

# Visual Speech Recognition for Seamless Communication with Hearing Impaired Persons

Md. Laraib Ahmad , Scholar ID: 2322207

## **Supervisor**

Dr. Debbrota Paul Chowdhury



Department of Computer Science and Engineering  
National Institute of Technology Silchar

December 13, 2024

- ① Introduction
- ② Research Advancement
- ③ Motivation
- ④ Literature Review
- ⑤ Research Gap
- ⑥ Problem Statement
- ⑦ Objective
- ⑧ Available Datasets
- ⑨ Dataset Preparation for Visual Speech Recognition System
- ⑩ Generalised Block Diagram
- ⑪ Methodology
- ⑫ conclusion
- ⑬ Future Work
- ⑭ References

## What is Lip Reading:



fig no - 1: <https://livingwithhearingloss.com/2016/04/19/lipreading-in-paradise/>

# Why Lip Reading is important?

Communication aid for the deaf and hard-of-hearing

- Primary Communication Tool
- Complement to hearing aids and cochlear implants



<https://www.inc.com/john-boitnott/this-entrepreneur-is-solving-one-of-the-biggest-problems-all-deaf-people-face.html>

<https://www.connecthear.org/post/all-about-cochlear-implants>

cont...

## Laryngeal cancer

- laryngeal cancer can result in loss of speech or changes in the voice, especially if the cancer or its treatment affects the vocal cords or requires the removal of parts of the larynx. In such case speaker can only move their lips to communicate.



<https://utswmed.org/medblog/cold-flu-allergy-hurt-your-voice/>

Time Period	Advancement
Early 20th century	The concept of lip reading as a skill for the deaf and hard-of-hearing began to be formally studied and taught. Schools for the deaf often included lip reading in their curricula.
1970s	Early research in automated lip reading began. Initial efforts focused on understanding the visual aspects of speech and how they could be captured and analyzed by computers.
1980s and 1990s	Advances in computer vision and pattern recognition led to more sophisticated experiments with automated lip reading systems. Algorithms to recognize visual speech elements were developed.
2000s	Deep learning and AI started to significantly improve automated lip reading accuracy. Researchers used statistical models like Hidden Markov Models (HMMs) to analyze visual speech data.
2010s	Deep learning techniques, especially CNNs and RNNs, began to be applied to lip reading, leading to substantial improvements. Large datasets and improved computational power also contributed to advancements.
2016	Google DeepMind's LipNet model demonstrated high accuracy in lip reading by leveraging deep learning techniques.
2020s	Ongoing research continues to refine and improve lip reading technologies, which are now applied in fields such as assistive technology for the deaf and hard-of-hearing, security, and human-computer interaction.

- Accessibility for Hearing-Impaired Individuals
- Speech Enhancement in Noisy Environments
- Multimodal Systems
- Advances in AI and Machine Learning

Author, year	Methodology	Dataset Used	Findings
Siddiqui <i>et al.</i> , [2022] [1]	A Multi-SVM classifier categorizes the lip movements to recognize spoken words.	custom-made by the authors.	<ul style="list-style-type: none"><li>The proposed lip reading system, based on visual cues alone, can effectively recognize words with an accuracy of 75%.</li></ul>



Author, year	Methodology	Dataset Used	Findings
Freitas <i>et al.</i> , 2016 [2]	LipNet model: spatiotemporal convolutional neural networks (STCNNs) to extract spatial and temporal features, followed by Bidirectional Gated Recurrent Units (Bi-GRUs) to capture temporal dependencies, and employs Connectionist Temporal Classification (CTC) loss for end-to-end training without the need for pre-segmented data	GRID corpus	<ul style="list-style-type: none"> <li>• Achieving a 95.2% sentence-level accuracy on the GRID corpus.</li> <li>• The study highlights the effectiveness of combining spatiotemporal convolutions, recurrent networks, and Connectionist Temporal Classification (CTC) for sentence-level prediction, marking a major improvement in automated lipreading</li> </ul>

Author, year	Methodology	Dataset Used	Findings
Zimmermann <i>et al.</i> , 2020[3]	The paper employs two methods Temporal Conditional GANs (TC-GANs) to generate lip movement videos for unseen utterances and a viseme-concatenation approach to synthesize videos by mapping phonemes to visemes to enable zero-shot learning in visual speech recognition.	OuluVS2 dataset	Using GANs for zero-shot learning significantly improves visual speech recognition accuracy for unseen utterances, effectively addresses the cold-start problem, and generalizes to new languages, with GANs outperforming the viseme-concatenation approach.

Author, year	Methodology	Dataset Used	Findings
XIAO <i>et al.</i> , 2020[4]	The methodology involves preprocessing video frames to extract lip regions, using a spatial-temporal CNN to generate features, applying a transformer-based model to classify visemes, and converting visemes to words through perplexity analysis for sentence prediction.	BBC LRS2 dataset.	The paper finds that the proposed viseme-based lip reading system significantly improves word accuracy with a 15% reduction in Word Error Rate (WER), achieves a Viseme Error Rate (VER) of 4.6%, and demonstrates robustness to varying lighting conditions, though further optimization is needed in converting visemes to words

Author, year	Methodology	Dataset Used	Findings
Xie <i>et al.</i> , 2024[5]	The methodology involves multi-scale lip motion video extraction, dynamic augmentation, and an end-to-end VSR system with multi-system fusion using diverse encoders for optimal visual speech recognition performance.	The paper uses the <b>**CN-CVS**</b> dataset for training, along with the development sets of <b>**CNVSRC-Single/Multi**</b> datasets from the Chinese Continuous Visual Speech Recognition Challenge (CNVSRC) 2023.	The paper finds that the proposed multi-system VSR approach with E-Branchformer encoder and ROVER fusion achieves leading performance with 34.76% CER in the Single-Speaker Task and 41.06% CER in the Multi-Speaker Task, securing first place in all three CNVSRC 2023 tracks.

Author, year	Methodology	Dataset Used	Findings
Pantic et al., 2022[6]	The methodology involves enhancing VSR performance through prediction-based auxiliary tasks, hyperparameter optimization, data augmentation (like time-masking), and pre-training/fine-tuning across multiple languages.	The paper uses the LRS2, LRS3, CMLR (Mandarin), and CMU-MOSEAS (Spanish) datasets for training and evaluation, with a focus on publicly available datasets for achieving state-of-the-art VSR performance across multiple languages. Additionally, the LRW and AVSpeech datasets are used in some experiments for further improvements.	The paper finds that careful model design, including prediction-based auxiliary tasks, data augmentation, and hyperparameter optimization, can significantly improve visual speech recognition performance, even surpassing models trained on much larger datasets.

Author, year	Methodology	Dataset Used	Findings
GUO <i>et al.</i> , 2020[7]	The methodology involves using a viseme-to-word conversion system with perplexity analysis, where visual speech input is processed through word lookup, chunkification, and iterative beam search to identify the most likely word sequences based on a pre-trained language model.	<p>The paper uses two datasets for experimentation:</p> <ul style="list-style-type: none"> <li>• OuluVS Dataset: This consists of short phrases like "hello," "excuse me," "I am sorry," etc.</li> <li>• BBC LRS2 Dataset: This contains longer and more varied sentences from BBC videos, making it more challenging due to a wide range of speakers and vocabulary</li> </ul>	The findings show that the model effectively predicts short phrases with 100% accuracy and performs reasonably well on longer sentences using perplexity analysis, though it struggles with increased errors when word boundaries are unknown.

Author, year	Methodology	Dataset Used	Findings
Ro <i>et al.</i> , 2024[8]	The methodology used in the paper combines vision-level adaptation (using padding prompts and LoRA for lip appearances, movements, and speaking speed) and language-level adaptation (using input prompt tuning to learn speaker-specific linguistic patterns) to adapt a pre-trained lip-reading model to target speakers effectively.	VoxLRS-SA dataset, which is derived from Vox-Celeb2 and LRS3 datasets.	The paper demonstrates that integrating vision and language-level adaptations improves speaker-specific lip reading performance, surpassing previous methods.

Author, year	Methodology	Dataset Used	Findings
Pantic et al., 2022[9]	The methodology combines advanced temporal models (like DC-TCN), effective data augmentations (Time Masking, mixup), and training strategies (self-distillation, word boundary indicators) to systematically enhance lip-reading performance.	LRW dataset	The paper finds that combining advanced temporal models (DC-TCN), Time Masking augmentation, and training strategies like self-distillation and word boundary indicators achieves state-of-the-art performance in lip-reading, particularly improving recognition of difficult words.



- Accuracy is not very high for word prediction.
- No dataset available for different accents.
- Viseme-based Challenges
- Cross-lingual Transfer Learning



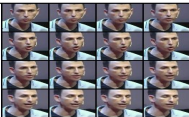

Despite significant progress, visual speech recognition faces challenges such as variability in lip movements, diverse speaking styles, and need for large and labeled datasets for training. Ongoing research aims to address these challenges and further refine technology making it more accurate, reliable and widely applicable.




- To create a dataset of different accents.
- To develop an algorithm that can recognize words correctly independent of accents.

Dataset	Type	Description
<b>GRID Corpus</b>	Sentence-level	Contains 34 speakers uttering structured sentences with fixed vocabulary (1000 unique sentences).
<b>LRS2 (Lip Reading Sentences 2)</b>	Sentence-level	Contains over 224 hours of data from BBC programs with spoken sentences for audio-visual speech recognition.
<b>LRS3 (Lip Reading Sentences 3)</b>	Sentence-level	Larger version of LRS2 with over 475 hours of videos for lip reading in challenging conditions.
<b>TCD-TIMIT</b>	Continuous Speech	Phonetically balanced dataset with 59 speakers reading 98 sentences, suitable for continuous speech recognition.
<b>AVLetters</b>	Alphabet-level	Dataset with speakers uttering letters A-Z multiple times for isolated letter recognition.
<b>LRW (Lip Reading in the Wild)</b>	Word-level	Contains over 500 different words spoken by various speakers extracted from TV broadcasts.
<b>OuluVS2</b>	Phrase-level	Contains 53 speakers saying 10 phrases, repeated 6 times per phrase, for small-scale phrase recognition.

Table 1: Available Datasets for Visual Speech Recognition by Lip Reading

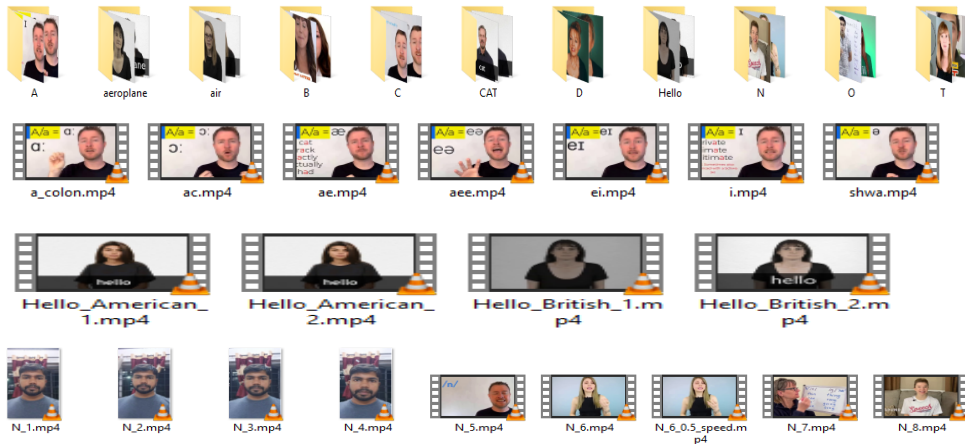
# Dataset Names and Images for Visual Speech Recognition

Dataset	Image
GRID Corpus	
LRS2 (Lip Reading Sentences 2)	
LRS3 (Lip Reading Sentences 3)	
TCD-TIMIT	

Dataset	Image
AVLetters	
LRW (Lip Reading in the Wild)	
OuluVS2	

# Dataset Preparation for Visual Speech Recognition System

I created dataset by clipping videos from youtube and recodring my own videos. For word pronunciation collins video on youtube is great resource.



# Generalized block diagram of visual speech recognition system

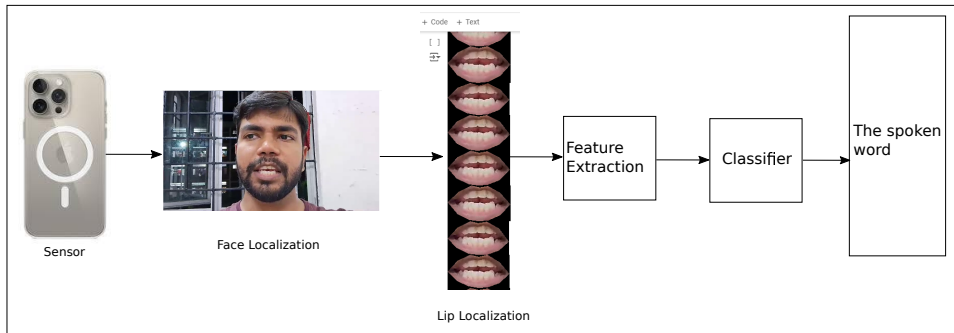


Figure 1: Generalized block diagram of visual speech recognition system

---

## Algorithm 1 Normalize Lip Points

---

**Require:** A map *lipPoints* of integers to coordinate pairs  $(x, y)$

**Ensure:** A normalized map of lip points

```
1: if  $48 \in \text{lipPoints}$  then
2:    $\text{origin} \leftarrow \text{lipPoints}[48]$  ▷ Set the origin point
3:    $\text{normalizedPoints} \leftarrow \{\}$  ▷ Initialize an empty map for normalized points
4:   for all  $(\text{key}, (x, y)) \in \text{lipPoints}$  do ▷ Iterate over each point
5:      $\text{normalizedPoints}[\text{key}] \leftarrow (x - \text{origin.first}, y - \text{origin.second})$ 
6:   end for
7:   return normalizedPoints
8: else
9:   return lipPoints ▷ Return original points if key 48 is not found
10: end if
```

---



---

**Algorithm 2** Euclidean Distance Calculation

---

**Require:**  $x_1, y_1, x_2, y_2$

**Ensure:**  $dx \leftarrow x_2 - x_1$ ,

1:  $dy \leftarrow y_2 - y_1$

2:  $distance \leftarrow \sqrt{dx^2 + dy^2}$

---

---

**Algorithm 3** Calculate Average Lip Point Distance

---

**Require:** Two maps *lipPoints1* and *lipPoints2* of integers to coordinate pairs  $(x,y)$

**Ensure:** The average Euclidean distance between corresponding points in *lipPoints1* and *lipPoints2*

```
1: totalDistance  $\leftarrow$  0.0
2: for  $i = 48$  to 67 do
3:   if  $i \in \text{lipPoints1}$  and  $i \in \text{lipPoints2}$  then
4:     totalDistance  $\leftarrow$  totalDistance + euclideanDistance(lipPoints1[ $i$ ], lipPoints2[ $i$ ])
5:   end if
6: end for
7: return totalDistance/20
```

---

---

## Algorithm 4 Extract and Normalize Lip Points from a Video

---

**Require:** videoPath: Path to the video file

detector: Frontal face detector

predictor: Shape predictor

**Ensure:** lipPointsAllFrames: Vector of normalized lip points for each frame

```
1: Initialize lipPointsAllFrames as an empty vector
2: Open the video using VideoCapture
3: if the video cannot be opened then
4:   Print error and return lipPointsAllFrames
5: end if
6: while frames are available do
7:   Read the current frame and convert to grayscale
8:   Detect faces using detector
9:   if at least one face is detected then
10:    Extract landmarks using predictor
11:    for each index  $i$  from 48 to 67 do
12:      Store landmarks in lipPoints
13:    end for
14:    Normalize lipPoints
15:    Append lipPoints to lipPointsAllFrames
16:  else
17:    Append empty map to lipPointsAllFrames
18:  end if
19: end while
20: Release the video resource
21: return lipPointsAllFrames
```

## Algorithm 5 Calculate Average Distance to Reference Videos

**Require:** newVideoLipPoints: Vector of lip points for the new video

referenceVideosLipPoints: Vector of lip points for the reference videos

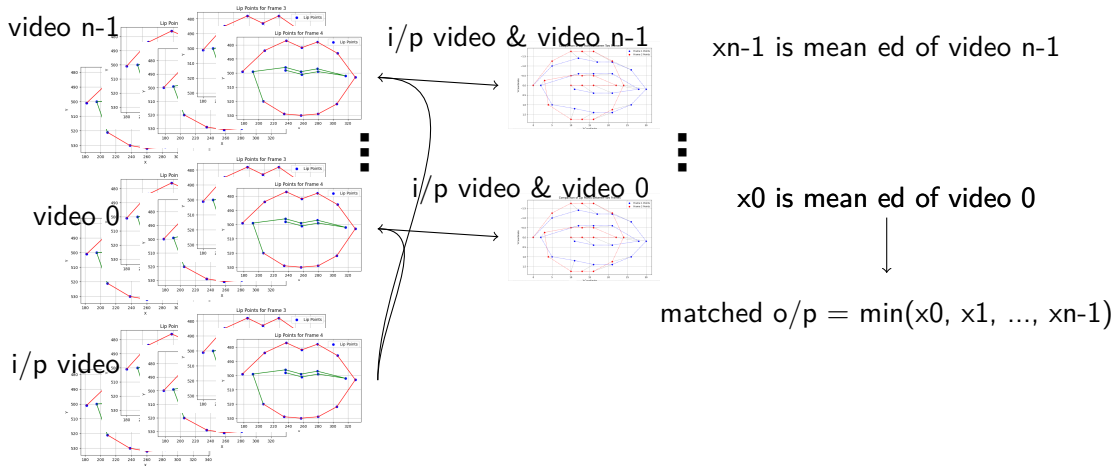
**Ensure:** averageDistance: The average distance to the reference videos

```

1: Initialize minFrames as the size of newVideoLipPoints
2: for each reference video refVideo in referenceVideosLipPoints do
3:   Update minFrames as the minimum between minFrames and the size of refVideo
4: end for
5: Initialize totalDistances as an empty vector
6: for each frame index  $i$  from 0 to minFrames - 1 do
7:   Get the lip points for the current frame from newVideoLipPoints
8:   if lip points for the current frame are not empty then
9:     Initialize distancesForFrame as an empty vector
10:    for each reference video refVideo in referenceVideosLipPoints do
11:      if  $i$  is less than the size of refVideo and lip points for this frame are not empty then
12:        Calculate the distance between lipPointsFrame and the corresponding frame in refVideo using calculateLipPointDistance
13:        Add the calculated distance to distancesForFrame
14:      end if
15:    end for
16:    if distancesForFrame is not empty then
17:      Calculate avgDistanceForFrame as the average of distancesForFrame
18:      Append avgDistanceForFrame to totalDistances
19:    end if
20:  end if
21: end for
22: Calculate averageDistance as the average of totalDistances
23: return averageDistance

```

# Visual representation of Methodology



## Frame-wise Euclidean Distance

For two frames  $F_1$  and  $F_2$ , with  $n$  feature points, the average Euclidean distance across all feature points is given by:

$$\Psi_{\text{frame}} = \frac{1}{n} \sum_{j=1}^n \sqrt{(x_{j1} - y_{j1})^2 + (x_{j2} - y_{j2})^2}$$

Here:

- $x_{j1}$  and  $x_{j2}$  are the coordinates of the  $j$ -th feature point in frame  $F_1$ ,
- $y_{j1}$  and  $y_{j2}$  are the coordinates of the  $j$ -th feature point in frame  $F_2$ ,
- $n$  is the total number of feature points in the frame.

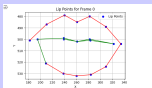
**Video-wise Euclidean Distance** For two videos  $V_1$  and  $V_2$ , each with  $M$  frames and  $n$  feature points per frame, the average Euclidean distance across all frames and their feature points is given by:

$$\psi_{\text{video}} = \frac{1}{M} \sum_{i=1}^M \frac{1}{n} \sum_{j=1}^n \sqrt{(x_{ij1} - x_{ij2})^2 + (y_{ij1} - y_{ij2})^2}$$

Here:

- $x_{ij1}$  and  $x_{ij2}$  are the coordinates of the  $j$ -th feature point in the  $i$ -th frame of video  $V_1$ ,
- $y_{ij1}$  and  $y_{ij2}$  are the coordinates of the  $j$ -th feature point in the  $i$ -th frame of video  $V_2$ ,
- $M$  is the total number of frames in each video,
- $n$  is the total number of feature points per frame.

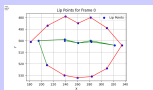
Dataset



Calculating  
Euclidean  
Distance with  
each frame of  
Hello video with  
datasets each  
video frames

```
→ Predicted word: Hello
→ Average distance to 'Cat': 3.8808612519925164
→ Average distance to 'Hello': 2.6640431595354536
→ Average distance to 'Aeroplane': 3.8843690174180057
→ Average distance to 'Air': 3.632254297618657
```

Hello spoken input video





- Successfully developed algorithm which can recognise the spoken words by reading lip without using Deep Learning.
- My algorithm is only applied on custom dataset which is created by me and accurately predicting 75% words correctly. However, Due to similar lip movements of different words(Homophone) the accuracy is falling.
- This algorithm can reduce the computational overhead as only deep learning is used for lip extraction from video not for prediction. For word prediction we are using our own developed algorithm.

# Future Work and Proposed Improvements

- Acquire videos of the same word with different accents.
- Extract 3D lip feature so that accuracy can be achieved more
- Apply our algorithm to our dataset as well publicly available dataset.
- Reading lip by using Deep Learning and comparing the result with my Algorithm
- Finding the solution for Homophones
- Compare the results with state-of-the-art.

- [1] Kunal Patil 1, Sandesh Patel 2, Harshad Rathod 3, Ashraf Siddiqui4, *LIP READING: VISUAL SPEECH RECOGNITION USING LIP READING*, International Research Journal of Engineering and Technology (IRJET), Apr 2022.
- [2] Yannis M. Assael , Brendan Shillingford, Shimon Whiteson & Nando de Freitas *LIP NET: END-TO-END SENTENCE-LEVEL LIPREADING*, Department of Computer Science, University of Oxford, Oxford, UK 1 Google DeepMind, London, UK 2 CIFAR, Canada 3, 16 Dec 2016.
- [3] Yaman Kumar, Dhruva Sahrawat, Shubham Maheshwari, Debanjan Mahata, Amanda Stent, Yifang Yin, Rajiv Ratn Shah, Roger Zimmermann, *Harnessing GANs for Zero-Shot Learning of New Classes in Visual Speech Recognition*, Cornell University, 2 Jan 2020.
- [4] SOUHEIL FENGHOUR, (Associate Member, IEEE), DAQING CHEN, (Member, IEEE),KUN GUO2 AND PERRY XIAO *Lip Reading Sentences Using Deep Learning With Only Visual Cues*, IEEE Access, 26 November 2020.
- [5] He Wang, Pengcheng Guo, Wei Chen, Pan Zhou, Lei Xie *The NPU-ASLP-LiAuto System Description for Visual Speech Recognition in CNVSRC 2023*,arXiv:29Feb2024.
- [6] Pingchuan Ma Stavros Petridis, Maja Pantic *Visual Speech Recognition for Multiple Languages in the Wild*,Imperial College London Meta AI, 13 Sep 2022.

- [7] SOUHEIL FENGHOUR, (Associate Member, IEEE), DAQING CHEN, (Member, IEEE), KUN GUO AND PERRY XIAO *DISENTANGLING HOMOPHEMES IN LIP READING USING PERPLEXITY ANALYSIS*, arXiv , 15 Dec 2020.
- [8] Jeong Hun Yeo, Chae Won Kim, Hyunjun Kim, Hyeongseop Rha, Seunghee Han, Wen-Huang Cheng, Yong Man Ro *Personalized Lip Reading: Adapting to Your Unique Lip Movements with Vision and Language*, arXiv , 2 Sep 2024.
- [9] Pingchuan Ma, Yujiang Wang, Stavros Petridis, Jie Shen, Maja Pantic *Training Strategies for Improved Lip-Reading by Koohestani and Hadfield*, arXiv , 29 Sep 2022.
- Grid Corpus Dataset - MDPI Journal  
Source: *MDPI Applied Sciences*, 2021.
  - LRS2 Dataset - Oxford VGG Group  
Source: *Visual Geometry Group*, University of Oxford.
  - LRS3 Dataset - ResearchGate  
Source: *ResearchGate*, LRS3 Dataset Overview.
  - TCD-TIMIT Dataset - ResearchGate  
Source: *ResearchGate*, TCD-TIMIT Results Overview.
  - AVLetters Database - ResearchGate  
Source: *ResearchGate*, AVLetters Database Example.

- LRW Dataset - ResearchGate  
Source: *ResearchGate*, LRW Dataset Frames.
- Oulu-VS2 Dataset - ResearchGate  
Source: *ResearchGate*, Oulu-VS2 Dataset Examples.

Thank you for listening !

Md. Laraib Ahmad

mdlaraib\_pg\_23@cse.nits.ac.in