Lip Reading System

A Project Based Learning Report Submitted in partial fulfilment of the requirements for the award of the degree

of

Bachelor of Technology

in The Department of Computer Science and Engineering

DEEP LEARNING (22AIP3305A)

Submitted by

2210030019: K. Harshita

2210030254: K. Bhavya Sri Sai

2210030012: C. Hasritha Reddy

Under the guidance of

Dr. G. Sai Sudha

Department of Electronics and Communication Engineering

Koneru Lakshmaiah Education Foundation, Aziz Nagar

Aziz Nagar – 500075

FEB - 2025.

Introduction

Lip reading systems have traditionally relied on conventional methodologies that involve manual or semi-automated feature extraction and classification processes. Early approaches utilized techniques like Principal Component Analysis (PCA), Hidden Markov Models (HMMs), and Support Vector Machines (SVMs) to manually extract and interpret features from the lip region. These methods often depended on high-resolution input data, typically captured at 640x480 pixels, to accurately identify and extract lip movements. How- ever, these conventional systems struggled with performance variability caused by changes in lighting, camera quality, and speaker diversity, making them unsuitable for real-world applications.

The absence of automation in conventional lip-reading systems created considerable challenges. Manual feature extraction requires extensive human effort and expertise, making the process slow, error-prone, and difficult to scale for larger datasets or real-time applications. Variations in speech speed, pronunciation, and facial expressions further complicated the accuracy of these systems, as they could not dynamically adapt to rapid or subtle changes in lip movements. Moreover, the limited diversity in conventional datasets, which often included controlled conditions with a small number of speakers, hindered the ability of these systems to generalize to new, unseen data.

To address these limitations, automation became crucial in advancing lip-reading technology. The integration of deep learning techniques, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), revolutionized the field by enabling end-to-end systems that could automatically learn relevant features from video data without manual intervention. CNNs are utilized to capture spatial features, such as lip shapes and movements, directly from video frames, while RNNs, specifically Long Short-Term Memory (LSTM) networks, handle the temporal dynamics of speech sequences. This combination allows automated systems to model the complex, sequential nature of speech, significantly improving accuracy and scalability. The shift towards automation not only enhances performance but also makes these systems more robust in diverse and challenging environments, paving the way for practical, real-time applications [1].

Literature Review/ Application Survey

Title	Methodology	Gaps
[1]	Twelve-layer CNN with two layers of Batch Normalization	Performance varies widely; struggles with unseen speakers and real-world conditions.
[3]	CNNs with RNN for temporal feature learning	High variability in real-world environments not addressed.
[4]	Self-attention and self-distillation with CNN front-end	Complexity due to model depth and high computational cost.
[5]	Spatial and temporal modeling using 3D CNNs	Difficulty in handling occlusions and distractions.
[6]	Seq2Seq model with CNNs and attention	Language-specific model, not tested for English.

TABLE I: Literature Review on few Lip-Reading Models

In this section, a comparison between different lip-reading models in accordance with methodologies and gaps/limitations is presented. Table I provides a structured comparison of various models, highlighting their methodological approaches and associated limitations. This comparative analysis focuses on key factors such as the dependency of models on controlled environments, challenges related to real-world variability, and specific areas that require further research and improvements.

Additionally, a comparison between various models based on their accuracy is presented in the bar graph (Figure 1). The primary objective of this analysis is to showcase the strengths and weaknesses of different lip-reading models, examining their effectiveness in diverse scenarios. A key emphasis is placed on understanding how well these models generalize beyond controlled environments, addressing aspects such as speaker diversity, lighting conditions, occlusions, and environmental noise. The structured overview serves as a foundation for understanding the current state of lip-reading technology and identifying opportunities for future development.

Linguistic analysis of lip-reading is both fascinating and conceptually rich. Many researchers have primarily relied on visual stimuli for speech recognition, focusing extensively on the analysis of lip movements. Over time, deep learning techniques have significantly influenced the development of lip-reading models. Early approaches leveraged traditional Hidden Markov Models (HMMs) combined with Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) classifiers. These methods, as evidenced by studies using datasets such as TCD-TIMIT, were primarily designed for word recognition rather than understanding sentence-level context. This limitation hindered their effectiveness, as these models lacked the ability to capture the structural and semantic relationships in continuous speech.

As research progressed, a shift towards more advanced architectures occurred. More recent approaches have favored speech sequence classification using Connectionist Temporal Classification (CTC) and Sequence-to-Sequence (Seq2Seq) models. These models improved upon previous methods by capturing greater temporal dependencies in speech, enhancing the ability to generate meaningful sentence-level predictions rather than just isolated words. This transition has allowed for more accurate and

contextually relevant lip-reading, making models more useful in real-world applications. Experiments have been conducted on a range of datasets, from basic pre-established datasets such as GRID, which provides simple audiovisual datasets, to more challenging and unconstrained datasets like Lip Reading in the Wild (LRW), which features diverse speakers and real-world conditions.

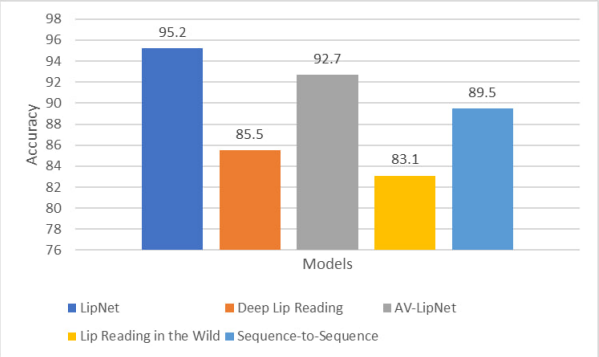
A. Performance Comparisons

Among the various models studied, Sequence-to-Sequence models appear to be the most effective for controlled sentence-level lip-reading tasks, making them a strong benchmark for further developments [4]. These models are particularly adept at handling structured speech but still face notable challenges when applied to real-world scenarios. One of the major limitations of all reviewed models is their difficulty in dealing with speaker variability, occlusions, and environmental noise. Real-world conditions introduce significant unpredictability that these models struggle to accommodate effectively. To address these issues, future research should focus on enhancing robustness by incorporating more sophisticated spatiotemporal feature extraction techniques.

One promising approach is the adoption of transformer-based architectures, which have demonstrated exceptional performance in capturing long-term dependencies in sequential data. Transformers, with their ability to model complex relationships between time steps, could significantly enhance lip-reading models' capacity to generalize across different contexts and speakers. Additionally, expanding the datasets to include more diverse speaker profiles, different lighting conditions, and real-world background noise will be crucial for improving model generalizability. The inclusion of such diverse datasets will help train models to perform reliably under varying conditions, making them more practical for real-world deployment.

Furthermore, integrating multimodal learning approaches that combine both audio and visual inputs is another key avenue for improving lip-reading accuracy. While purely visual models have made substantial progress, the addition of complementary audio information can provide more context and disambiguate challenging cases where lip movements alone may not be sufficient. The fusion of multiple modalities has the potential to create more robust and accurate lip-reading systems, enabling applications in fields such as assistive technology, security, and human-computer interaction.

These advancements will help overcome the current limitations faced by lip-reading models and push the boundaries of performance beyond what existing systems have achieved. By improving robustness to real-world variability, incorporating cutting-edge machine learning architectures, and leveraging multimodal learning techniques, future lip-reading models will be more adaptable and practical for widespread real-world applications.



References

[1] NadeemHashmi, Saquib & Gupta, Harsh & Mittal, Dhruv & Ku- mar, Kaushtubh & Nanda, Aparajita & Gupta, Sarishty. (2018). ”A Lip Reading Model Using CNN with Batch Normalization”. 1-6. 10.1109/IC3.2018.8530509.

[2] Sheng, Changchong & Kuang, Gangyao & Bai, Liang & Hou, Chenping & Guo, Yulan & Xu, Xin & Pietikainen, Matti & Liu, Li. (2024). Deep Learning for Visual Speech Analysis: A Survey. IEEE transactions on pattern analysis and machine intelligence. PP. 10.1109/TPAMI.2024.3376710.

[3] Fenghour, Souheil & Chen, Daqing & Guo, Kun & Li, Bo & Xiao, Perry. (2021). Deep Learning-Based Automated Lip-Reading: A Survey. IEEE Access. 9. 121184 - 121205. 10.1109/ACCESS.2021.3107946.

[4] Xue, Junxiao & Huang, Shibo & Song, Huawei & Shi, Lei. (2023). Fine-grained sequence-to-sequence lip reading based on self-attention and self-distillation. Frontiers of Computer Science. 17. 10.1007/s11704- 023-2230-x.

[5] Chung, Joon Son & Senior, Andrew & Vinyals, Oriol & Zisserman, Andrew. (2016). Lip Reading Sentences in the Wild. 10.48550/arXiv.1611.05358.

[6] Zhao, Ya & Xu, Rui & Song, Mingli. (2019). A Cascade Sequence-to-Sequence Model for Chinese Mandarin Lip Reading. 10.48550/arXiv.1908.04917.

[7] Patil, Rutuja. (2024). Enhancing Lip Reading: A Deep Learning Approach with CNN and RNN Integration. Journal of Electrical Systems. 20. 463-471. 10.52783/jes.1367.

[8] Chung, J.S., et al. ”Lip Reading Sentences in the Wild.” Proceedings of CVPR, 2017.

[9] Afouras, T., et al. ”LRS3-TED: A Large-Scale Dataset for Visual Speech Recognition.” Interspeech, 2018.

[10] http://www.robots.ox.ac.uk/ vgg/data/lip reading/lrs3.html

[11] https://sites.google.com/site/achrafbenhamadou/-datasets/miracl-vc1

[12] http://www.robots.ox.ac.uk/ vgg/data/lip reading/lrs2.html

[13] http://spandh.dcs.shef.ac.uk/gridcorpus/

[14] Chung, J.S., et al. ”Deep Learning Approaches to Visual Speech Recognition.” Speech Comm., 2020.

[15] Koller, O., et al. ”Quantitative Analysis of Lip Reading Accuracy in Noisy Environments.” IEEE Transactions on Audio, 2022.

[16] Fernandez-Lopez, A., et al. ”End-to-End Lipreading Models for Visual Speech Recognition.” ACM Multimedia Conf., 2019.

[17] Athanasios, F., et al. ”Speaker-independent Lipreading Accuracy in CNN Models.” Speech Processing, 2020.

[18] Momeni, H., et al. ”A Survey on Visual Speech Recognition (Lipreading) Systems.” IEEE Journal, 2020.

[19] https://www.robots.ox.ac.uk/ vgg/data/lip reading/lrw1.html

[20] https://www.kaggle.com/datasets/mohamedbentalb/lipreading-dataset