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

September 27, 2024

Index

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

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/

Research Advancement

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.

Motivation

- 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 *et al.*, [2022] [1]	A Multi-SVM classifier categorizes the lip movements to recognize spoken words.	custom-made by the authors.	• The proposed lip reading system, based on visual cues alone, can effectively recognize words with an accuracy of 75%.

Author, year	Methodology	Dataset Used	Findings
Freitas *et al.*, 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	• Achieving a 95.2% sentence-level accuracy on the GRID corpus. • 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

Author, year	**Methodology**	**Dataset Used**	**Findings**
Zimmermann *et al.*, 2020[3]	The paper employs two methodsTemporal Conditional GANs (TC-GANs) to generate lip movement videos for unseen utterances and a viseme-concatenation approach to synthesize videos by mapping phonemes to visemesto 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 *et al.*, 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 *et al.*, 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 **CN-CVS** dataset for training, along with the development sets of **CNVSRC-Single/Multi** 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 *et al.*, 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.	The paper uses two datasets for experimentation: • OuluVS Dataset: This consists of short phrases like "hello," "excuse me," "I am sorry," etc. • 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	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.

Research Gap

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

Problem Statement

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.

Objective

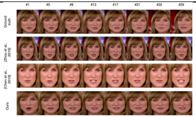
- Create a dataset of different accents.
- Develop an algorithm that can recognize words correctly independent of accents.

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

Generalized block diagram of visual speech recognition system

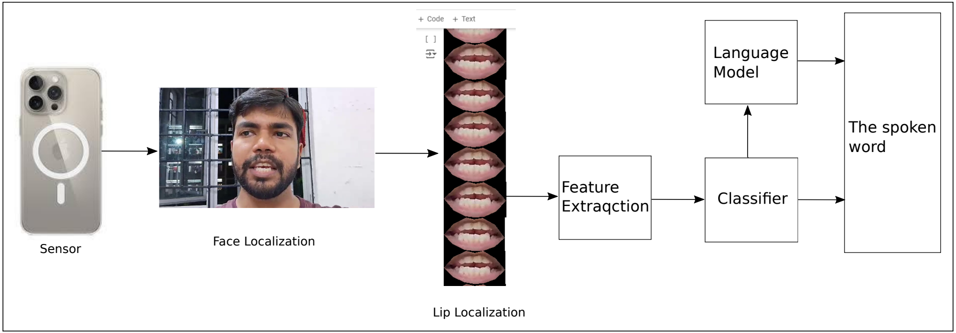


Figure 1: Generalized block diagram of visual speech recognition system

Futute Work

- Acquire videos of the same word with different accents.
- Develop algorithm for correct lip reading
- Apply our algorithm to our dataset as well publicly available dataset.
- 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.

- 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