AI Driven Lip Reading System for Assistive Communication

Project Work Synopsis

Submitted in the partial fulfilment for the award of the degree of

BACHELOR OF ENGINEERING
IN
COMPUTER SCIENCE WITH SPECIALIZATION IN
ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Submitted by:

21BCS6634 RASHAZ RAFEEQUE
21BCS6589 JEEVAN A.J
21BCS6272 RHISHITHA T.S

Under the Supervision of:

Dr. Preet Kamal

CHANDIGARH UNIVERSITY, GHARUAN, MOHALI - 140413, PUNJAB

May, 2025

Abstract

Lip-reading technology has made significant progress with deep learning-based models, enhancing accuracy, efficiency, and real-time performance. Traditional assistive communication methods for individuals with speech impairments, such as text-based devices or sign language interpreters, often come with limitations in speed, accessibility, and ease of use. This project aims to address these challenges by developing a deep learning-powered system that converts lip movements into natural speech, allowing non-verbal individuals to communicate effortlessly.

The system utilizes the LipNet model, which combines Convolutional Neural Networks (CNNs) for feature extraction and Recurrent Neural Networks (RNNs) with Connectionist Temporal Classification (CTC) for sequence prediction. This enables the recognition of continuous speech directly from silent video clips, making it a highly effective tool for real-time lip-reading. To enhance accuracy, a text processing module incorporating Natural Language Processing (NLP) techniques refines the transcribed text by correcting errors and predicting contextually appropriate words. This ensures that the generated text is coherent, reducing ambiguities in communication.

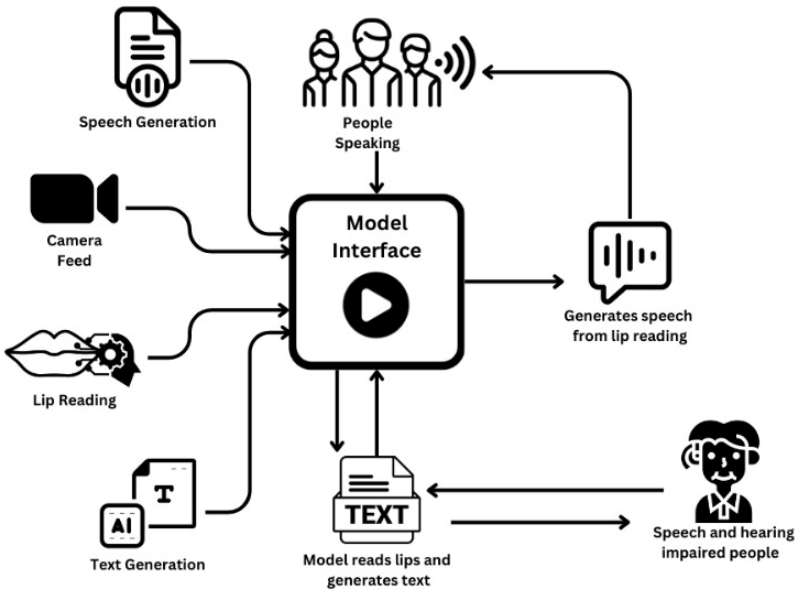


Table of Contents

1. INTRODUCTION

Effective communication is fundamental to human interaction, yet individuals with speech impairments face significant challenges in expressing themselves fluently. Existing assistive communication methods, such as text-based input devices or sign language interpreters, can be slow, inconvenient, and often require external assistance. These limitations create barriers in social interactions, education, and professional environments, reducing the independence of non-verbal individuals. With advancements in artificial intelligence, particularly in deep learning and computer vision, there is an opportunity to develop a more efficient and natural communication system that translates lip movements into speech in real-time.

This project proposes a deep learning-based lip-reading system that captures and interprets lip movements, converting them into meaningful text and subsequently into speech. The core of the system is built on the LipNet model, which employs Convolutional Neural Networks (CNNs) for feature extraction and Recurrent Neural Networks (RNNs) with Connectionist Temporal Classification (CTC) for sequence prediction. Unlike traditional phoneme-based speech recognition systems, this approach processes entire sentences from video frames, enabling more accurate and fluid recognition of speech from lip movements. Additionally, a Natural Language Processing (NLP)-based text refinement module enhances the accuracy of the generated text by correcting errors and improving contextual coherence, ensuring seamless conversation flow.

To facilitate natural communication, the processed text is converted into speech using a Text-to-Speech (TTS) engine, producing clear and expressive vocal output. The system is optimized for real-time performance, ensuring minimal delay in generating responses, making it suitable for everyday conversations in homes, schools, and workplaces. By eliminating the need for manual input and external communication aids, this project provides a hands-free, intuitive, and accessible solution for individuals with speech disabilities. Beyond improving communication, this innovation paves the way for future advancements in AI-driven assistive technologies, fostering a more inclusive and connected society

1.1 Problem Definition

Individuals with speech impairments face significant challenges in expressing themselves due to the limitations of existing assistive communication methods, such as text-based devices or sign language. These methods can be slow, inconvenient, and dependent on external assistance, making real-time communication difficult. The lack of an intuitive and hands-free solution restricts their ability to engage effectively in daily interactions. This project aims to develop a deep learning-based system that recognizes lip movements, converts them into text, and then synthesizes speech, enabling seamless and natural communication for non-verbal individuals. Lip-reading systems face challenges in accurately predicting words due to variations in lip movements, speaker differences, and environmental conditions.

1.2 Problem Overview

Speech is a primary mode of human communication, but for individuals who are mute or have speech disorders, expressing thoughts can be a challenge. Current assistive technologies, including typing-based devices and sign language interpretation, often require significant effort and are not universally accessible. Recent advancements in deep learning, particularly in lip-reading and speech synthesis, provide an opportunity to bridge this communication gap.

This project proposes a real-time lip-reading system that leverages deep learning techniques to analyze lip movements, generate corresponding text, and convert it into natural speech using a Text-to-Speech (TTS) engine. By integrating LipNet for lip-reading, NLP for text refinement, and TTS for speech synthesis, the system ensures high accuracy and fluid communication. The real-time nature of the system makes it highly practical for daily interactions, allowing non-verbal individuals to express themselves effortlessly in social, educational, and professional settings. This innovation not only enhances accessibility but also contributes to the advancement of AI-driven assistive technologies, fostering inclusivity and independence.

1.3 Hardware Specification

- 11th Gen Intel® i7-11800H @ 2.30GHz
- 16 GB RAM. 256GB SSD 1TB HDD

1.4 Software Specification

- Operating System: Windows with appropriate drivers for GPU.
- Development Environment: Pycharm/Jupyter/Google Colab
- Python: Programming language for implementing neural networks and TTS integration.
- Libraries and Tools:
  - TensorFlow / Keras / PyTorch for building and training CNN & Bi-Directional LSTM models.
  - OpenCV: For image processing and real-time video capture.
  - gTTS (Google Text-to-Speech) or pyttsx3 for converting object labels to audio.
  - Pandas / NumPy for dataset manipulation.
  - Matplotlib / Seaborn for visualizing model performance.
- Dataset: GRID CORPUS & OuluVS2

2. LITERATURE SURVEY

2.1 Literature Review Summary

Year and Citation	Article/ Author	Technique	Source	Evaluation Parameter
Exarchos, Themis, Georgios N. Dimitrakopoulos, Aristidis G. Vrahatis, Georgios Chrysovitsiotis, Zoi Zachou, and Efthymios Kyrodimos. "Lip-Reading Advancements: A 3D Convolutional Neural Network/Long Short-Term Memory Fusion for Precise Word Recognition." BioMedInformatics 4, no. 1 (2024): 410-422.	Exarchos, Themis, Georgios N. Dimitrakopoulos, Aristidis G. Vrahatis, Georgios Chrysovitsiotis, Zoi Zachou, and Efthymios Kyrodimos.	3D CNN + LSTM fusion for lip-reading	BioMedInformatics	High precision in word recognition, useful for improving lip-reading accuracy
Prajwal, K.R., Mukhopadhyay, R., Namboodiri, V.P. and Jawahar, C.V., 2020. Learning individual speaking styles for accurate lip to speech synthesis. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 13796-13805).	Prajwal, K.R., Mukhopadhyay, R., Namboodiri, V.P. and Jawahar, C.V.	Lip-to-speech synthesis using deep learning models for speaking style adaptation	IEEE/CVF Conference on CVPR	Speaking style adaptation, valuable for natural speech synthesis from lip movements
Chandra, A., Paruchuri, C., Karthika, A. and Yogitha, P., 2024. Lip Reading Using Neural networks and Deep Learning. Available at SSRN 4825936	Chandra, A., Paruchuri, C., Karthika, A. and Yogitha, P	Deep learning and neural networks for lip reading	SSRN	General performance of deep learning models in lip-reading tasks.
Paul, Suraj, Dhanesh Lakhani, Divyanshu Aryan, Shudhashekhar Das, and Rohit Varshney. "Lip Reading System for Speech-Impaired Individuals."	Paul, Suraj, Dhanesh Lakhani, Divyanshu Aryan, Shudhashekhar Das, and Rohit Varshney.	Lip-reading system for speech-impaired individuals	IJFMR	Application-oriented insights for assisting hearing-impaired individuals

Jishnu, T. S., and Anju Antony. "LipNet: End-to-End Lipreading." Indian Journal of Data Mining (IJDM) 4, no. 1 (2024): 1-4.	Jishnu, T. S., and Anju Antony.	LipNet (End-to-End lip-reading model)	Indian Journal of Data Mining (IJDM)	End-to-end efficiency, beneficial for real-time lip-reading implementation.
Kholiev, V. O., and O. Yu Barkovska. "Improved Speaker Recognition System Using Automatic Lip Recognition." Control systems & computers 1 (2024): 38-49.	Kholiev, V. O., and O. Yu Barkovska.	Automatic lip recognition for speaker identification	Control Systems & Computers	Speaker recognition enhancement through lip movements
Shahed, Md Tanvir Rahman, Md Tanjil Islam Aronno, Hussain Nyeem, Md Abdul Wahed, Tashrif Ahsan, R. Rafiul Islam, Tareque Bashar Ovi, Manab Kumar Kundu, and Jane Alam Sadeef. "LipBengal: Pioneering Bengali Lip-Reading Dataset for Pronunciation Mapping through Lip Gestures." Data in Brief (2024): 111254.	Shahed, Md Tanvir Rahman, Md Tanjil Islam Aronno, Hussain Nyeem, Md Abdul Wahed, Tashrif Ahsan, R. Rafiul Islam, Tareque Bashar Ovi, Manab Kumar Kundu, and Jane Alam Sadeef.	LipBengal dataset for Bengali lip-reading and pronunciation mapping	ivySCI	Dataset availability, useful for training multilingual lip-reading models.
Wang, Huijuan, Gangqiang Pu, and Tingyu Chen. "A lip-reading method based on 3D convolutional vision transformer. " *IEEE Access* 10 (2022): 77205-77212.	Huijuan Wang, Gangqiang Pu, Tingyu Chen	3D Convolutional Vision Transformer	IEEE	The paper evaluates the proposed 3D Convolutional Vision Transformer (3DCvT) model for lip reading by measuring its word recognition accuracy on the LRW and LRW-1000 datasets.
Sarhan, Amany M., Nada M. Elshennawy, and Dina M. Ibrahim. "HLR-net: a hybrid lip-reading model based on deep convolutional neural networks." *Computers, Materials and Continua* 68, no. 2 (2021): 1531-49.	Amany M. Sarhan, Nada M. Elshennawy and Dina M. Ibrahim	HLR-Net, Encoder & Decoder	Tech Science	HLR-Net uses inception, gradient, and GRU layers in its encoder and attention and fully connected layers in its decoder, with performance evaluated using CER, WER, and BLEU score.

Al-Qurishi, Muhammad, Thariq Khalid, and Riad Souissi. "Deep learning for sign language recognition: Current techniques, benchmarks, and open issues." *IEEE Access* 9 (2021): 126917-126951.	Muhammad Al-Qurishi, Thariq Khalid, Riad Souissi	ML (Naive Bayes, Random Forest, SVM) & DL (CNNs, RNNs, HMMs)	IEEE	The review analyzes SLR benchmark datasets and performance, noting the difficulty of direct comparison due to varied datasets and metrics. While not specifying metrics, it implies standard SLR evaluations are used, focusing on approaches and frameworks, not a meta-analysis.
Guliani, Dhruv, Françoise Beaufays, and Giovanni Motta. "Training speech recognition models with federated learning: A quality/cost framework." In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3080-3084. IEEE, 2021.	Dhruv Guliani, Françoise Beaufays, Giovanni Motta	Federated learning	IEEE	A novel metric evaluates the trade-off between model quality and computational cost. Hyperparameter optimization and variational noise are used to compensate for non-IID data effects.
Reddy, V. Madhusudhana, T. Vaishnavi, and K. Pavan Kumar. "Speech-to-Text and Text-to-Speech Recognition Using Deep Learning." In *2023 2nd International Conference on Edge Computing and Applications (ICECAA)*, pp. 657-666. IEEE, 2023.	V. Madhusudhana Reddy, T. Vaishnavi, K. Pavan Kumar	CNNs, RNNs and transformer-based models	IEEE	The review covers advancements in STT and TTS, from traditional methods to deep learning It discusses challenges like accuracy, accent diversity, and context awareness, implying standard evaluation metrics are used in the field, but focuses on approaches and future directions.
Matsui, Kenji, Kohei Fukuyama, Yoshihisa Nakatoh, and Yumiko O. Kato. "Speech enhancement system using lip-reading." In *2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, pp. 1-5. IEEE, 2020.	Kenji Matsui, Kohei Fukuyama, Yoshihisa Nakatoh, Yumiko O. Kato	Variational Autoencoder (VAE)	IEEE	Word recognition accuracy. The experiments achieved 65% accuracy, and 100% when considering the top two candidate words, using a dataset of 20 Japanese words.

Prashanth, B. S., MV Manoj Kumar, B. H. Puneetha, R. Lohith, V. Darshan Gowda, V. Chandan, and H. R. Sneha. "Lip Reading with 3D Convolutional and Bidirectional LSTM Networks on the GRID Corpus." In *2024 Second International Conference on Networks, Multimedia and Information Technology (NMITCON)*, pp. 1-8. IEEE, 2024.	B S Prashanth, M V Manoj Kumar, B H Puneetha, R Lohith, V Darshan Gowda, V Chandan	3D Convolutional Neural Networks, bidirectional Long Short-Term Memory	IEEE	Character Error Rate (CER) and Word Error Rate (WER). The best model achieved a CER of 1.54% and a WER of 7.96% on benchmark datasets.
G. C, R. J. D, S. K. A and S. S. V. P. Reddy, "AI Lip Reader Detecting Speech Visual Data with Deep Learning," 2024 4th International Conference on Intelligent Technologies (CONIT), Bangalore, India, 2024, pp. 1-6, doi: 10.1109/CONIT61985.2024.10627537. keywords: {Deep learning;Visualization;Adaptation models;Accuracy;Lips;Speech recognition;Linguistics;Bidirectional LSTM;Lipreading;Convolutional Neural Network (CNN);latent space}	G. C, R. J. D, S. K. A, S. S. V. P. Reddy	3DCNN, BiLSTM, Multilingual Dataset	IEEE Explore	Evaluation parameters include 98.4% accuracy, along with ROC curves, confusion matrices, precision, recall, and F1-score, ensuring comprehensive assessment of model performance in multilingual lip-reading.
E. de la Guía, V. L. Camacho, L. Orozco-Barbosa, V. M. Brea Luján, V. M. R. Penichet and M. Lozano Pérez, "Introducing IoT and Wearable Technologies into Task-Based Language Learning for Young Children," in IEEE Transactions on Learning Technologies, vol. 9, no. 4, pp. 366-378, 1 Oct.-Dec. 2016, doi: 10.1109/TLT.2016.2557333.	E. de la Guía, V. L. Camacho, L. Orozco-Barbosa, V. M. Brea Luján, V. M. R. Penichet and M. Lozano Pérez	IoT	IEEE Org	Evaluation parameters include student engagement, language acquisition, instructor ease of use, scenario effectiveness, and task-based performance accuracy.
Mevlüde Akdeniz, Fatih Özdinç, Maya: An artificial intelligence based smart toy for pre-school children, International Journal of Child-Computer Interaction, Volume 29, 2021, 100347, ISSN 2212-8689	Mevlüde Akdeniz, Fatih Özdinç	AI, Image Processing, NLP	ScienceDirect	Evaluation parameters include engagement, learning outcomes, usability, adaptability, and overall satisfaction.

Khondaker A. Mamun, Rahad Arman Nabid, Shehan Irteza Pranto, Saniyat Mushrat Lamim, Mohammad Masudur Rahman, Nabeel Mahammed, Mohammad Nurul Huda, Farhana Sarker, Rubaiya Rahtin Khan, Smart reception: An artificial intelligence driven bangla language based receptionist system employing speech, speaker, and face recognition for automating reception services, Engineering Applications of Artificial Intelligence, Volume 136, Part A, 2024, 108923, ISSN 0952-1976	Khondaker A. Mamun, Rahad Arman Nabid, Shehan Irteza Pranto, Saniyat Mushrat Lamim, Mohammad Masudur Rahman, Nabeel Mahammed, Mohammad Nurul Huda, Farhana Sarker, Rubaiya Rahtin Khan	AI, Face Recognition, Speech Recognition, ASR, TTS	ScienceDirect	The evaluation parameters for the study include the accuracy of face and speaker recognition, the Word Error Rate (WER) for the ASR model, the Mean Opinion Score (MOS) for TTS, validation loss for the question-answering system, and overall user satisfaction rates from real-world testing among participants.
Amara, K., Boudjemila, C., Zenati, N., Djekoune, O., Aklil, D., & Kenoui, M. (2022). AR Computer-Assisted Learning for Children with ASD based on Hand Gesture and Voice Interaction. IETE Journal of Research, 69(12), 8659–8675.	Amara, K., Boudjemila, C., Zenati, N., Djekoune, O., Aklil, D., & Kenoui, M	AR, Gesture Recognition, Voice Recognition	Taylor & Francis Online	Evaluation parameters include engagement, vocabulary, social interaction, feedback, and AR vs. non-computer outcomes.
arXiv:2401.05459 (cs) [Submitted on 10 Jan 2024 (v1), last revised 8 May 2024 (this version, v2)] Personal LLM Agents: Insights and Survey about the Capability, Efficiency and Security Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, Rui Kong, Yile Wang, Hanfei Geng, Jian Luan, Xuefeng Jin, Zilong Ye, Guanjing Xiong, Fan Zhang, Xiang Li, Mengwei Xu, Zhijun Li, Peng Li, Yang Liu, Ya-Qin Zhang, Yunxin Liu	Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, Rui Kong, Yile Wang, Hanfei Geng, Jian Luan, Xuefeng Jin, Zilong Ye, Guanjing Xiong, Fan Zhang, Xiang Li, Mengwei Xu, Zhijun Li, Peng Li, Yang Liu, Ya-Qin Zhang, Yunxin Liu	IPAs, LLMs, IoT	Cornel University (arxiv.org)	The evaluation parameters for the study likely include the effectiveness of Personal LLM Agents in understanding user intent, the efficiency of task execution, user satisfaction levels, security and privacy measures in handling personal data, and the overall capability of the agents to provide intelligent and personalized assistance.

2.2 Existing System

Several deep learning-based systems have significantly advanced lip-reading technology, improving accuracy and real-time performance. LipNet was one of the first end-to-end deep learning models, utilizing CNNs for feature extraction and RNNs with Connectionist Temporal Classification (CTC) for sequence prediction. Unlike traditional phoneme-based recognition methods, LipNet predicts entire sentences from silent video clips, making it highly effective for continuous speech recognition.

Another notable system, Lip2Wav, focuses on converting lip movements into speech. It employs a transformer-based phoneme-to-speech generation model, producing realistic and intelligible speech. This approach is particularly beneficial for assistive communication and applications such as dubbing and silent speech conversion.

AV-HuBERT (Audio-Visual Hidden Unit BERT) introduces a multimodal approach by integrating both visual and audio inputs. Using self-supervised learning, it improves speech recognition even in noisy environments or when audio signals are weak. This model demonstrates the advantage of combining modalities to enhance lip reading accuracy.

Other architectures, such as Deep Speech-based lip-reading systems, incorporate CNNs with Attention-based RNNs for improved sequence modeling and text generation from lip movements. Additionally, recent research explores Vision Transformers (ViTs) to enhance feature extraction and temporal modeling. These advancements play a crucial role in assistive communication, security applications, and silent speech interfaces, paving the way for more robust and real-time lip-reading solutions.

2.3 Problem Formulation

The project addresses a crucial gap in assistive communication technology for individuals who are mute or have speaking disabilities. Traditional methods, which often involve typing or using text-based devices, can be slow and cumbersome. This new system, powered by deep learning and computer vision, will enable real-time communication through lip-reading, allowing users to express themselves naturally without relying on traditional input methods.

By integrating a speech generator with the lip-reading model, the project ensures that the text generated from lip movements is converted into natural-sounding speech. This will allow non-verbal individuals to communicate effectively in a variety of settings, whether at home, school, or work, making interactions smoother and less dependent on external devices.

The system will be optimized for real-time performance, crucial for practical use in everyday conversations, while focusing on high accuracy in detecting lip movements and converting them into the correct words or phrases. Additionally, attention will be given to speech synthesis quality to ensure that the generated speech sounds realistic and engaging, helping users feel more connected in social situations.

The project will not only aim to improve accessibility for individuals with speech impairments but also set the stage for future advancements in AI-driven assistive technologies, contributing to a more inclusive society.

2.4 Proposed System

The proposed system is designed to assist individuals with speech impairments by converting their lip movements into speech, enabling seamless communication. This system integrates deep learning-based lip reading with speech synthesis to provide a real-time, accessible solution. The core functionality revolves around capturing lip movements, analyzing them to generate corresponding text, and converting this text into speech.

At the foundation of this system is a lip-reading model built using deep learning techniques, specifically the LipNet model. The model processes video frames captured in real time from a webcam, focusing on the movement of the lips. Using a trained neural network, it maps these movements to textual representations of spoken words. This model is designed to recognize a broad vocabulary while being adaptable to different users and variations in lip motion.

Once the lip-reading model generates text, a text processing module refines the output to enhance accuracy. This step involves error correction and contextual prediction, ensuring that incomplete or misinterpreted words are corrected based on linguistic patterns. Natural Language Processing (NLP) techniques can further improve the quality of the transcribed text, making it more coherent and readable.
The final stage of the system is speech synthesis, where the processed text is converted into spoken words using a Text-to-Speech (TTS) engine. This ensures that the generated speech is clear and natural, closely resembling human conversation. The integration of TTS makes it possible for individuals who are mute or have speaking disabilities to express themselves effortlessly, eliminating the need for manual text input or external communication aids.

By combining real-time lip reading, text processing, and speech synthesis, this system provides an innovative and practical solution for those with speech impairments. The hands-free approach enhances usability, making the technology accessible to a wider range of users. This assistive tool has the potential to greatly improve communication for people with speaking disabilities, fostering better social inclusion and independence.

3. OBJECTIVES

- To train a LipNet model using the GRID dataset to recognize and transcribe lip movements into text with high accuracy.

- To develop a real-time lip-to-speech system by capturing live video, processing frames for lip reading, and converting the predicted text into speech using TTS synthesis.

- To provide a seamless communication tool for mute individuals and those with speech disabilities by transforming lip movements into audible speech.

4. METHODOLOGY

The development of the lip-reading system involves a structured methodology comprising data preprocessing, model training, inference processing, and speech synthesis. The system is trained on the GRID corpus dataset, a widely used dataset for lip-reading research, which consists of videos of speakers pronouncing structured sentences. The training phase focuses on learning visual speech patterns, while the inference engine enables real-time prediction and speech generation.

Model Training

The lip-reading model is built using LipNet, a deep learning-based architecture that leverages Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to analyze sequential lip movements. The training process begins with data preprocessing, where each video is converted into individual frames and resized to match the model's input dimensions. The corresponding alignment files containing the transcriptions of the spoken words are processed to map each frame sequence to its respective text label.

To ensure efficient feature extraction, CNN layers are used to capture spatial information from each frame, while Bidirectional Long Short-Term Memory (BiLSTM) networks help model the temporal dependencies between consecutive lip movements. The training is performed using CTC (Connectionist Temporal Classification) loss, which allows the model to align variable-length input sequences with text outputs without requiring precise frame-to-character mapping. The optimizer, typically Adam or RMSprop, is used to adjust the model's weights for improved accuracy.

Inference Engine

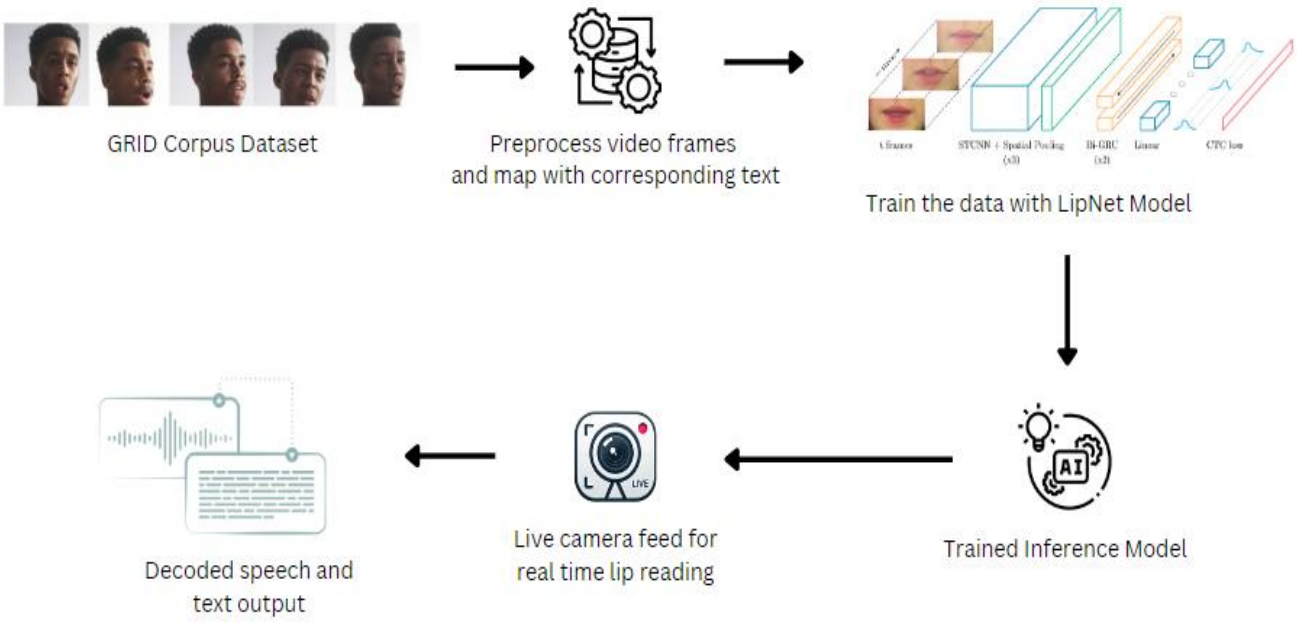
Once trained, the model is deployed for real-time lip reading. The inference process starts with capturing live video frames from a webcam. Each frame undergoes preprocessing, including resizing, normalization, and grayscale conversion if necessary. A sequence of 75 frames (matching the trained model's input requirement) is stored in a buffer and passed through the trained LipNet model to predict the corresponding text.

To improve accuracy, a text processing module is integrated to refine the predicted words. This module applies error correction algorithms, such as spelling correction and context-aware word suggestions, ensuring a more coherent and accurate transcription.

Speech Generation

Once the final text output is obtained, it is fed into a Text-to-Speech (TTS) engine, such as Google Text-to-Speech (gTTS) or Tacotron, to convert the text into synthesized speech. The generated speech is played in real time using an audio output system. The integration of lip reading with speech synthesis allows users with speaking disabilities to communicate effectively, transforming silent lip movements into audible speech.

This end-to-end system, combining deep learning-based visual speech recognition with real-time speech synthesis, provides an innovative assistive technology for individuals with speech impairments, enabling natural and effortless communication.



5. EXPERIMENTAL SETUP

The experimental setup for the real-time lip-reading and speech synthesis system involves the integration of multiple components that work in tandem to capture, process, and convert lip movements into audible speech. The setup is designed to ensure accuracy in lip-reading and high-quality speech synthesis, optimizing performance for real-time applications. Below is a detailed description of the experimental setup:

1. Hardware Setup

- **Camera:** A high-definition camera (preferably with at least 30 FPS) will be used to capture live video of the subject's face. The camera will be placed at a fixed position in front of the subject, ensuring a clear view of the lips during speech. This setup is crucial for accurately tracking and extracting lip movements.
- **Microphone:** A microphone will be used to capture any background audio or to detect whether the subject has initiated the lip-reading system using voice commands. This will help in making the system fully hands-free.
- **Computing Hardware:** The system will require a powerful computing device, likely equipped with a GPU (e.g., NVIDIA T4 or equivalent) to handle the deep learning computations in real-time. The GPU will accelerate both the lip-reading model and speech synthesis process.

2. Software Setup

- **Development Environment:**
  - **Python** will be used as the primary programming language for implementing the model and its components.
  - **TensorFlow/Keras or PyTorch** will be used to develop and train the deep learning models for lip-reading.
  - **OpenCV** will be employed for real-time video capture and preprocessing of video frames.
  - **gTTS (Google Text-to-Speech)** or **pyttsx3** will be used to synthesize the generated text into speech.
- **Libraries and Frameworks:**
  - **OpenCV:** For video capture and preprocessing, including face detection and lip region extraction.
  - **NumPy and Pandas:** For data manipulation and handling.
  - **TensorFlow or PyTorch:** For training and deploying the deep learning models for lip reading.
  - **gTTS/pyttsx3:** For text-to-speech synthesis to convert the detected text into speech.

3. Dataset

The GRID dataset will be used to train the lip-reading model. The dataset contains audio-visual data, where each video corresponds to a spoken word or phrase. The dataset includes a variety of speakers and captures a range of phonetic sounds, which helps in training the model to recognize lip movements from different individuals. The vocabulary used for training includes the characters abcdefghijklmnopqrstuvwxyz'?!0123456789.

4. Model Architecture

- **Lip-Reading Model:**
  - The model will be based on LipNet or similar convolutional-recurrent architectures capable of recognizing lip movements and associating them with textual representations. The model will be trained to recognize visual cues from the lip region and classify them into characters or words.
  - The architecture will include a combination of Convolutional Neural Networks (CNNs) for feature extraction and Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) layers for sequence modeling.
- **Inference Engine:** Once trained, the model will be deployed as an inference engine capable of performing real-time lip-reading. This engine will take input from live video capture, process frames to detect lip movements, and produce corresponding text output.

- **Text-to-Speech Model:** The text output generated by the lip-reading model will be passed to a speech synthesis system (such as gTTS or pyttsx3) to convert the text into natural-sounding speech.

5. Evaluation Metrics

The following metrics will be used to evaluate the performance of the system:

- **Accuracy of Lip-Reading:** Measured by comparing the predicted text with the ground truth from the dataset (e.g., character-level or word-level accuracy).
- **Speech Naturalness:** Assessed using **mean opinion scores (MOS)** from human evaluators to evaluate the quality of the synthesized speech.
- **Real-Time Performance:** Evaluated by measuring the system's frame rate (FPS) and the delay between video input and speech output.

This experimental setup will provide a comprehensive and effective platform to train and develop the real-time lip-reading and speech synthesis system, ensuring its potential to assist individuals with speech disabilities in everyday communication.

6. CONCLUSION

The current phase of the project has successfully established a solid foundation for the real-time lip-reading system integrated with speech synthesis. We have identified the problem and outlined a structured solution that combines advanced deep learning techniques with real-time video processing. The methodology leverages the GRID dataset for training, focusing on the conversion of lip movements into text, followed by the synthesis of natural-sounding speech through text-to-speech technology.

The system is designed to handle live video input, capturing frames from a camera and preprocessing them for lip-reading analysis. The deep learning model, trained on the GRID dataset, will recognize lip patterns and translate them into textual representation. Once the text is generated, a speech synthesis component will transform it into spoken words, enabling individuals with speech disabilities to communicate effortlessly.

While we are still in the preparation phase, the potential of this system to revolutionize communication for individuals with speech disabilities is immense. By offering a seamless conversion of lip movements into audible speech, the system will make communication more accessible and efficient, enhancing social interactions and overall quality of life for users. This tool not only improves communication but also opens the door to further developments in AI-powered assistive technologies.

.

REFERENCES

[1] Exarchos, Themis, Georgios N. Dimitrakopoulos, Aristidis G. Vrahatis, Georgios Chrysovitsiotis, Zoi Zachou, and Efthymios Kyrodimos. "Lip-Reading Advancements: A 3D Convolutional Neural Network/Long Short-Term Memory Fusion for Precise Word Recognition." BioMedInformatics 4, no. 1 (2024): 410-422.

[2] Prajwal, K.R., Mukhopadhyay, R., Namboodiri, V.P. and Jawahar, C.V., 2020. Learning individual speaking styles for accurate lip to speech synthesis. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 13796-13805).

[3] Chandra, A., Paruchuri, C., Karthika, A. and Yogitha, P., 2024. Lip Reading Using Neural networks and Deep Learning. Available at SSRN 4825936.

[4] Paul, Suraj, Dhanesh Lakhani, Divyanshu Aryan, Shudhashekhar Das, and Rohit Varshney. "Lip Reading System for Speech-Impaired Individuals."

[5] Jishnu, T. S., and Anju Antony. "LipNet: End-to-End Lipreading." Indian Journal of Data Mining (IJDM) 4, no. 1 (2024): 1-4.

[6] Kholiev, V. O., and O. Yu Barkovska. "Improved Speaker Recognition System Using Automatic Lip Recognition." Control systems & computers 1 (2024): 38-49.

[7] Shahed, Md Tanvir Rahman, Md Tanjil Islam Aronno, Hussain Nyeem, Md Abdul Wahed, Tashrif Ahsan, R. Rafiul Islam, Tareque Bashar Ovi, Manab Kumar Kundu, and Jane Alam Sadeef. "LipBengal: Pioneering Bengali Lip-Reading Dataset for Pronunciation Mapping through Lip Gestures." Data in Brief (2024): 111254.

[8] Wang, Huijuan, Gangqiang Pu, and Tingyu Chen. "A lip reading method based on 3D convolutional vision transformer." *IEEE Access* 10 (2022): 77205-77212.

[9] Sarhan, Amany M., Nada M. Elshennawy, and Dina M. Ibrahim. "HLR-net: a hybrid lip-reading model based on deep convolutional neural networks." *Computers, Materials and Continua* 68, no. 2 (2021): 1531-49.

[10] Al-Qurishi, Muhammad, Thariq Khalid, and Riad Souissi. "Deep learning for sign language recognition: Current techniques, benchmarks, and open issues." *IEEE Access* 9 (2021): 126917-126951.

[11] Guliani, Dhruv, Françoise Beaufays, and Giovanni Motta. "Training speech recognition models with federated learning: A quality/cost framework." In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3080-3084. IEEE, 2021.

[12] Reddy, V. Madhusudhana, T. Vaishnavi, and K. Pavan Kumar. "Speech-to-Text and Text-to-Speech Recognition Using Deep Learning." In *2023 2nd International Conference on Edge Computing and Applications (ICECAA)*, pp. 657-666. IEEE, 2023.

[13] Matsui, Kenji, Kohei Fukuyama, Yoshihisa Nakatoh, and Yumiko O. Kato. "Speech enhancement system using lip-reading." In *2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, pp. 1-5. IEEE, 2020.

[14] Prashanth, B. S., MV Manoj Kumar, B. H. Puneetha, R. Lohith, V. Darshan Gowda, V. Chandan, and H. R. Sneha. "Lip Reading with 3D Convolutional and Bidirectional LSTM Networks on the GRID Corpus." In *2024 Second International Conference on Networks, Multimedia and Information Technology (NMITCON)*, pp. 1-8. IEEE, 2024.

[15] G. C, R. J. D, S. K. A and S. S. V. P. Reddy, "AI Lip Reader Detecting Speech Visual Data with Deep Learning," 2024 4th International Conference on Intelligent Technologies (CONIT), Bangalore, India, 2024, pp. 1-6, doi: 10.1109/CONIT61985.2024.10627537. keywords: {Deep learning;Visualization;Adaptation models;Accuracy;Lips;Speech recognition;Linguistics;Bidirectional LSTM;Lipreading;Convolutional Neural Network (CNN);latent space}

[16] E. de la Guía, V. L. Camacho, L. Orozco-Barbosa, V. M. Brea Luján, V. M. R. Penichet and M. Lozano Pérez, "Introducing IoT and Wearable Technologies into Task-Based Language Learning for Young Children," in IEEE Transactions on Learning Technologies, vol. 9, no. 4, pp. 366-378, 1 Oct.-Dec. 2016, doi: 10.1109/TLT.2016.2557333.

[17] Mevlüde Akdeniz, Fatih Özdinç, Maya: An artificial intelligence based smart toy for pre-school children, International Journal of Child-Computer Interaction, Volume 29, 2021, 100347, ISSN 2212-8689, https://doi.org/10.1016/j.ijcci.2021.100347. (https://www.sciencedirect.com/science/article/pii/S221286892100060X)

[18] Khondaker A. Mamun, Rahad Arman Nabid, Shehan Irteza Pranto, Saniyat Mushrat Lamim, Mohammad Masudur Rahman, Nabeel Mahammed, Mohammad Nurul Huda, Farhana Sarker, Rubaiya Rahtin Khan, Smart reception: An artificial intelligence driven bangla language based receptionist system employing speech, speaker, and face recognition for automating reception services, Engineering Applications of Artificial Intelligence, Volume 136, Part A, 2024, 108923, ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2024.108923.

[19] Amara, K., Boudjemila, C., Zenati, N., Djekoune, O., Aklil, D., & Kenoui, M. (2022). AR Computer-Assisted Learning for Children with ASD based on Hand Gesture and Voice Interaction. IETE Journal of Research, 69(12), 8659–8675. https://doi.org/10.1080/03772063.2022.2101554

[20] arXiv:2401.05459 (cs) [Submitted on 10 Jan 2024 (v1), last revised 8 May 2024 (this version, v2)] Personal LLM Agents: Insights and Survey about the Capability, Efficiency and Security Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, Rui Kong, Yile Wang, Hanfei Geng, Jian Luan, Xuefeng Jin, Zilong Ye, Guanjing Xiong, Fan Zhang, Xiang Li, Mengwei Xu, Zhijun Li, Peng Li, Yang Liu, Ya-Qin Zhang, Yunxin Liu