. Rajkumar and L. S. Kumar, "Sign Languages to Speech Conversion Prototype using the SVM Classifier," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, India, 2019, pp. 1803-1807, doi: 10.1109/TENCON.2019.8929356.
2. N. Zhao and H. Yang, "Realizing speech to gesture conversion by keyword spotting," 2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP), Tianjin, China, 2016, pp. 1-5, doi: 10.1109/ISCSLP.2016.7918458.
3. Navneet Upadhyay, Abhijit Karmakar, Speech Enhancement using Spectral Subtraction-type Algorithms: A Comparison and Simulation Study, Procedia Computer Science, Volume 54, Pages 574-584, ISSN 1877-0509,https://doi.org/10.1016/j.procs.2015.06.066. (https://www.sciencedirect.com/science/article/pii/S1877050915013903)
4. T. Bohra, S. Sompura, K. Parekh and P. Raut, "Real-Time Two-Way Communication System for Speech and Hearing-Impaired Using Computer Vision and Deep Learning," 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2019, pp. 734-739, doi: 10.1109/ICSSIT46314.2019.8987908.
5. Sanyam Jain, ADDSL: Hand Gesture Detection and Sign Language Recognition on Annotated Danish Sign Language. arXiv:2305.09736. https://doi.org/10.48550/arXiv.2305.09736
6. R. Kumar, M. Gupta and S. R. Sapra, "Speech to text Community Application using Natural Language Processing," 2021 5th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2021, pp. 1-6, doi: 10.1109/ISCON52037.2021.9702428.
7. V. Chunduru, M. Roy, D. R. N. S and R. G. Chittawadigi, "Hand Tracking in 3D Space using MediaPipe and PnP Method for Intuitive Control of Virtual Globe,"2021 IEEE 9th Region 10 Humanitarian Technology Conference (R10-HTC), Bangalore, India, 2021, pp. 1-68)
8. Caliwag, A.C.; Hwang, H.-J.; Kim, S.-H.; Lim, W. Movement-in-a- vVdeo Detection Scheme for Sign Language Gesture Recognition Using Neural Network. Appl. Sci. 2022, 12, 10542.
9. Savas, Serkan & Erguzen, Atilla. (2023). HAND GES TURE RECOGNITION WITH TWO STAGE APPROACH USING TRANSFER LEARNING AND DEEP ENSEMBLE LEARNING. W. Movement-in-a-Video Detection Scheme for Sign Language Gesture Recognition Using Neural Network. Appl. Sci. 2022, 12, 10542. https://doi.org/10.3390/app122010542
10. Kumud Tripathi, Neha Baranwal G.C. Nandi, Continuous Indian Sign Language Gesture Recognition and Sentence Formation, Procedia Computer Science, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2015.06.060
11. H. Muthu Mariappan and V. Gomathi, "Real-Time Recognition of Indian Sign Language," 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), Chennai, India, 2019, pp. 1-6, doi:10.1109/ICCIDS.2019.8862125.
12. Paula Escudeiro, Nuno Escudeiro, Rosa Reis, Jorge Lopes, Marcelo Norberto, Ana Bela Baltasar, Maciel Barbosa, Jos´e Bidarra, Virtual Sign – A Real Time Bidirectional Translator of Portuguese Sign Language, Procedia Computer Science, Volume 67, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2015.09.269.
13. Caixia Wu, Chong Pan,YufengJin, Shengli Sun, and Guangyi Shi Shaoxing, "Improvement of Chinese Sign Language Translation System based on Collaboration of Arm and Finger Sensing Nodes", At The 6th Annual IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems June 19- 22, 2016
14. Srujana Gattupalli, Ashwin Ramesh Babu, James Robert Brady, Fillia Makedon, Vassilis Athitsos. arXiv:1804.01174, Towards Deep Learning based Hand Keypoints Detection for Rapid Sequential Movements from RGB Images. https://doi.org/10.48550/arXiv.1804.01174
15. Daniil Osokin, Real-time 2D Multi-Person Pose Estimation on CPU: Lightweight OpenPose. follow the bottomup approach from OpenPose, with proposed network design and optimized post-processing code the full solution runs at 28 frames per second. arXiv:1811.12004v1. https://doi.org/10.48550/arXiv.1811.12004, Thu, 29 Nov 2018 08:05:05.
16. Liu, Y. Liu, Y. Wang, V. Prinet, S. Xiang and C. Pan, "Decoupled Representation Learning for Skeleton- Based Gesture Recognition," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 5750-5759