

---

---

# Trace Talk - A Lip Reading AI

---

---

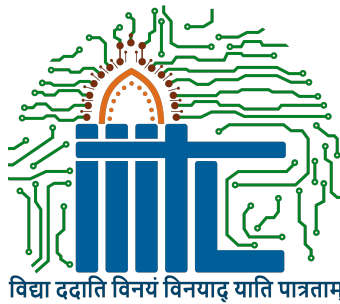
*A project report submitted in partial fulfillment of the requirements for the  
award of the degree of*

**B.Tech. in Computer Science**

by

**Adithi | Ashutosh | Nabeela | Sphoorthy | Vaibhav**  
**LCI2021019 | LCS2021024 | LCS2021018 | LCS2021053 | LCI2021033**

under the guidance of  
**Dr.Saurabh Shukla**



**Indian Institute of Information Technology, Lucknow**  
**May 2024**

© Indian Institute of Information Technology, Lucknow 2024.








# Declaration of Authorship

We declare that the work presented in “Trace Talk” is our own. We confirm that:

- This work was completed entirely while in candidature for B.Tech. degree at Indian Institute of Information Technology, Lucknow.
- Wherever we have consulted the published work of others, it is always cited.
- Wherever we have cited the work of others, the source is always indicated. Except for the aforementioned quotations, this work is solely our work.
- We have acknowledged all major sources of information.

Signed:

      
(Adithi) (Ashutosh) (Nabeela) (Sphoorthy) (Vaibhav)

Date: 2nd May 2024

---



# CERTIFICATE

This is to certify that the work entitled “**Trace Talk**” submitted by **Adithi, Ashutosh, Nabeela, Sphoorthy, Vaibhav** who got his/her name registered on **Dec 2021** for the award of B.Tech. degree at Indian Institute of Information Technology, Lucknow is absolutely based upon his/her own work under the supervision of **Dr.Saurabh Shukla**, HOD of Computer Science, Indian Institute of Information Technology Lucknow - 226 002, U.P, India and that neither this work nor any part of it has been submitted for any degree or any other academic award anywhere before.



Dr Saurabh Shukla  
Department of Computer Science  
Indian Institute of Information Technology, Lucknow  
Pin - 226 002, INDIA



# Acknowledgements

"We would like to express our sincere gratitude to all those who have contributed to the successful completion of this B.Tech project. First and foremost, We would like to thank our project supervisor Dr. Saurabh Shukla, for his invaluable guidance, support, and encouragement throughout the project. His insights and expertise were crucial in shaping the direction of the project. We are grateful to our institute IIIT LUCKNOW for providing us with an opportunity to undertake this project. The knowledge and skills we have acquired during this project will undoubtedly be invaluable in our future endeavors." We extend our heartfelt thanks to all those who have played a part in making this project a success.

Lucknow  
May 2024

Adithi | Ashutosh | Nabeela | Sphoorthy | Vaibhav





# ABSTRACT

The goal of this project is to build a speech recognition system that can accurately recognize spoken words from a set of predefined words and also has an added feature of web/file browsing based on the spoken words. The algorithm used to recognize words uses computer vision and deep learning, and is trained on a large dataset generated by us. The model architecture includes several convolutional and dense layers, and was trained using TensorFlow( Keras). The training process achieved a training accuracy of 97.4%, and testing accuracy was 96%, demonstrating strong classification performance. Once trained, the system can be used to recognize spoken words in a live setting.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	1
1.2	Features . . . . .	2
1.2.1	Real Time Prediction . . . . .	2
1.2.2	Web/File Browsing . . . . .	2
<b>2</b>	<b>Literature Review</b>	<b>3</b>
2.1	Background and Evolution of Speech Recognition Systems .	3
2.2	Integration of Computer Vision in Speech Recognition . . . .	3
2.3	Deep Learning Approaches in Speech Recognition . . . . .	4
2.4	Datasets and Training Methodologies . . . . .	4
2.5	Tools and Frameworks . . . . .	4
2.6	Future Directions and Challenges . . . . .	4
<b>3</b>	<b>Methodology</b>	<b>5</b>
3.1	Technical Specifications . . . . .	5
3.1.1	Data Collection and Preprocessing . . . . .	5
3.1.2	Model Architecture . . . . .	6
3.1.3	Image Processing Techniques . . . . .	6
3.2	Model Training and Evaluation . . . . .	6
3.2.1	Challenges and Solutions . . . . .	7
3.2.2	Conclusion . . . . .	7
<b>4</b>	<b>Simulation and Results</b>	<b>9</b>
4.1	Model Evaluation Metrics . . . . .	9
4.1.1	Accuracy . . . . .	9
4.1.2	Balanced Accuracy . . . . .	9
4.1.3	Precision . . . . .	10
4.1.4	Recall . . . . .	10
4.1.5	F1 Score . . . . .	10
4.2	Model Training . . . . .	10

4.2.1	Epochs and Training Progress . . . . .	10
4.2.2	Training and Testing Accuracy . . . . .	11
4.3	Confusion Matrix Analysis . . . . .	11
4.4	Performance Metrics . . . . .	12
4.4.1	Precision, Recall, and F1 Score . . . . .	12
4.4.2	Balanced Accuracy . . . . .	12
4.5	ROC AUC Curve Analysis . . . . .	12
<b>5</b>	<b>Conclusion and Future Scope</b>	<b>13</b>
5.1	Conclusion . . . . .	13
5.2	Future Scope . . . . .	14
5.2.1	Deep Learning for Vocabulary Expansion . . . . .	14
5.2.2	Contextual Awareness with Attention Mechanisms .	14
5.2.3	Multimodal Fusion: Speechreading and Sign Lan- guage Recognition . . . . .	14
<b>A</b>	<b>Appendix</b>	<b>15</b>
A.1	Training a custom dlib shape predictor: . . . . .	15
A.2	Web Browsing Functionality Integration: . . . . .	15
A.3	Additional Resources . . . . .	15

# Chapter 1

## Introduction

In a world where lip reading is traditionally a challenging skill even for humans, our model 'Trace Talk' marks a significant breakthrough. We have developed a model using artificial intelligence, machine learning, and deep learning algorithms to decipher spoken words based solely on the visual movements of a speaker's lips.

At the heart of Trace Talk lies a complex network of algorithms meticulously trained with our own datasets of lip movements and corresponding speech patterns. Through iterative refinement and optimization, our model has achieved remarkable accuracy and robustness, surpassing traditional human lip-reading capabilities in many cases.

As we continue to refine and expand the capabilities of Trace Talk, we envision a future where communication barriers are overcome, and inclusivity is prioritized. With each advancement, we move closer to a world where everyone, regardless of their auditory abilities, can fully participate and engage in the rich tapestry of human communication.

### 1.1 Motivation

Human lipreading performance suffers due to ambiguity in lip movements, speech variability, visual clarity issues, and high cognitive load. Previous lipreading methods lacked deep learning techniques, often relying on heavy preprocessing for video feature extraction.

We have incorporated a real-time prediction feature into our project which significantly enhances functionality, enabling instantaneous recognition and interpretation of spoken words through lip movements. Utilizing a convolutional neural network (CNN) architecture trained on labeled lip images, our model, Tracetalk, dynamically analyzes live video streams to

discern spoken words in real-time by continuously processing frames and assessing lip movement patterns.

## **1.2 Features**

### **1.2.1 Real Time Prediction**

We have incorporated a real-time prediction feature into our project which significantly enhances functionality, enabling instantaneous recognition and interpretation of spoken words through lip movements. Utilizing a convolutional neural network (CNN) architecture trained on labeled lip images, our model, Tracetalk, dynamically analyzes live video streams to discern spoken words in real-time by continuously processing frames and assessing lip movement patterns.

### **1.2.2 Web/File Browsing**

Our model leverages the web browser library for application navigation and utilizes the OS library for file browsing, all powered by lip reading. This method offers the advantage of quicker interaction compared to typing, and it proves particularly useful in environments where microphone usage is impractical due to noise or other constraints.

# Chapter 2

## Literature Review

Speech recognition systems have witnessed significant advancements in recent years, driven by the integration of computer vision and deep learning techniques. This review aims to contextualize the development of speech recognition systems and identify key insights and methodologies relevant to the project's goal of building an accurate word recognition system.

### 2.1 Background and Evolution of Speech Recognition Systems

Early speech recognition systems relied primarily on statistical modeling and pattern recognition algorithms, which often struggled with variations in speech patterns and background noise. However, with the advent of deep learning, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), significant improvements in accuracy and robustness have been achieved.

### 2.2 Integration of Computer Vision in Speech Recognition

The integration of computer vision techniques in speech recognition systems has emerged as a promising approach to enhance accuracy and reliability. By analyzing visual cues such as lip movements and facial expressions, researchers have demonstrated improved performance in word recognition tasks, especially for individuals with speech impairments.

## **2.3 Deep Learning Approaches in Speech Recognition**

Deep learning has revolutionized speech recognition, offering superior performance compared to traditional statistical methods. CNNs are commonly used to extract spatial features from visual input, such as lip images, while RNNs, including LSTM networks, capture temporal dependencies in speech sequences. The combination of these architectures enables robust and context-aware word recognition.

## **2.4 Datasets and Training Methodologies**

The availability of large, annotated datasets is critical for training robust speech recognition models. While public datasets such as the TIMIT dataset have been widely used in the past, recent advancements have seen researchers generating custom datasets tailored to specific tasks and domains. Training methodologies, including data augmentation and transfer learning, have also been instrumental in improving model generalization and performance.

## **2.5 Tools and Frameworks**

Frameworks like TensorFlow and Keras have become the go-to tools for building and training deep learning models, offering flexibility, scalability, and ease of use. These frameworks provide a wide range of pre-trained models and optimization algorithms, streamlining the development process for speech recognition systems.

## **2.6 Future Directions and Challenges**

While significant progress has been made in speech recognition technology, challenges remain, particularly in handling variations in accents, languages, and speech styles. Future research directions include exploring multi-modal approaches integrating audio and visual cues, as well as developing more robust and efficient models for real-time applications.



# Chapter 3

## Methodology

This chapter delineates the methodology employed in constructing a robust lip reading model aimed at accurately identifying spoken words from a predefined set. The system harnesses computer vision and deep learning techniques to achieve its objectives, with a substantial focus on model training and evaluation.

### 3.1 Technical Specifications

The lip reading model is underpinned by a convolutional neural network (CNN) architecture, augmented with dense layers for classification. TensorFlow and Keras serve as the primary libraries for model development, providing a robust framework for implementing deep learning algorithms. The model is trained on a meticulously curated dataset prepared by our group.

#### 3.1.1 Data Collection and Preprocessing

**Libraries:**

The project leverages various Python libraries, including TensorFlow, Keras, OpenCV, PIL, numpy, scikit-learn, matplotlib, and seaborn. These libraries facilitate data collection, preprocessing, model development, and performance evaluation.

**Dataset:**

A bespoke dataset is assembled for model training, consisting of several video clips capturing word enunciations. Each video clip is manually labeled with a word from a predefined set, covering a diverse range of vo-

cabulary. The dataset is evenly distributed across different words, ensuring adequate representation for each class.

### 3.1.2 Model Architecture

#### **3D Convolutional Neural Network (CNN):**

The heart of the speech recognition system is a 3D CNN architecture tailored to capture spatiotemporal features from video frames. The model comprises three Conv3D layers followed by MaxPooling3D layers for feature extraction. Subsequent dense layers, including Dropout layers, facilitate classification. L2 regularization is incorporated into Conv3D layers to mitigate overfitting..

### 3.1.3 Image Processing Techniques

#### **Lip Segmentation:**

A crucial preprocessing step involves segmenting lip movements from the video frames using DLIB's Face Detector model. This segmentation isolates the lips and surrounding areas, essential for subsequent processing.

#### **Gaussian Blurring:**

Gaussian blurring is applied to the segmented lips to reduce noise and enhance image quality, facilitating smoother gradient descent during model training.

#### **Contrast Stretching:**

Contrast stretching enhances the contrast between darker and lighter pixels, making lip movements more discernible and aiding the model in feature extraction.

#### **Bilateral Filtering:**

Bilateral filtering smoothens the image while preserving edges, reducing noise and improving overall image clarity, thereby enhancing the model's performance.

#### **Sharpening:**

The sharpening technique accentuates object edges, making them more defined and aiding in object recognition. This step enhances the model's ability to detect and classify lip movements accurately.

## 3.2 Model Training and Evaluation

#### **Training Process:**

The model is trained using the curated dataset, with 80% allocated for training and 20% for testing. Training parameters, including batch size, learning rate, and optimizer choice, are optimized to maximize performance..

**Evaluation Metrics:**

Training and testing accuracy serve as primary evaluation metrics, providing insights into the model's classification performance. Quality assessment involves comparing generated images with industry standards and user expectations.

### **3.2.1 Challenges and Solutions**

**Technical Challenges:**

Several challenges, such as hardware limitations and dataset creation, were encountered during the project. Solutions involved adjusting threshold values for optimal lip distance and maintaining accuracy while implementing iterative training with augmented datasets to enhance model robustness.

### **3.2.2 Conclusion**

The methodology outlined in this chapter lays the groundwork for developing a sophisticated lip reading model. By integrating computer vision, deep learning, and image processing techniques, the system demonstrates promising capabilities in accurately recognizing spoken words, contributing to the advancement of assistive technologies for the hearing impaired.



# Chapter 4

## Simulation and Results

This chapter presents the simulation and results of the key components of our lip reading project, "**TraceTalk**". Each subsection unveils the outcomes of distinct features and methodologies employed within the platform. Through a combination of quantitative analysis and qualitative assessment, we delve into the accuracy, robustness, and real-world applicability of our lip reading model. Additionally, accompanying visual representations provide a comprehensive view of the results, offering insights into the efficacy and performance of our system across various scenarios and environments.

### 4.1 Model Evaluation Metrics

#### 4.1.1 Accuracy

Accuracy stands as a fundamental metric, gauging the model's performance on unseen data. It measures the proportion of correct predictions against the actual labels. In this lip reading project, we rely on accuracy to ascertain the model's general effectiveness.

#### 4.1.2 Balanced Accuracy

Balanced Accuracy offers a nuanced perspective by considering the disparities in data distributions. Although our dataset primarily maintains balance, this metric ensures a fair evaluation, particularly crucial in scenarios where class imbalances are present.

### **4.1.3 Precision**

Precision serves as a marker of the model's confidence, calculated as the ratio of True Positives to the sum of False Positives plus True Positives ( $P = TP / (TP + FP)$ ). A higher precision indicates greater confidence in the model's predictions.

### **4.1.4 Recall**

Similar to precision, recall quantifies the model's ability to capture positive instances, expressed as the ratio of True Positives to the sum of False Negatives plus True Positives ( $R = TP / (TP + FN)$ ). It highlights the model's sensitivity to positive samples.

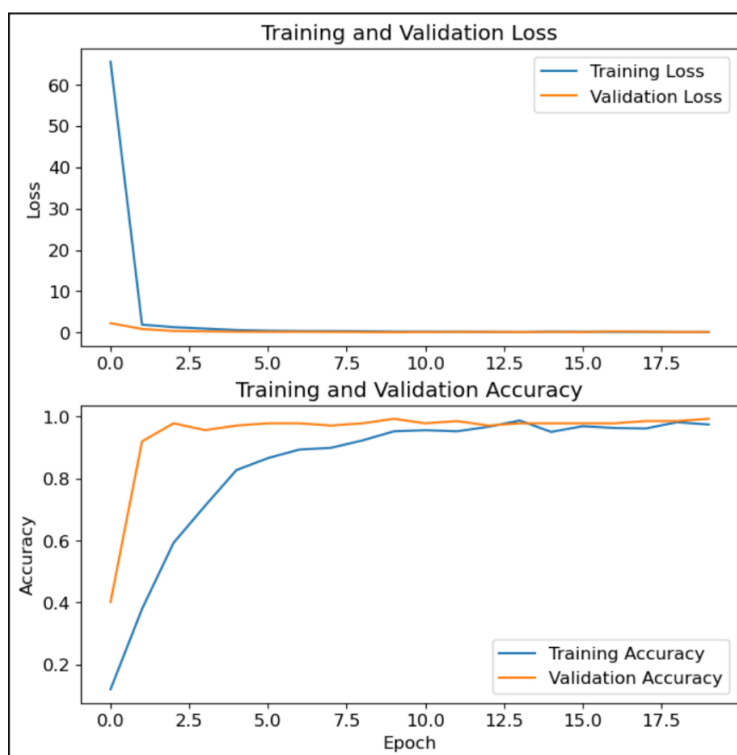
### **4.1.5 F1 Score**

The F1 Score harmonizes precision and recall, offering a balanced assessment of the model's performance. Calculated as the weighted average of precision and recall ( $2 * (Precision * Recall) / (Precision + Recall)$ ), it serves as a middle ground between the two metrics, ensuring a comprehensive evaluation.

## **4.2 Model Training**

### **4.2.1 Epochs and Training Progress**

The 3D CNN underwent training for several epochs, allowing for iterative refinement of its parameters. This extensive training period facilitated the model's ability to capture intricate patterns in lip movements, enhancing its overall performance. The accompanying graph illustrates the model's progression throughout the training phase, providing insights into its learning trajectory and the convergence of its training process.



## 4.2.2 Training and Testing Accuracy

Upon completion of training, the model achieved a final training accuracy of 97.4 percent and a testing accuracy of 96 percent. Testing accuracy serves as a pivotal metric, indicating the model's efficacy in handling unseen data, crucial for real-world applications.

The Testing accuracy of our model can be seen in the demo link provided below:

<https://youtu.be/Embb-n3pUPA?si=LD7do1H1FPNXk3z1>

## 4.3 Confusion Matrix Analysis

If we examine the confusion matrix generated using the testing data, we can observe several instances where the model made incorrect predictions. Notably, the model frequently confused visually similar words such as "hello" with "demo" or "my" with "bye". These challenges underscore the importance of evaluating the model's ability to discern subtle differences in lip movements, particularly in contexts where similar phonemes occur.

## **4.4 Performance Metrics**

### **4.4.1 Precision, Recall, and F1 Score**

The model exhibits exceptional precision, recall, and F1 scores, predominantly at 1.0 for most classes. These metrics underscore the model's accuracy in identifying positive samples and distinguishing between different lip movements.

Additionally, we achieved a balanced accuracy of 98 percent, indicating the model's robustness in handling varying data distributions across different scenarios.

### **4.4.2 Balanced Accuracy**

With a balanced accuracy of 98 percent, the model demonstrates robustness in handling varying data distributions, reaffirming its reliability across different scenarios.

## **4.5 ROC AUC Curve Analysis**

The ROC AUC curve analysis provides further insights into the model's discriminatory power and overall performance. The consistently high AUC values suggest that the model effectively distinguishes between positive and negative instances with minimal false positives.

Moreover, the steep rise towards the top-left corner of the curve indicates high true positive rates at low false positive rates, highlighting the model's efficacy in various lip reading tasks.



# Chapter 5

## Conclusion and Future Scope

### 5.1 Conclusion

This project utilizes computer vision and deep learning techniques to develop an algorithm capable of translating lip movements into spoken words and also has an additional feature of web/file browsing making it easy to open any web application or file in your system without the need to type or even use your voice for it.

By leveraging transfer learning, a dataset comprising 700 video clips depicting words being spoken was curated. Each clip underwent various image processing techniques, including lip segmentation, Gaussian blurring, contrast stretching, bilateral filtering, and sharpening.

The model architecture, consisting of convolutional and dense layers, achieved impressive training and validation accuracies of 97.4 percent and 96 percent, respectively. Its lightweight nature allows for real-time recognition of spoken words once trained.

This project serves as a testament to the potential of technology in addressing real-world challenges and promoting inclusivity. TraceTalk has the capability to revolutionize interactions for individuals with hearing and voice impairments, enabling them to engage in meaningful conversations and participate actively in various social settings. Through the integration of deep learning and computer vision, TraceTalk lays the groundwork for future advancements in communication accessibility, fostering improved connection and understanding.

## 5.2 Future Scope

### 5.2.1 Deep Learning for Vocabulary Expansion

**Current Approach:** Utilizing a predefined word set.

**Future Direction:** Implement deep learning architectures such as recurrent neural networks (RNNs) or transformers to process sequential lip movements and recognize a broader vocabulary. This entails training on extensive datasets of labeled speech videos.

### 5.2.2 Contextual Awareness with Attention Mechanisms

**Current Approach:** Limited contextual understanding.

**Future Direction:** Integrate attention mechanisms into deep learning models to focus on relevant contextual cues such as surrounding text or facial expressions. This could involve leveraging natural language processing (NLP) techniques and exploring transformers with self-attention for exploiting contextual relationships.

### 5.2.3 Multimodal Fusion: Speechreading and Sign Language Recognition

**Current Approach:** Standalone functionality.

**Future Direction:** Integrate TraceTalk with sign language recognition systems. This requires exploring multimodal learning techniques capable of effectively processing both lip movements and hand gestures, potentially utilizing deep learning architectures designed for multimodal data.

# Appendix A

## Appendix

This appendix contains additional information and resources related to our lip reading model.

### A.1 Training a custom dlib shape predictor:

Details regarding creation of shape predictor for our lip reading application are provided here.

<https://pyimagesearch.com/2019/12/16/training-a-custom-dlib-shape-predictor/>

### A.2 Web Browsing Functionality Integration:

Details regarding the integration of web browsing technology into our lip reading application are provided here.

<https://docs.python.org/3/library/webbrowser.html>

### A.3 Additional Resources

For additional information about the technology utilized in the project and relevant resources, please refer to the provided link:

<https://www.tensorflow.org/guide/keras>

<https://docs.streamlit.io/develop/tutorials>



# Bibliography

- [1] N. Deshmukh, A. Ahire, S. H. Bhandari, A. Mali, and K. Warkari. Vision based lip reading system using deep learning. In *2021 International Conference on Computing, Communication and Green Engineering (CCGE)*, pages 1–6, 2021. doi: 10.1109/CCGE50943.2021.9776430.
- [2] A. H. Kulkarni and D. Kirange. Artificial intelligence: A survey on lip-reading techniques. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–5, 2019. doi: 10.1109/ICCCNT45670.2019.8944628.
- [2] A. H. Kulkarni and D. Kirange, "Artificial Intelligence: A Survey on Lip-Reading Techniques," 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, 2019, pp. 1-5, doi: 10.1109/ICCCNT45670.2019.8944628. keywords: Lips;Hidden Markov models;Visualization;Speech recognition;Feature extraction;Deep learning;Data models;Learning systems;Neural networks;Artificial intelligence;Speech Recognition;Databases,
- [1] N. Deshmukh, A. Ahire, S. H. Bhandari, A. Mali and K. Warkari, "Vision based Lip Reading System using Deep Learning," 2021 International Conference on Computing, Communication and Green Engineering (CCGE), Pune, India, 2021, pp. 1-6, doi: 10.1109/CCGE50943.2021.9776430. keywords: Location awareness;Deep learning;Lips;Mouth;Computer architecture;Feature extraction;Robustness;CNN;RNN;LSTM;Attention Mechanism;automatic lip reading;deep learning,