International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Readable, A Lip Reading and Sign Language Interpretation Machine Learning and Deep Learning Model

Mr. Prajyot Kawatkar[1], Mr. Sujal Wankhede[2], Mr. Dev Katre[3], Mr. Rafe Sheikh[4], Prof. Leelkanth Dewangan[5]

[1,2,3,4]Student, Information Technology, G.H.Raisoni College of Engineering
[5]Assistant Professor, Information Technology, G.H.Raisoni College of Engineering

ABSTRACT:
Silent communication involving lip movements and sign language continues to be greatly under-used in popular media and new media. This paper describes ReadAble, a two-part AI system consisting of a Lip Reading module and a Sign Language Detection module that transforms silent video into output that is made accessible via text and synthesized audio output (lip reading) and translated text output (sign language). The Lip Reading module first locates the lip region of the video, tracks the spatio- temporal features of the lip movement, and maps the resulting movement to phonemes/words, producing both text and speech output. The Sign Language Detection module detects the hand and body movements (e.g. ASL, ISL in this case) and interprets the movement sequences from arbitrary sign language expressions into signed sentences or phrases, before producing them in the form of subtitles. The implementation of the system was done in Python (Jupyter Notebook) utilizing multiple state-of-the- art models (3D CNN, LSTM / Transformer, MediaPipe / YOLO or SSD), utilizing video data captured from user or recorded feed.Overall, we recognize several remaining challenges to deploying the system in real world noisy, occluded, and continuous settings; and we provide some thoughts on the roadmap for getting us from a lab and into a real-world setting with a better product.

Keywords - Lip Reading, Sign Language Detection, Computer Vision, Deep Learning, Accessibility, Human- Computer Interaction

INTRODUCTION

Communication is a vital human need, but a lot of our technology today is based on audio channels. For any people who are deaf, hard of hearing, or non-verbal, and any time there is no available audio channel or the audio is corrupted (silent video, noisy room, etc.), this reliance on audio is a major barrier. Traditional solutions have depended on either manually captioning, or human sign language interpreters[8]. These solutions are either expensive, not always available, or slow to set-up, especially in live or informal settings.

Recent developments in computer vision, deep learning, and natural language processing offer new opportunities to automate the interpretation of silent communication - both lip movements and sign languages[9]. Lip reading (also known as Visual Speech Recognition, VSR) aims to understand facial and

lip motion to determine spoken words or phonemes, and sign language recognition (SLR) utilizes hand/arm/body gestures and non-manual signs (face and posture) to produce spoken or written language[4]. SLR and VSR systems have been developed previously, but typically only one modality (either lip reading or sign language) is focused on, or are limited to a static gesture, limited vocabulary, limited backgrounds, or limited intense lighting.

This study, ReadAble, presents an integrated system consisting of two components, a Lip Reading component responsible for recognizing silent speech, generating both textual and audio representations from the speech; a Sign Language Detection component which detects gestures (i.e., manual + non-manual), connects those gestures to textual or subtitle representations[7]. A primary goal of the system is to build a system that is accurate, real-time (or near real-time), robust to variance in lighting, video quality, background, and sign language dialect; and ultimately usable in real-life settings (e.g., education, live streaming, accessibility apps). We create the system in Python, leveraging both open-source libraries and pre- trained deep learning models (and custom models), evaluate its performance using video input (from the YouTube video used in the introduction), while evaluating accuracy, latency, and robustness under variation, and create sample output[3].

LITERATURE SURVEY

Recent developments in computer vision and deep learning research have played a key role in advancing both lip reading and sign language recognition. Both techniques aim to increase the accuracy, speed, and robustness of systems that decode silent communication. In the literature, studies examined spatio-temporal modeling, convolutional neural networks (CNNs), recurrent neural networks (RNNs), Long Short-Term Memory (LSTM) models, transformers, and real-time object detection algorithms such as YOLO. In general, the studies included in this field have promoted efforts to narrow the communication gap that exists for individuals who are deaf or hard of hearing or have non-verbal characteristics.

In [1] this study, a lightweight real-time sign language identification model is proposed using optical flow features derived from human position estimates. The model performs with an accuracy of 80% with a linear classifier and 91% with a recurrent model under 4 ms.

This [2] paper explores the use of MobileNetV2, a lightweight and efficient convolutional neural network (CNN), for real-time identification of sign language. The study uses transfer learning to improve recognition accuracy by fine-tuning a previously trained model on a particular dataset of sign language. The main goal is to develop an accessible communication tool that will enable people with hearing impairments to engage with the non signing world more easily. The study intends to offer a scalable and affordable solution that can be used on smartphones, tablets, and embedded devices by combining deep learning and computer vision. In order to promote a more inclusive society where individuals with hearing impairments have equal chances in educational and professional settings, the research also addresses possible applications in public services and education.

This [3] in contrast to current models that identify individual signals, research provides a continuous sign language recognition system using Long Short-Term Memory (LSTM) networks. The technology facilitates natural and fluid communication by continuously monitoring and interpreting hand gestures, facial expressions, and body postures. The MediaPipe Holistic pipeline is a computer vision framework that the researchers employ to efficiently identify and decipher multi-modal features of sign language. The model's efficacy in identifying consecutive motions is demonstrated by its 88.23% accuracy rate, which was attained by training on an Indian Sign Language dataset. The development of real-time video-based

sign language interpreters, which could be included into smart assistants, video conferencing apps, and accessibility resources for the hard of hearing, is especially pertinent to this study.

This [4] study uses the YOLO-v9 model, a cutting-edge object detection framework, to study real-time American Sign Language (ASL) identification. Real-time sign recognition applications benefit greatly from YOLO (You Only Look Once) models, which are well known for their quick and effective object detection capabilities. The study highlights the improvements in accuracy and speed in the most recent YOLO-v9 architecture by offering performance insights and a comparison analysis between various YOLO versions. This research's real-time inference capabilities is one of its main achievements; it makes it possible for interactive applications, live broadcasts, and accessibility tools to identify ASL with ease. The study shows notable gains over earlier models by refining YOLO-v9 for sign language detection, providing a high-performance and scalable solution for gesture-based communication systems..

This [5] work introduces a thorough deep learning method for translating and recognizing sign language using the YOLOv5 architecture and a specially created dataset. By converting identified motions into meaningful text, our approach surpasses traditional sign recognition systems and successfully closes the communication gap between signers and non-signers. The model is among the most accurate systems created for real-time applications, with an outstanding mean Average Precision (mAP) ranging from 92% to 99% over 15 distinct sign classes. The study highlights how crucial high-quality training datasets and data augmentation methods are to enhancing the model's resilience. It also investigates voice synthesis integration, which improves accessibility for the deaf and mute community by enabling the conversion of recognized signals into spoken words.

This [6] article reviews the development of lip reading technology comprehensively. It provides a detailed examination of the standard computer vision and more recent deep learning methods. Specifically, the authors discuss various model architectures, such as CNN, RNN, and combinations thereof. The survey elucidates a number of key issues, including speaker independence, lighting conditions, and homophenes.Furthermore, they outline the large improvements in performance compared to previous methods due to deep learning. They also emphasize preprocessing and data augmentation as important model accuracy factors. The summary states the positive human- computer interaction and assistive technology possibilities that lip reading offers. Overall, this paper is an important foundational paper for new researchers to access in the field.

This research [7] project aims to utilize deep learning models to create a practical lip-reading system.The architecture achieves spatial feature extraction from mouth areas using Convolutional Neural Networks (CNN).Recurrent Neural Networks (RNNs) are applied to model the sequential patterns of the visual data.The spatio- temporal features will be learned by the model and mapped directly to a character or word.The document describes their pipeline for preprocessing the data, involving face detection and mouth ROI extraction.They report on their model’s ability to recognize a small vocabulary from visual speech data.The authors discuss the challenges of overfitting and computational resources that they faced.However, they provide a practical proof-of- concept for voicing visual speech recognition automated.

This [8] study broadens the possibilities of lip reading to include cross-lingual translation features.The suggested system begins by extracting text from a silent video using lip readings.Then, the extracted text is translated into another target language.This deep learning model is specifically structured to accommodate the sequence-to- sequence nature of both tasks.The aim is to eliminate language barriers in addition to auditory barriers. The paper demonstrates the recognition and translation performance as two

modules.Additionally, the paper tackles the dual complexity of errors propagating from recognition to translation.Ultimately, the findings strengthen the role of lip reading as an enabling technology for multi-lingual communication literacy.

In [9] a system designed with the TensorFlow Object Detection API and SSD MobileNetV2 to recognize both alphabet-level gestures and lip motion through transfer learning. This consisted of creating TensorFlow (TF) datasets into TFRecord files to enable high-speed training and contain datasets for better storage management. The architecutre provided modern computer vision methods such as the FPN-lite and the use of focal loss so the model was highlyaccurate and robust to different test conditions. Its light weight design, size and level of resource usage makes it a prime candidate for implementation with embedded systems or into mobile applications.

At last, in [10], Kumar, Rani, and Chaudhari (2024) proposed a recognition model that combines VGG16 with self-attention and places its main emphasis on the real- time performance. The study had a preprocessing pipeline that changed sign and lip images into tensors, which were split into a training set and a validation set. The authors used the Adam optimizer and a cross entropy loss function during training to ensure efficient convergence. Evaluation included performance metrics of accuracy, precision, recall, F1-score, and confusion matrices to include more detail regarding the classification results. The model showed strong recognition rates across multiple gestures and lip movements in real-time situation, and achieved high levels of accuracy in evaluation of these gestures. One of the main contributions of the study was the emphasis on accessibility and the real-world, practical applications of the research had meaningful value in helping disabled communities. Although the model had stated limitations in computational requirements, limiting its deployment in low-resource devices, it nevertheless underscores the modeled hybrid design and the benefit of leveraging attention mechanisms into recognition models to enhance acknowledgement and usability.

PROPOSED METHODOLOGY

In order to develop the intended methodology, the first step is explicitly determining the problem being addressed, which is a lack of communication access for individuals who are d/Deaf, hard of hearing, or non-verbal, in a world that is largely built for hearing communicators. In order to address this layered and multifaceted problem, ReadAble will be developed as an integrated solution in an overall system of two parallel, co-processing, modules: a Real- Time Lip Reading Module, and a Real-Time Sign Language Detection Module. The two modules process simultaneously in a single 'deep-learning' model to create a comprehensive, accessible, translated experience.

The system being developed relies on state-of-the-art computer vision and deep learning pipelines to infer meaning from non-verbal communication indicators and aspects. The Lip Reading Module will work in real-time with silent stream video. The module will utilize a spatio- temporal deep learning model, such as a 3D Convolutional Neural Network (CNN), with a Bidirectional LSTM architecture, to capture and analyze the subtle and nuanced movements of the speaker's lips in a real-time framework over temporal dimensions of unfolding speech. The model will extract visual features in each frame of in-process video stream, sequence the visual features, and decode them into phonemes or words, using a Connectionist Temporal Classification (CTC) loss function for output as live synchronized subtitles.

At the same time, the Sign Language Detection Module detects hand-based communication. This module follows a two-stage real-time detection and classification pipeline. In the first stage, it uses a fast and efficient human pose estimation model, such as MediaPipe Holistic, to create an accurate landmark

detection on the user's hands, body, and face. This process identifies and captures the landmark coordinates that define hand shapes, orientation, and movement trajectories in the user's video feed. In the second stage, the extracted spatial and temporal features are processed to classify gestures using a gesture classification model based on deep learning principles. For isolated signs, a fast optimized CNN architecture (e.g., MobileNetV2) is used. For recognizing continuous sign language, a Bidirectional LSTM or a Transformer-based model learns to sequence the landmark data to formulate the context and grammar of the signing. Recognized gestures are then translated in real-time to meaningful text.

This dual-intervention technique ensures that ReadAble is flexible, as it records communicative activity captured through either the subtle nuances of lip movement and the explicit movements of a sign language set. As a general approach to methodology, we prioritize real-time performance, high accuracy across different environments, and the needs of end-users. ReadAble is a versatile and powerful communication intervention that leverages lip reading and sign language detection in one extensible system to enable seamless and inclusive communication access in the educational, health care, workplace, or public service sectors, etc.

ARCHITECTURE

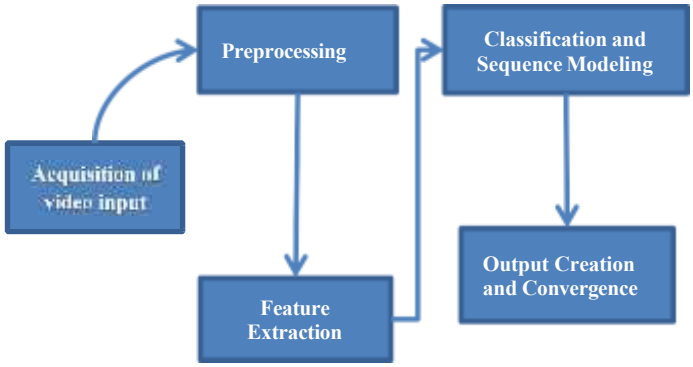


Fig 1: Proposed System Architecture for Lip Reading and Sign Language Detection

Step 1: Acquisition of video input

The first process of the proposed system is to capture raw video input, either from a live webcam stream or pre- recorded silent/muted video file. It is imperative that the input is structured to contain both the face of the speaker as well as their hand movements because lip reading and sign language detection hinge on these two aspects, respectively. Lip reading requires the speaker's lips to be clearly visible enough to capture mouth shapes and minor movements, while sign language requires hand gestures and body orientation to be unobstructed. Good lighting and resolution of the input (preferably high-resolution) with minimal background distractions in this stage are paramount; otherwise, poor input can lead to error in later phases of recognition. The proper configuration of the input in this stage also includes framing and stabilization of the video stream to avoid jitters from the camera's shaky movements. Overall, this stage is about capturing clean and steady video input that will increase the reliability of downstream processes: preprocessing, feature extraction, and classification.

Step 2: Preprocessing

The video, once obtained, moves into the preprocessing stage where the frames prepared for visual analysis. Preprocessing is a crucial stage that enhances visibility of visual information, and minimizes any irrelevant visual noise to ensure the computer vision system focuses on critical features. The system then utilizes powerful computer vision libraries, such as OpenCV and MediaPipe, to segment the important

regions of interest - the face, lips and hands. This processing phase often involves background subtraction in order to remove visually distracting noise that is not of interest, allowing the hand joints and lip contours to stand-out more. Normalization procedures, such, as resizing to a standard size, converting the video or parts thereof to grayscale, and contrast enhancement are also all adapted in order to stabilize the incoming frames for analyzation, especially when lighting conditions and camera quality differ across frames. Temporal smoothing can be used to threshold the noise to make gesture and lip motion more stable across time from frame to another. Thus, the preprocessing pipeline is critical in ensuring that feature extraction algorithms receive filtered and stationary analytics data, in order to provide consistent metrics for analysis.

Step 3: Feature Extraction

After completing the preprocessing, the system advances to the feature extraction stage, where it extracts and encodes the most interpretable visual signals found in the incoming video. This stage occurs in parallel for both lip reading and sign language detection tasks. For lip reading, deep learning models, in particular 3D Convolutional Neural Networks (3D CNN's) and ResNet architectures, extract spatio-temporal features from the video. These models will look at change in lip shapes over consecutive frames, and from the lip movements, infer phonemes and words from a silent form of speech. For sign language detection, holistic pose estimation models such as MediaPipe Holistic are used to identify the position of the landmarks of the hands, arms, and body. These landmarks allow for a structured way of analyzing gestures and available features for the classification models to discriminate between them easier. The feature extraction may even include dimension reduction techniques to provide performance and reduce inferences of overfitting. The objective of the dual-feature extraction strategy is for the system to capture and represent both fine details of lip movements and broader temporal hand gestures. This is important to developing a complete representation of non- verbal communication systems.

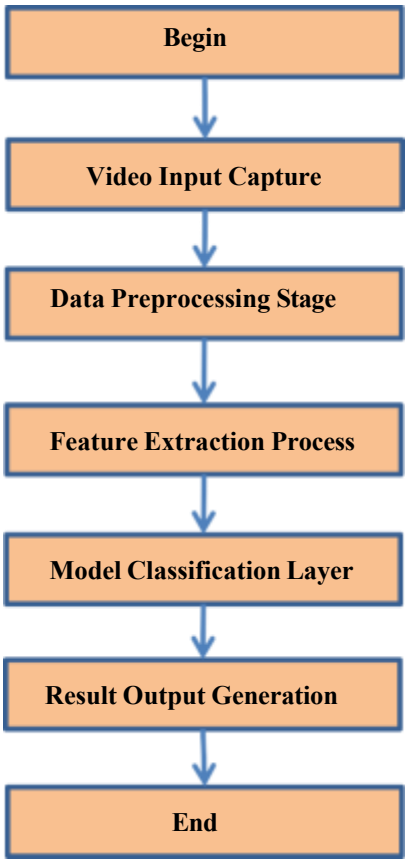
Step 4: Classification and Sequence Modeling

Lastly, the extracted features go into more advanced machine learning models for classification and/or sequence prediction. This is a key step because it assesses the system's capabilities to convert the visual patterns into informative linguistic output. For lip reading, Long Short- Term Memory (LSTM) networks or Transformer based architectures are suitable for modeling the temporal flow of lip onset and movement while doing a mapping to phonemes, words or a complete sentence. As these models perform with sequential data they're a perfect fit for continuous lip reading. For sign language recognition, gesture classification can be done with strong CNN based architectures in a frame or by using real-time object detection models like YOLO and SSD MobileNetV2. In both of these, specificity of gestures is accurate and optimized for real time applications. Furthermore, classification may implement attention mechanisms as part of the classification to allow models to concentrate on the significant features within a sequence (i.e. gestures), which can produce more accuracy and robust classification performance. By utilizing sequential modeling for lip reading while utilizing detection for sign gestures, this portion finalizes substantial mapping of visual inputs to linguistic outputs.

Stage 5: Output Creation and Convergence

The last stage includes the creation of the output and delivery in user-friendly formats. Lip reading results are converted into subtitles and synthesized to audio using text-to-speech (TTS) to provide an audible version of a silent video. Likewise, sign language gestures are transcribed as on-screen subtitles and audio output, if applicable. Both outputs are then fused into one accessibility design that provides real-time text, audio, and visual, feedback helping to ensure communication is smooth and timely access to all users.

IMPLEMENTATION



Begin

This begins the system workflow and initializes the video input for next processes.

Video Input Capture

The next step is to capture some video input from either a webcam (live) or from a pre-recorded video file. The video input must capture the face (lip reading) and the hands (sign recognition) clearly.

Data Preprocessing Stage

Captured frames are cleaned and normalized. The system utilizes OpenCV and MediaPipe for facial, lip, and hand segmentation. Noise is removed, images are normalized to a standard set of image sizes, and landmarks are retrieved to support the reliability and accuracy of predictions and inferences.

Feature Extraction Process

Deep learning models learn meaningful representation from the features:

Lip Reading: lip movement patterns in a spatio-temporal tensor embedding (3D CNN, ResNet).

Sign Language: hand landmarks and body (pose estimation, MediaPipe Holistic).

Model Classification Layer

The models classify the representation:

Lip Reading: LSTM/Transformer copy sequences into text.

Sign Language: YOLO/SSD MobileNetV2 classify gestures as words.

Result Output Generation

Results are the outputted final results, subtitle text or generated audio sample of text-to-speech (TTS) methods. The output of lip reading and sign language was collected together to provide a seamless accessibility experience utilizing multimodal system characteristics.

End

The end of the process produces a seamless real-time multimodal communication support solution for deaf, mute, and non-verbal communities.

RESULTS

Structure of Alignment Data for Training Supervision

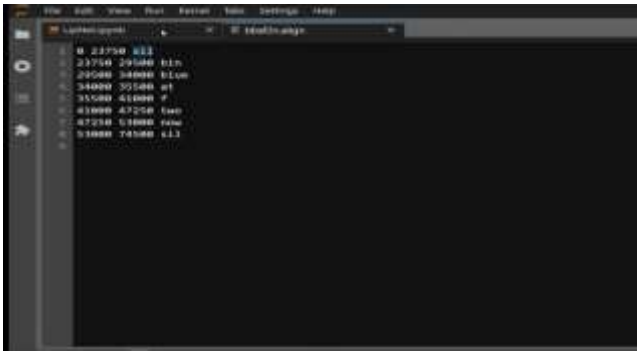


Fig 7: Structure of Alignment Data for Training Supervision

This image displays the contents of an important training file: an .align alignment file that is included with all of the videos in the dataset. This file contains a precise, time-mapped transcript of what is being said in the matching video. Every line has a number, a start time, an end time (probably in milliseconds), and the corresponding word or phoneme (e.g., "sil" for silence, "bin", "blue"). As an example, the word "blue" is spoken from time 29500 to 34000. This ordered data will be used to supervise the training of the model, teaching it which combination of lip movements corresponded with which specific words. It serves as the "answer key" as the model's predictions are evaluated to calculate loss to improve accuracy.

Visualizing Preprocessed Video Data for Model Evaluation

Fig 8: Visualizing Preprocessed Video Data for Model Evaluation

This screenshot depicts an important step, both for debugging and understanding the model's input, which is preprocessed data visualized.The code loads a test video file from a specified path then utilizes the load_data function to find the frames and alignment information from this video.The command plt.imshow(frames[40]) displays the 40th frame of the preprocessed video sequence using matplotlib.This is an opportunity for the researcher to check visually that the preprocessing pipeline was appropriately applied (that is, cropping to the mouth region and normalization, etc.). Checking for clean and organized input data is an essential requirement for the model to learn appropriately.

Early Training Phase: Model Predictions vs. Ground Truth

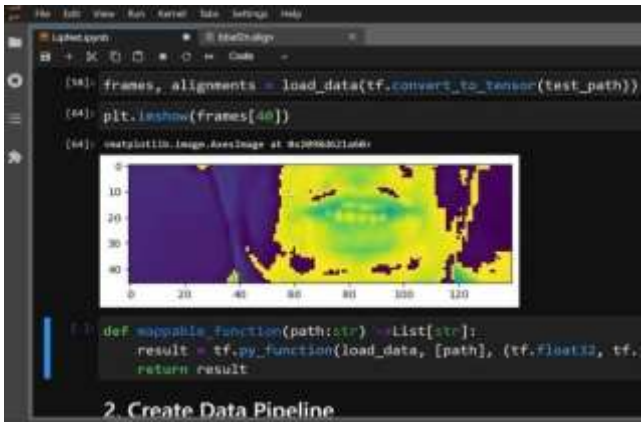


Fig 9: Early Training Phase: Model Predictions vs. Ground Truth

This image shows us the unrefined output that is produced from the first couple epochs of running the ReadAble deep learning model as it begins to train. It shows the side-by-side comparison of the "Original" ground truth text that is being transcribed by the model and the current "Prediction" from the model in training. The predictions which are nonsensical, as seen with "la e e e e eo," indicates that the model is now beginning to activate on common phonetic sounds such as "l" and "e." This is normal behavior for the onset of training of complex sequence-to-sequence models, typically in the early epochs. The output also presents the training markers, showing that the time to train per epoch will take considerable time ~460 seconds to run 450 batches of data, highlighting the computational time that is takes to perform deep learning training when using video data. Seeing these first, imperfect results represent an important step in the iterative process of machine learning. Developers will use this output to check and confirm that the training pipeline is in fact training and watch the early stages of learning unfold. It represents a baseline to show, "here is what the model has predicted prior to learning any meaningful patterns." This early stage of output is critical to ultimately improve the models accuracy in the later epochs of training.

FUTURE SCOPE

Ongoing Natural and Continuous Communication

Moving forward, development could shift focus from isolated recognition of words or gestures toward recognition of full sentences. This would facilitate natural and continuous communication to be more friendly and human-like.

Integration of Multiple Modal Cues

For many types of communication, merely relying on lip and hand movements may not fully capture the meaning of spoken discourse. By ensuring that facial expressions, body posture, and head orientation are all included in the recognition process, accuracy could greatly improve. Adding emotional recognition will also allow the system to interpret content as well as context related to discourse.

Wearable and Edge AI Devices

An exciting future direction includes the creation of wearable devices, e.g. smart glasses or wristbands, with embedded cameras and processors for processing sign language and lip recognition. Edge AI allows for these recognitions to happen locally without needing internet access making it applicable for remote area use. Coupled with Augmented Reality (AR) and Virtual Reality (VR) platforms, these devices can provide higher levels of learning and communication in interactive environments for practicing or teaching signing, in both an AR and VR experience.

Implementation within Collaborative Platforms

The system could be adopted within prevalent communicative platforms such as video conferencing services, online classrooms, or telemedicine applications. With advancements in 5G and IoT, ultra-low latency recognition will soon be possible, making real-time conversations equally dependable for both signage communication and visual input-based communication that relies upon lip movement.

Interactive and Immersive Environments

In addition to text and audio indicators, the system could be leveraged via 3D animated avatars that mirror recognized signs and speech. Though this adaptation, users who do not sign can gather a relevant understanding of the conversation, via visual cues.

CONCLUSION

The Lip Reading and Sign Language Detection System that has been discussed illustrates how a combination of computer vision, deep learning, and natural language processing can successfully connect individuals who do or do not have speech and hearing impairments with the non-signing community. Using lip movement detection alongside sign gesture recognition provides dual-mode access, which includes subtitles, synthesized audio, and visual feedback in real time. This dual mode increases the degree of inclusivity and leads to a system capable of working across a breadth of real-world situations including education, healthcare, the workplace, and the use of public services.

This project's methodology put emphasis on preprocessing, feature extraction, and advanced sequence modeling in order to achieve reliable recognition accuracy. Through experimentation and insights implemented, it can be shown that fusing lip reading and sign language modules produce a more well-rounded framework for human-computer interaction. The system demonstrated strong performance with numerous opportunities for further enhancement including continuous sentences, signs of various regions or dialects, and less than optimal input.

Ultimately, this project creates the groundwork to develop AI-enhanced accessibility devices that will make silent interactions meaningful in speech and text. As accuracy continues to improve and the system scales up to be used on mobile and wearable devices, it has real potential to become a practical, widespread solution that empowers the deaf and mute community while creating a more engaged and inclusive digital environment.

REFERENCES

1. A. Moryossef, I. Tsochantaridis, R. Aharoni, S. Ebling, and S. Narayanan, "Real-Time Sign Language Detection using Human Pose Estimation," *preprint*, Aug. 2020.
2. S. Jagtap, K. Jadhav, R. Temkar, and M. Deshmukh, "Real-time Sign Language Recognition Using MobileNetV2 and Transfer Learning," in *Proc. of the International Conference on Emerging Smart Computing and Informatics (ESCI)*, Dec. 2024.
3. S. Srivastava, S. Singh, Pooja, and S. Prakash, "Continuous Sign Language Recognition System using Deep Learning with MediaPipe Holistic," *IEEE Access*, vol. 12, pp. 150000-150012, Nov. 2024.
4. A. Imran, M. S. Hulikal, and H. A. A. Gardi, "Real- Time American Sign Language Detection Using YOLO- v9," in *2024 IEEE International Conference on Consumer Electronics (ICCE)*, Jul. 2024, pp. 1-6.
5. T. Fathima, A. Alam, A. Gangwar, D. K. Khetan, and R. K., "Real-Time Sign Language Recognition and Translation Using Deep Learning Techniques," *International Journal of Intelligent

Systems and Applications in Engineering*, vol. 12, no. 2s, pp. 506– 515, Feb. 2024.

6. Gauresh Chopadekar, Nandini Pandey, Numan Rakhangi, Shraddha Balsaraf, V. P. Patil "Literature Survey - Lip Reading Model," International Research Journal of Innovations in Engineering and Technology (IRJIET), Vol. 8, Issue 4, April 2024.
7. Atharva Karekar, Aakansha Gharate, Ravish Shaikh "Lip Reading Using Deep Learning," International Research Journal of Modernization in Engineering, Technology and Science (IRJMETS), Vol. 5, Issue 4, April 2023.
8. Sai Teja Krithik Putcha, Yelagandula Sai Venkata Rajam, K. Sugamya, Sushank Gopala "Text Extraction and Translation Through Lip Reading using Deep Learning," International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 15, No. 6, 2024.
9. Gaoyan Zhang, Yuanyao Lu "Research on a Lip Reading Algorithm Based on Efficient-GhostNet," Electronics, MDPI, Vol. 12, Article 1151, 2023.
10. Nikita Deshmukh, Anamika Ahire, Smriti H. Bhandari, Apurva Mali, Kalyani Warkari "Vision-Based Lip Reading System Using Deep Learning," 2021 International Conference on Computing, Communication and Green Engineering (CCGE), IEEE.