Study on Various Techniques of Video Lip Reading Using Deep Learning

Ramya Thavva[1], Ali Hasan[2], and Sadwale Shivaji[3]

[1,2,3]Department of AI & ML, Chaitanya Bharathi Institute of Technology, Gandipet, Hyderabad, India.

[1]ramyat_cse@cbit.ac.in, [2]ah14hasan@gmail.com, [3]shivajisadwale53@gmail.com

Abstract

The field of video lip reading has experienced a lot of development as a result of the innovations in deep learning and computer vision technologies. The ability to lip read from videos has proven to be a crucial advancement in several real life applications including defense, education, and in aiding the persons with hearing impairments. This study aims to provide an insight into the deep learning models and methodologies that have been increasingly used in the lip reading domain over the years. Another objective of the study is to provide an overview of each deep-learning based lip reading method by comparing their datasets, models, and performance on several evaluation metrics like accuracy, WER, CER, and others. This study also aims to highlight the challenges and research gaps to be bridged in the future.

Keywords: Lip reading, Deep Learning, Video Lip reading, Recurrent Neural Networks(RNNs) , Convolutional Neural Networks (CNNs), Computer Vision.

1 Introduction

Video lip reading refers to understanding what a subject is speaking in a video without listening to the audio and by just observing the movements of the subject's lips. This was generally performed by expert professionals until the last few years. However, progress in deep learning technologies and increase in large and diverse datasets has led to the advancement in automated lip reading systems. Over the past few years, Lip reading systems have progressed from simply recognizing small units of speech like individual characters to decoding complete sentences.

1.1 Methodology

The broad structure of lip-reading systems contains a frontend for extracting features, a backend for classification, and some pre-processing steps [3]. The various datasets used for implementation of lip reading using deep learning are described in Table 1. The automated lip-reading process consists of the following steps shown in Figure 1 [3].

- **Input videos:** Videos capturing different speakers are divided into individual frames that must be interpreted and decoded.

- **Region of Interest:** This pre-processing step involves identifying and extracting the Region of Interest (ROI). It includes the lips from the image data. It involves face detection, lip localization, and lip extraction from the frames. Basic computer vision manipulations like cropping are applied to the ROI to streamline the process of training and validation [14].

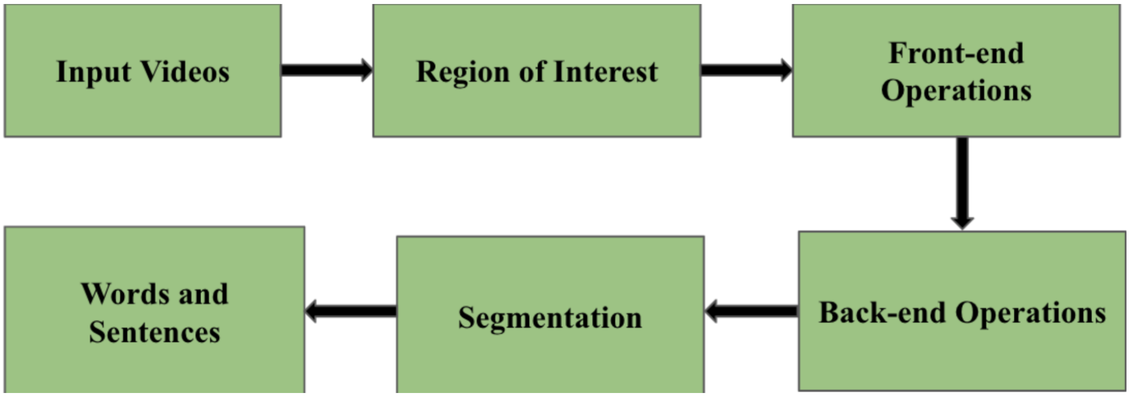

Figure 1: Block diagram of a typical Lip-Reading architecture.

Dataset Name	**Description**
Lip Reading in the Wild (LRW)	It contains 500 English words from BBC programs, with over 538,766 sequences.
BBC Lipreading Sentences 2 (LRS2)	It contains about 46,000 video tracks, 2 million word instances, and includes over 40,000 words from a variety of BBC broadcasts.
Lipreading Sentences 3 (LRS3)	It contains videos from more than 400 hours of TED and TEDx talks.
Japanese Audiovisual dataset	It comprises speech from six male speakers, covering 300 words - 216 phonetically balanced and 84 important words.
GRID corpus	It contains recordings of 1000 sentences spoken by 34 speakers (18 male, 16 female).
LRW-1000	It contains 1,000 Mandarin words across 718,018 videos by more than 2,000 speakers.

Table 1: Descriptions of the most common datasets used in lip reading architectures.

- **Front-End operations:** This step focuses on extracting pertinent features from redundant information and converting large-dimensional data into a more compact representation.

- **Back-end operations:** Speech is assigned to movements on the face that have been transformed to smaller-dimensional feature vectors.

- **Words and Sentences:** The interpreted speech is classified into segments or categories, eventually being converted into spoken words or sentences.

1.2 Evaluation Metrics

The following are the benchmark metrics through which the performance of a lip reading model is evaluated:

1. **Accuracy:** It measures how correctly the model recognizes the words in a given video. It is defined as the ratio of correct predictions (C) to the total number of words (T) multiplied by 100 [5]. It is given by the formula:

$$Accuracy = \frac{C}{T} \times 100 \tag{1}$$

2. **Word Error Rate (WER):** It is defined as the ratio of wrong predictions to the total number of words. WER ranges from 0 to 1. Lower WER values indicate better accuracy in recognizing the spoken words. WER is given by the formula:

$$WER = \frac{S + D + I}{N} \tag{2}$$

where, S is the number of substitutions, D is the number of deletions, I represents the number of insertions, and N is the total number of words [8].

3. **Character Error Rate (CER):** It is defined as the ratio of incorrectly predicted characters to the total number of characters. A lower CER indicates better performance. It is given by the formula:

$$CER = \frac{S + D + I}{C} \tag{3}$$

where, S is the number of substitutions, D is the number of deletions, I represents the number of insertions, and N is the total number of characters [8].

4. **BLEU score:** It compares the prediction of the model to one or more references by measuring the overlap between sequences of words generated by the model. A higher BLEU score indicates a closer match to the original reference.

5. **Overlap (OL):** It measures a lip segmentation model's accuracy. The overlap between the ground truth regions and the segmented regions of the lips is calculated. OL is given by the formula:

$$OL = 2 \times \frac{A1 \cap A2}{A1 + A2} \times 100 \tag{4}$$

where A1 is the ground truth lip region and A2 is the segmented lip region [9].

6. **Segmentation Error (SE):** It evaluates the errors in the outer lip boundaries and inner lip boundaries. It is given by the formula:

$$SE = 2 \times \frac{OLE \cap ILE}{2 \times TL} \times 100 \tag{5}$$

where OLE represents outer lip error, ILE refers to the inner lip error, and TL represents all the lip pixels in the region of ground truth [9].

2 Literature Survey

A visual speech recognition (VSR) system using a convolutional neural network (CNN) was developed for feature extraction in images of the mouth. It was implemented on the Japanese audiovisual dataset. Visual data was captured at 100 Hz, and images were cropped to 32x32 pixels. The audio data was recorded at 16 kHz. The CNN architecture had three convolutional layers. It also used response normalization and pooling for feature refinement. A fully connected layer for producing phoneme probabilities was used. The model was trained on over 100,000 frames. The goal was to maximize the value of the log-probability of correct phoneme labels [10].

LipNet, an advanced neural network model for complete sentence level lipreading, was proposed by combining spatio-temporal convolutional neural networks (STCNNs) and bidirectional gated recurrent units

(Bi-GRUs). It used connectionist temporal classification (CTC) loss for training. It featured three STCNN layers for extracting spatial and temporal features, along with Bi-GRUs that used temporal context from past and future frames. It was able to handle input and output of variable lengths without explicitly aligning the sequences. Training was done on the GRID corpus. Some dataset augmentation techniques like image mirroring, word clips, and frame alterations were done to prevent overfitting [1].

A novel method to predict spoken sentences from silent videos using lip movements from the LRS2 dataset was proposed. Lip regions were first extracted using facial landmarks. The frames were then converted to grayscale and resized to 112 x 112 pixels. Data augmentation techniques such as horizonatal flipping and random shifts were used. A spatial-temporal visual front-end was created to obtain features which consisted of a 3D convolutional network and 2D ResNet. Viseme sequences were produced by an attention transformer-based viseme classifier which used these features. Seventeen classes, that included visemes and sentence markers, were defined. The Carnegie Mellon Pronouncing Dictionary was then used to convert words into phonemes and map them to the visemes [4].

A DC-TCN (Densely Connected Temporal Convolutional Network) model was proposed to enhance lip reading of words that are isolated. LRW-1000 and LRW datasets were used for training and evaluation. It was better at capturing temporal features compared to traditional TCNs. 3D convolutional layers were used for spatial-temporal feature extraction. A 2D ResNet-18 was used for refinement. Spatial information was summarized with average pooling. It was then given as input to the model. Using identical dilation rates, the model explored Fully Dense (FD) and Partially Dense (PD) block variants. The model enhanced classification by using the Squeeze-and-Excitation (SE) blocks which emphasized informative features [7].

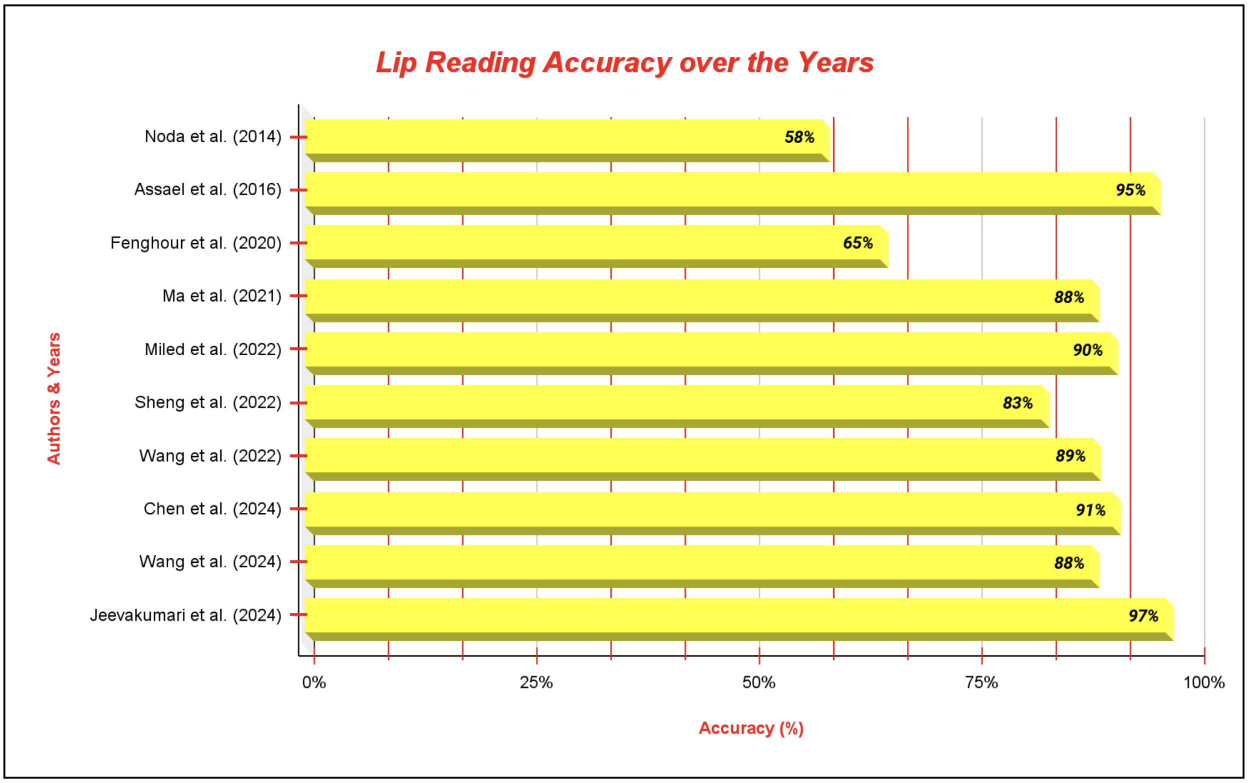

Figure 2: Bar graph illustrating the trends of accuracies from 2014 to 2024 in various lip reading architectures.

A two-block lip reading system was developed by combining lip segmentation and deep learning. The first block had a hybrid active contour model combining Geodesic Active Contours (GAC) with Distance Regularized Level Set Evolution (DRLSE) for precise mouth region segmentation in video frames. The second block used a spatio-temporal model with CNNs and Bi-GRUs to recognize spoken words. The CNN extracted features that were spatial in nature. The Bi-GRU captured the temporal dependencies. The Softmax layer performed classification. The model analyzed 22 frames per video in the LRW dataset to focus on key spoken word segments [9].

An Adaptive Semantic-Spatio-Temporal Graph Convolutional Network (ASST-GCN) aimed to enhance visual speech recognition by modeling dynamic mouth contours. The model had a global stream that captured overall lip motion and a local stream that focused on detailed mouth contour dynamics using Lip Reading related Landmark Points (LRLPs). The ASST-GCN module learned both semantic and spatio-temporal relationships among these points. The model utilized three primary datasets: LRW, LRS2 and LRS3 [11].

Another method using a 3D Convolutional Vision Transformer (3DCvT) was proposed, combining 3D convolution and vision transformers to extract features that are spatio-temporal in nature from videos. For sequence modeling, it had a frontend network with a 3D CNN, a transformer block for feature extraction, and a network in the backend using a Bi-directional Gated Recurrent Unit (BiGRU). The 3D CNN captured temporal correlations, and the transformer block enhanced global and local feature extraction. The model included a Squeeze-and-Excitation (SE) structure and used Mixup data augmentation and label smoothing for improving performance. The LRW-1000 and the LRW datasets were used for evaluation [13].

Three frameworks: viseme subword modeling (VSM), hybrid viseme subword and end-to-end modeling (HVSEM), and collaborative viseme subword and end-to-end modeling (CVSEM) were also proposed. VSM segmented words into sequences of viseme subwords and used deep neural networks and hidden Markov models to match each lip frame with its corresponding viseme. HVSEM incorporated both VSM and end-to-end approaches using multi-task learning, using their combined strengths especially in handling head movements. CVSEM enhanced this by including a state-mapped temporal mask to filter out noisy frames, permitting the end-to-end branch to focus on pivotal time steps. The LRW-1000 and LRW datasets were used for evaluation [2].

Another innovative approach called Mini-3DCvT, which integrated 3D convolution with visual transformers to effectively extract spatiotemporal features from videos, was proposed. It involved two main steps: In weight transformation, the parameters were shared among neighboring transformer layers and weight transformations were applied to enhance parameter diversity while maintaining performance. In weight distillation, three distillation techniques—prediction-logit, self-attention, and hidden-state— were used to transfer data from a large pre-trained model to a smaller model, thereby improving accuracy while reducing model size and computational complexity. The LRW-1000 and LRW datasets were used for evaluation [12].

Another approach for VSR in noisy environments was proposed using the GRID corpus. Pre-processing was done by extracting the region of interest (ROI) from videos, converting them to grayscale, normalizing them, and then scaling the frames to enhance model performance. The model used a 3D-CNN and EfficientNetB0 for feature extraction. For classification, the back-end part used Bidirectional Long Short-Term Memory (Bi-LSTM) layers with Connectionist Temporal Classification (CTC) loss. To train the model, a 0.0001 learning rate in the Adam optimizer was used, focusing on minimizing the CTC loss function [6].

Figure 2 illustrates the accuracies achieved by different lip reading approaches. Table 2 summarizes these approaches and highlights key aspects such as their strengths, limitations, and performance on corresponding evaluation metrics.

Authors	**Approach**	**Strengths**	**Limitations**	**Performance**
Noda, Kuniaki, et al. (2014)	Visual Speech Recognition (VSR) system with CNN for extracting features from mouth area images.	Application of deep learning with reasonable phoneme recognition and robustness against image variances.	Use of a small dataset affecting generalizability, reliance on speaker-dependent models, limited consonant recognition.	Achieved a phoneme recognition rate of 58% and word recognition rate of 37%.
Assael, Yannis M., et al. (2016)	Complete sentence level lip reading using STCNNs, Bi-GRUs, and CTC loss.	High accuracy on the GRID dataset, surpassing human lip reading performance and former models.	Use of a limited dataset, uncertain performance in uncontrolled, noisy environments.	The model achieved 6.4% CER and 11.4% WER for unseen speakers and 1.9% CER and 4.8% WER for overlapped speakers.
Fenghour, Souheil, et al. (2020)	Predicting spoken sentences from silent videos using 3D convolutional network and 2D ResNet.	Significant decrease in error rate and increase in accuracy from previous models, no need for audio input.	Significant gap between Viseme and Word error rates, indicating problems in converting viseme to words.	Decreased the WER to 35.4% compared to the past models.
Ma, Pingchuan, et al. (2021)	DC-TCN for lip reading of words that are isolated using 3D convolutional layers and 2D ResNet-18 for refinement.	High accuracy on LRW-1000 and LRW datasets, surpassing previous models.	Increased computational complexity, scope for improving accuracy on the LRW-1000 dataset.	Achieved 43.65% and 88.36% accuracies on the LRW-1000 and the LRW dataset respectively.
Miled, Malek, Mohammed Anouar Ben Messaoud, and Aicha Bouzid. (2022)	Lip segmentation and Lip reading using Haar Cascade classifier (for segmentation) and CNN and BiGRU (for lip reading).	Development of an accurate and robust lip reading system that uses lip segmentation and deep learning to achieve high accuracy on the LRW dataset.	Potential challenges in noisy environments and reliance on a specific dataset, affecting generalizability.	Achieved an accuracy of 90.38%, OL of 91%, and SE of 7.8%.

Authors	**Approach**	**Strengths**	**Limitations**	**Performance**
Sheng, Changchong, et al. (2022)	Adaptive Semantic-Spatio-Temporal Graph Convolutional Network (ASST-GCN), Transformers, and Convolutional Networks.	Enhancement of dynamic and automatic lip reading using novel frameworks, high experimental validation.	Over-reliance on precise facial landmark detection leads to challenges in complicated real-world scenarios.	Achieved an accuracy of 82.6% on the LRW dataset.
Wang, Huijuan, Gangqiang Pu, and Tingyu Chen. (2022)	3D Convolutional Vision Transformer and BiGRU (bidirectional gated recurrent unit).	Advancement of lip-reading technology using Transformers, effective feature extraction, comprehensive evaluation on LRW dataset.	Complexity of the model leading to high computational demands, and limiting real-world applicability.	Achieved 57.5% and 88.5% accuracies on the LRW-1000 and the LRW dataset respectively.
Chen, Hang, et al. (2024)	VSM (Hybrid and Collaborative) and End to End Modeling.	Integration of temporal mask module in CVSEM, filtering noise and improving model's decision making.	Computational complexity and reliance on large dataset limiting real-world applicability.	Achieved 58.89% and 90.75% accuracies on the LRW-1000 and the LRW dataset respectively.
Wang, Huijuan, et al. (2024)	3D Convolution and Visual Transformer techniques such as BiGRU.	Innovative use of weight sharing and distillation techniques, enhancing model efficiency without sacrificing performance.	Challenges in accurately capturing subtle lip movements and potential performance degradation during compressing of models.	Achieved 57.1% and 88.3% accuracies on the LRW-1000 and the LRW dataset respectively.
Jeevakumari, SA Amutha, and Koushik Dey. (2024)	3D Convolutional Neural Networks (3D CNN), Bi-Directional Long Short-Term Memory (BiLSTM), and Connectionist Temporal Classification (CTC) loss.	Comprehensive dataset utilization, successful integration of temporal and spatial data processing.	Reliance on a specific dataset affecting generalizability, need for larger datasets and alternate architectures to enhance real-world performance.	Achieved an accuracy of 96.7% and a WER of 8.2%.

Table 2: Overview of the strengths, limitations, and performance of various lip reading approaches by different authors over the years.

3 Conclusion

In conclusion, this study highlights the progress in video lip-reading systems because of the advancements in deep learning models. The study reviews various architectures and methodologies across datasets, revealing both strengths and limitations. The analysis shows models' accuracy and the importance of dataset traits in model selection.

Emerging trends in sophisticated architectures and hybrid paradigms have the promise of delivering improved accuracy but pose challenges like high computational demands and the need for diverse datasets. The study identifies gaps in adaptability and evaluation metrics, and tracks the trends in the accuracy of the models over the past decade. Continued research and collaboration are essential to advance video lip-reading capabilities through deep learning and overcome the current limitations.

References

[1] Assael, Y.M., Shillingford, B., Whiteson, S., De Freitas, N.: Lipnet: End-to-end sentence-level lipreading. arXiv preprint arXiv:1611.01599 (2016)

[2] Chen, H., Wang, Q., Du, J., Wan, G.S., Xiong, S.F., Yin, B.C., Pan, J., Lee, C.H.: Collaborative viseme subword and end-to-end modeling for word-level lip reading. IEEE Transactions on Multimedia (2024)

[3] Fenghour, S., Chen, D., Guo, K., Li, B., Xiao, P.: Deep learning-based automated lip-reading: A survey. IEEE Access **9**, 121184–121205 (2021)

[4] Fenghour, S., Chen, D., Guo, K., Xiao, P.: Lip reading sentences using deep learning with only visual cues. IEEE Access **8**, 215516–215530 (2020)

[5] Ferraro, A., Galli, A., La Gatta, V., Postiglione, M.: Benchmarking open source and paid services for speech to text: an analysis of quality and input variety. Frontiers in big Data **6**, 1210559 (2023)

[6] Jeevakumari, S.A., Dey, K.: Lipsyncnet: A novel deep learning approach for visual speech recognition in audio-challenged situations. IEEE Access (2024)

[7] Ma, P., Wang, Y., Shen, J., Petridis, S., Pantic, M.: Lip-reading with densely connected temporal convolutional networks. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2857–2866 (2021)

[8] Mehta, K., Qader, R., Labbé, C., Portet, F.: Fine-grained control of sentence segmentation and entity positioning in neural nlg. In: 1st Workshop on Discourse Structure in Neural NLG (2019)

[9] Miled, M., Messaoud, M.A.B., Bouzid, A.: Lip reading of words with lip segmentation and deep learning. Multimedia Tools and Applications **82**(1), 551–571 (2023)

[10] Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H.G., Ogata, T., et al.: Lipreading using convolutional neural network. In: Interspeech. vol. 1, p. 3 (2014)

[11] Sheng, C., Zhu, X., Xu, H., Pietikäinen, M., Liu, L.: Adaptive semantic-spatio-temporal graph convolutional network for lip reading. IEEE Transactions on Multimedia **24**, 3545–3557 (2021)

[12] Wang, H., Cui, B., Yuan, Q., Pu, G., Liu, X., Zhu, J.: Mini-3dcvt: a lightweight lip-reading method based on 3d convolution visual transformer. The Visual Computer pp. 1–13 (2024)

[13] Wang, H., Pu, G., Chen, T.: A lip reading method based on 3d convolutional vision transformer. IEEE Access **10**, 77205–77212 (2022)

[14] Zhang, J.X., Wan, G., Pan, J.: Is lip region-of-interest sufficient for lipreading? In: Proceedings of the 2022 International Conference on Multimodal Interaction. pp. 368–372 (2022)