A
Mini Project Report on

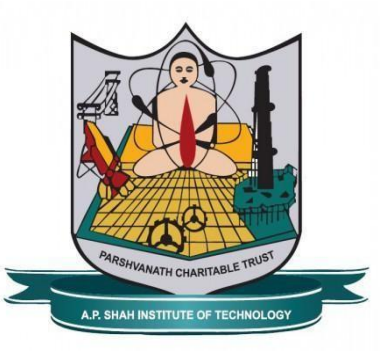
AI BASED LIP READING

Submitted in partial fulfillment of the requirements
for the degree of
BACHELOR OF ENGINEERING
IN
Computer Science & Engineering
Artificial Intelligence & Machine Learning

by

Sarang Bahikar (22106129)
Krishit Doshi (22106001)
Under the guidance of

Prof. Viki Tukaram Patil



Department of Computer Science & Engineering
(Artificial Intelligence & Machine Learning)
A. P. Shah Institute of Technology
G. B. Road, Kasarvadavali, Thane (W)-400615
University Of Mumbai
2024-2025

A. P. SHAH INSTITUTE OF TECHNOLOGY

CERTIFICATE

This is to certify that the project entitled "**AI Based Lip Reading"** is a bonafide work Sarang Bahikar (22106129), Krishit Doshi (22106001) submitted to the University of Mumbai in partial fulfillment of the requirement for the award of **Bachelor of Engineering** in **Computer Science & Engineering (Artificial Intelligence & Machine Learning).**

Prof. Viki Tukaram Patil	Dr. Jaya Gupta
Mini Project Guide	Head of Department

A. P. SHAH INSTITUTE OF TECHNOLOGY

Project Report Approval

This Mini project report entitled "**AI Based Lip Reading"** by **Author Sarang Bahikar (22106129), Krishit Doshi (22106001)** is approved for the degree of ***Bachelor of Engineering*** in ***Computer Science &Engineering***, (AIML) ***2024-25***.

External Examiner: _____

Internal Examiner: _____

Place: APSIT, Thane

Date:

Declaration

We declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission hasnot been taken when needed.

Sarang Bahikar (22106129)

Krishit Doshi (22106001)

ABSTRACT

Lip reading, the visual interpretation of speech movements, has gained significant attention in recent years due to its potential applications in various fields such as security, communication aids for the hearing impaired, and human-computer interaction. This mini project focuses on leveraging machine learning techniques to develop an automated lip reading system that can transcribe spoken words based on video input.

The core objective of this project is to build and evaluate a model capable of accurately recognizing spoken words by analysing lip movements from video frames. The approach involves collecting a dataset of synchronized video and audio recordings, where each video frame corresponds to a specific phoneme or word. Pre-processing techniques, including frame extraction and face alignment, are employed to enhance the quality of input data.

For feature extraction, convolutional neural networks (CNNs) are utilized to capture the spatial features of the lip movements, while recurrent neural networks (RNNs), specifically Long Short-Term Memory (LSTM) networks, are used to model the temporal dependencies across frames. The combined use of CNNs and LSTMs allows the system to learn both the spatial characteristics of lip movements and the sequential nature of speech.

The model's performance is evaluated using standard metrics such as accuracy and word error rate (WER). Preliminary results indicate that the system achieves a promising level of accuracy, demonstrating the feasibility of using machine learning for effective lip reading. This project not only contributes to the advancement of automated speech recognition technologies but also provides a foundation for future research in enhancing the robustness and scalability of lip reading systems.

Index

CHAPTER 1

INTRODUCTION

1. INTRODUCTION

Lip reading, or visual speech recognition, is a process by which speech is understood by interpreting the movement of the lips, face, and tongue. This capability is invaluable in various scenarios, such as assisting individuals with hearing impairments, enhancing security systems, and improving human-computer interaction. Traditional lip reading relies on human ability, which can be limited by factors such as environmental noise, speaker variability, and the speed of speech. To address these limitations, there has been increasing interest in developing automated systems that can accurately interpret lip movements using machine learning technologies.

Recent advancements in machine learning and computer vision have opened new avenues for automating the lip reading process. By harnessing the power of deep learning algorithms, researchers and engineers can develop systems that not only recognize lip patterns but also adapt to diverse speaking styles and contexts. These systems are designed to convert visual input from video frames into text, effectively transcribing spoken words without relying on auditory information.

This project aims to explore the application of machine learning techniques to create an automated lip reading system. The primary goal is to develop a model that can accurately transcribe spoken words based on visual data captured from video recordings.

A critical first step is to gather a diverse and high-quality dataset of video and audio recordings. This dataset will serve as the foundation for training and evaluating the lip reading model. The project will employ advanced computer vision techniques to extract relevant features from video frames. This involves isolating the lip region and capturing its movement patterns over time. The core of the project involves designing and training a machine learning model. Convolutional Neural Networks (CNNs) will be used to analyse spatial features of the lip movements, while Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, will capture the temporal dynamics of speech. The model's performance will be assessed using metrics such as accuracy and word error rate (WER). Continuous evaluation and optimization will be carried out to improve the model's effectiveness and robustness.

High-quality datasets are essential for training effective machine learning models. These datasets typically consist of video recordings with synchronized audio that includes various speakers, speech styles, and environments. Examples include the LRW (Lip Reading in the Wild) and GRID datasets.
Pre-processing Before training, videos must be processed to extract relevant frames and align facial features. Techniques such as face detection, landmark extraction, and lip region cropping are employed to prepare the data for feature extraction

Convolutional Neural Networks (CNNs) are used to analyse individual frames and extract spatial features related to lip movements. CNNs can learn to identify distinct patterns and textures associated with different phonemes and words. Since speech is a temporal sequence, Recurrent Neural Networks (RNNs) are used to capture the dynamic changes over time. Long Short-Term Memory (LSTM) networks are particularly suited for this purpose due to their ability to handle long-range dependencies and avoid issues like vanishing gradients.

Combining CNNs and LSTMs can effectively model both the spatial and temporal aspects of lip movements. The CNN extracts features from video frames, while the LSTM processes sequences of these features to understand the context and sequence of speech. The model is trained using a supervised learning approach, where it learns to map visual features to corresponding spoken words or phonemes. Loss functions such as cross-entropy are used to optimize the model's performance.

Accuracy and Word Error Rate (WER) are commonly used to evaluate the performance of the lip reading model. These metrics assess how well the model transcribes spoken words and how closely its output matches the actual text. To improve performance, hyper parameter tuning, data augmentation, and advanced optimization techniques are applied. Strategies like transfer learning, where pre-trained models are fine-tuned on specific datasets, can also enhance the model's effectiveness.

This project aims to develop a machine learning-based lip reading system that can transcribe spoken words from video input. By leveraging advanced computer vision and deep learning techniques, the project seeks to enhance the accuracy and applicability of automated lip reading, contributing to fields like accessibility, security, and human-computer interaction. The outcomes of this project will not only demonstrate the potential of machine learning in visual speech recognition but also pave the way for further advancements in the domain.

CHAPTER 2

LITERATURE SURVEY

2. LITERATURE SURVEY

2.1-HISTORY

The concept of interpreting speech through visual cues has ancient roots. Historical accounts suggest that ancient civilizations, including the Greeks and Romans, recognized the importance of lip movements in understanding speech, though formal methods were not developed. In the 17th century, educators for the deaf, like Juan Pablo Bonet, began exploring how to teach speech through visual cues.4

The 19th century saw significant advancements in education for the deaf, with pioneers like Laurent Clerc and Thomas Hopkins Gallaudet developing formal systems for teaching deaf students. Lip reading became an integral part of these educational approaches. Researchers began studying how lip movements relate to speech, laying the groundwork for more systematic approaches.

In the early 20th century, systems like the "Visible Speech" developed by Alexander Melville Bell and "Speech Reading" developed by Sarah Fuller began formalizing lip reading techniques. These methods aimed to teach deaf individuals how to read lips more effectively. With the rise of video technology, researchers could better analyze and record lip movements, improving their understanding of visual speech. The 1980s and 1990s saw the development of early computer models for lip reading. These models used basic image processing techniques to recognize lip movements, although their accuracy was limited by the technology of the time.

The 2000s marked a significant leap with the introduction of machine learning algorithms and deep learning techniques. Researchers began applying Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to lip reading tasks, leading to improvements in accuracy and robustness. Pioneering work such as the development of the GRID corpus and the LRS (Lip Reading Sentences) dataset provided large-scale, labelled data necessary for training modern models.

Future research aims to further improve the accuracy and robustness of lip reading systems, especially in challenging conditions such as low lighting or noisy environments. The integration of lip reading with artificial intelligence (AI) and augmented reality (AR) or virtual reality (VR) could lead to new applications and enhanced user experiences. As lip reading technology advances, ethical and privacy concerns related to surveillance and consent will need to be addressed to ensure responsible use.

2.2-LITERATURE REVIEW

LipNet: End-to-End Sentence-Level Lipreading. Assael, Y., Shillingford, B., & de Freitas, N. (2016).

LipNet represents a pioneering approach in the field of lip reading, offering an end-to-end solution for sentence-level prediction that integrates both visual feature learning and sequence modeling. Its success in achieving high accuracy on challenging tasks demonstrates the effectiveness of deep learning techniques in capturing the complexities of visual speech. This advancement highlights the potential for further innovations in automated lip reading and its applications in assistive technologies and communication systems.

A Study of Word-Level Lip Reading Accuracy. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Gergen, J., et al. (2016).

Lip-reading is the task of decoding text from the movement of a speaker's mouth. Traditional approaches separated the problem into two stages: designing or learning visual features, and prediction. More recent deep lip-reading approaches are end-to-end trainable (Wand et al., 2016; Chung & Zisserman, 2016a). However, existing work on models trained end-to-end perform only word classification, rather than sentence-level sequence prediction. Studies have shown that human lip-reading performance increases for longer words (Easton & Basala, 1982), indicating the importance of features capturing temporal context in an ambiguous communication channel. Motivated by this observation, we present LipNet, a model that maps a variable-length sequence of video frames to text, making use of spatiotemporal convolutions, a recurrent network, and the connectionist temporal classification loss, trained entirely end-to-end. To the best of our knowledge, LipNet is the first end-to-end sentence-level lip reading model that simultaneously learns spatiotemporal visual features and a sequence model. On the GRID corpus, LipNet achieves 95.2% accuracy in sentence-level, overlapped speaker split task, outperforming experienced human lip readers and the previous 86.4% word-level state-of-the-art accuracy

Lip Reading in the Wild. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Chung, J. S., & Zisserman, A. (2016a).

The goal of this work is to recognize phrases and sentences being spoken by a talking face, with or without the audio. Unlike previous works that have focused on recognizing a limited number of words or phrases, we tackle lip reading as an open-world problem - unconstrained natural language sentences, and in the wild videos. This lip reading performance beats a professional lip reader on videos from BBC television, and we also demonstrate that visual information helps to improve speech recognition performance even when the audio is available.

CHAPTER 3

Problem Statement

3. Problem Statement

Lip reading, or visual speech recognition, is the process of interpreting spoken language by analyzing the movements of a speaker's lips. The primary challenge in lip reading lies in accurately decoding speech from visual input alone, which is inherently ambiguous and affected by various factors such as lighting, facial expressions, and individual speaking styles. Traditional lip reading methods have addressed this challenge in two separate stages: feature extraction and prediction. While these approaches have demonstrated some success, they often struggle with limitations in handling temporal dynamics and contextual information.

Recent advancements have introduced end-to-end deep learning models that aim to integrate both stages into a unified framework. However, most existing end-to-end models are designed for word-level classification rather than full sentence-level sequence prediction. This limitation restricts their ability to leverage the temporal context of longer sequences, which is crucial for accurate and meaningful interpretation of spoken language.

Developing a robust, end-to-end lip reading system that effectively integrates spatiotemporal features and sequence modeling has significant implications. It could enhance assistive technologies for individuals with hearing impairments, improve security systems that rely on visual speech data, and contribute to advancements in human-computer interaction. By addressing the current limitations and pushing the boundaries of automated lip reading, this research aims to make meaningful contributions to the field of visual speech recognition.