CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY

An Autonomous Institute | Affiliated to Osmania University

Kokapet Village, Gandipet Mandal, Hyderabad, Telangana-500075, www.cbit.ac.in

Approved by | Affiliated to | UGC Autonomous | 10 Programs Accredited by | Grade A++ in | All India Ranking 151-200 Band

COMMITTED TO RESEARCH, INNOVATION AND EDUCATION 46 years

Department of Artificial Intelligence and Machine Learning

CBIT ICICT 2024 Conference

Presentation on:

“Study on Various Techniques of Video Lip Reading Using Deep Learning”

By: Paper ID - 227

Authors:

160121729020: Ali Hasan

160121729053: Sadwale Shivaji

Guided By:

Ms. Ramya Thavva, M. Tech (CSE)

Assistant Professor, Dept. of AIML

LIST OF CONTENTS

- Abstract
- Introduction
- Existing Works
  - Datasets Description
  - Evaluation Metrics
  - Literature Review Table
- Drawbacks of Existing Systems
- Proposed System
- Advantages
- Results
- Conclusion
- References

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.

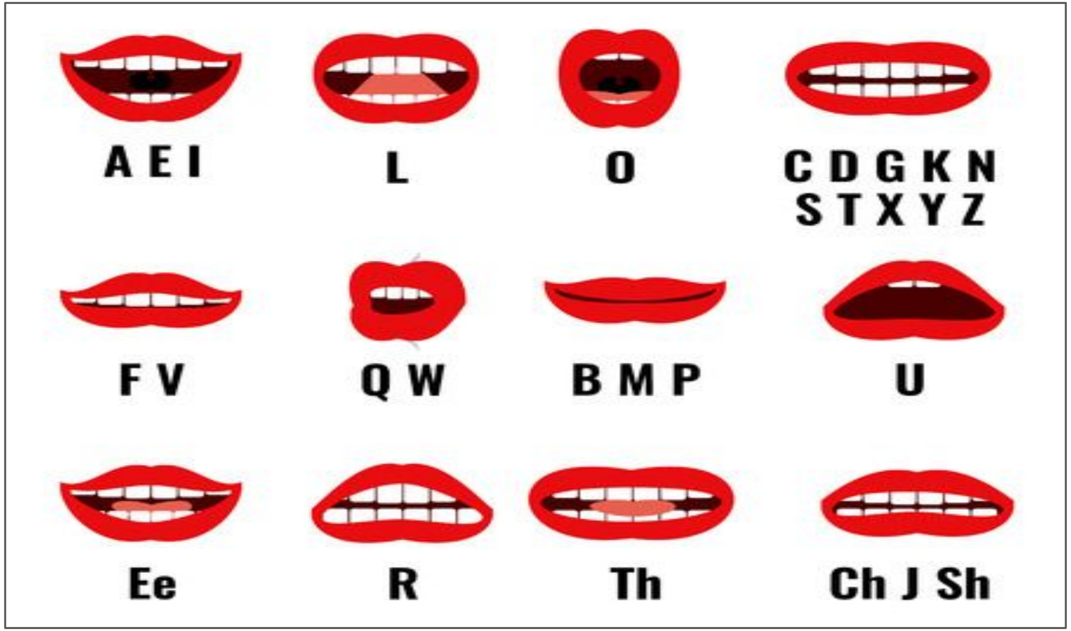
- This study aims to provide an insight into the deep learning models and methodologies that have been increasingly used in the lip reading domain over the years.

- 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.

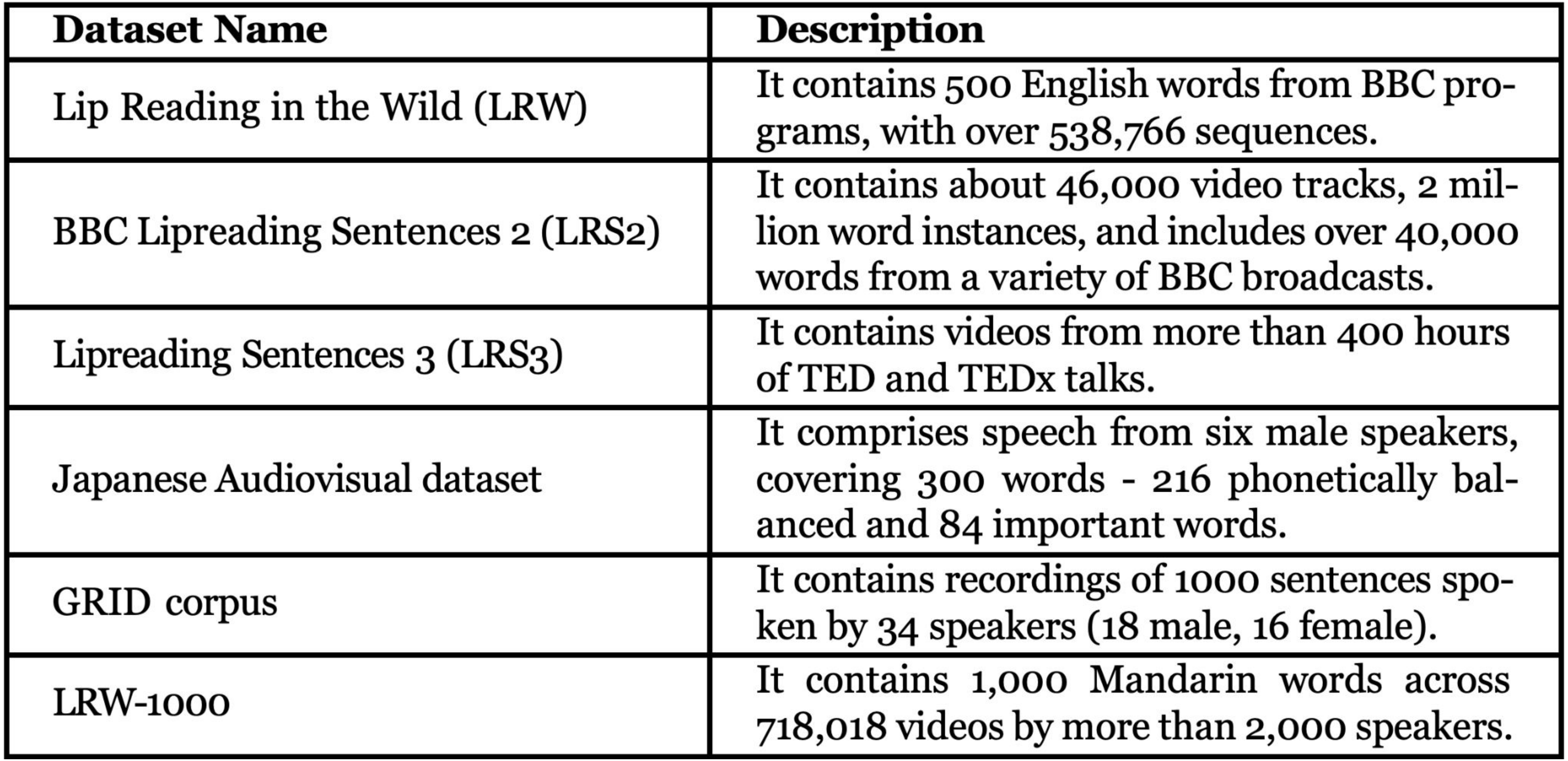
- This study also aims to highlight the challenges and research gaps to be bridged in the future and proposes a novel solution to overcome these challenges.

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.



Existing Works



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.

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

EVALUATION METRICS

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

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:

$$\textbf{Accuracy = (C/T) x 100}$$

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:

$$\textbf{WER = (S+D+I)/N}$$

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].

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:

$$\textbf{CER = (S+D+I)/C}$$

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].

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.

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:

$$\mathbf{OL = 2 \times [(A1 \cap A2)/(A1+A2)] \times 100}$$

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

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

$$\mathbf{SE = 2 \times [(OLE \cap ILE)/(2 \times TL)] \times 100}$$

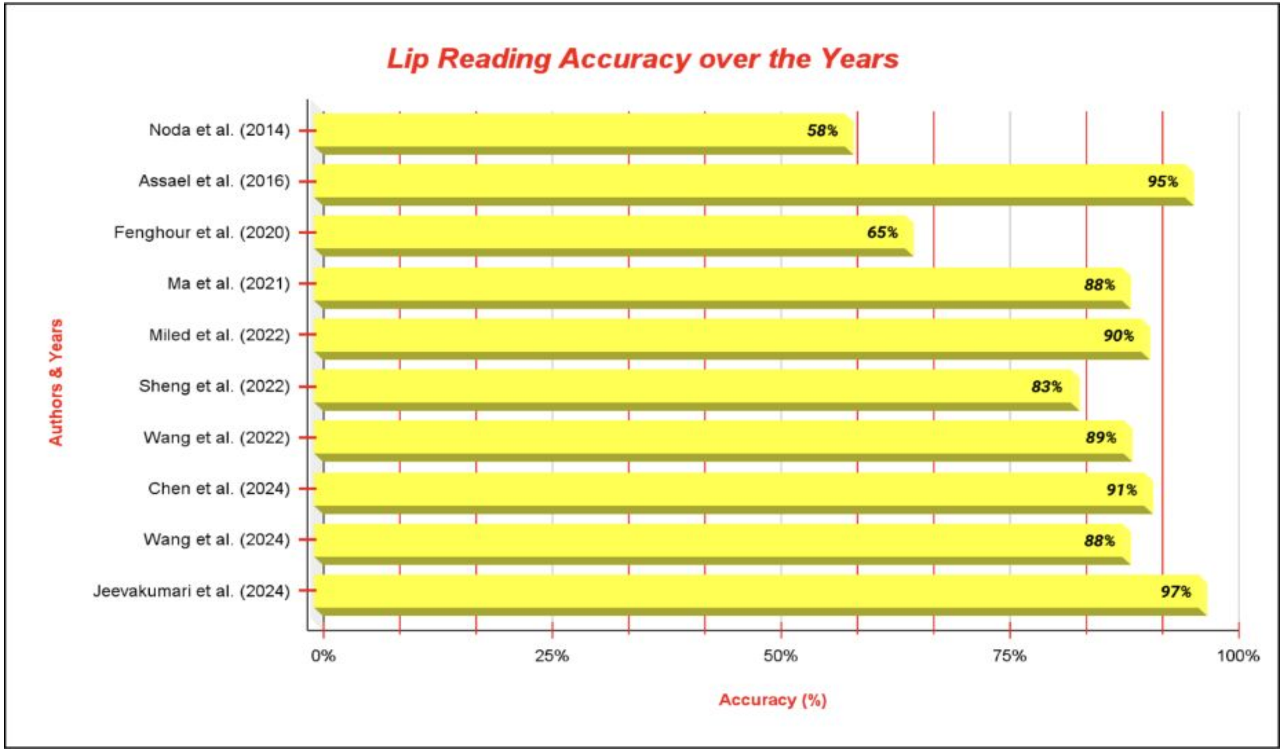
where OLE represents outer lip error, ILE refers to the inner lip error, and TL represents all the lip pixels in the region of ground truth [9].

Authors	Approach	Strengths	Limitations	Performance
Noda, Ku- niaki, 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.

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%.
Sheng, Chang- chong, 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, SAAmutha, 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 alternative architectures to enhance real-world performance.	Achieved an accuracy of 96.7% and a WER of 8.2%.

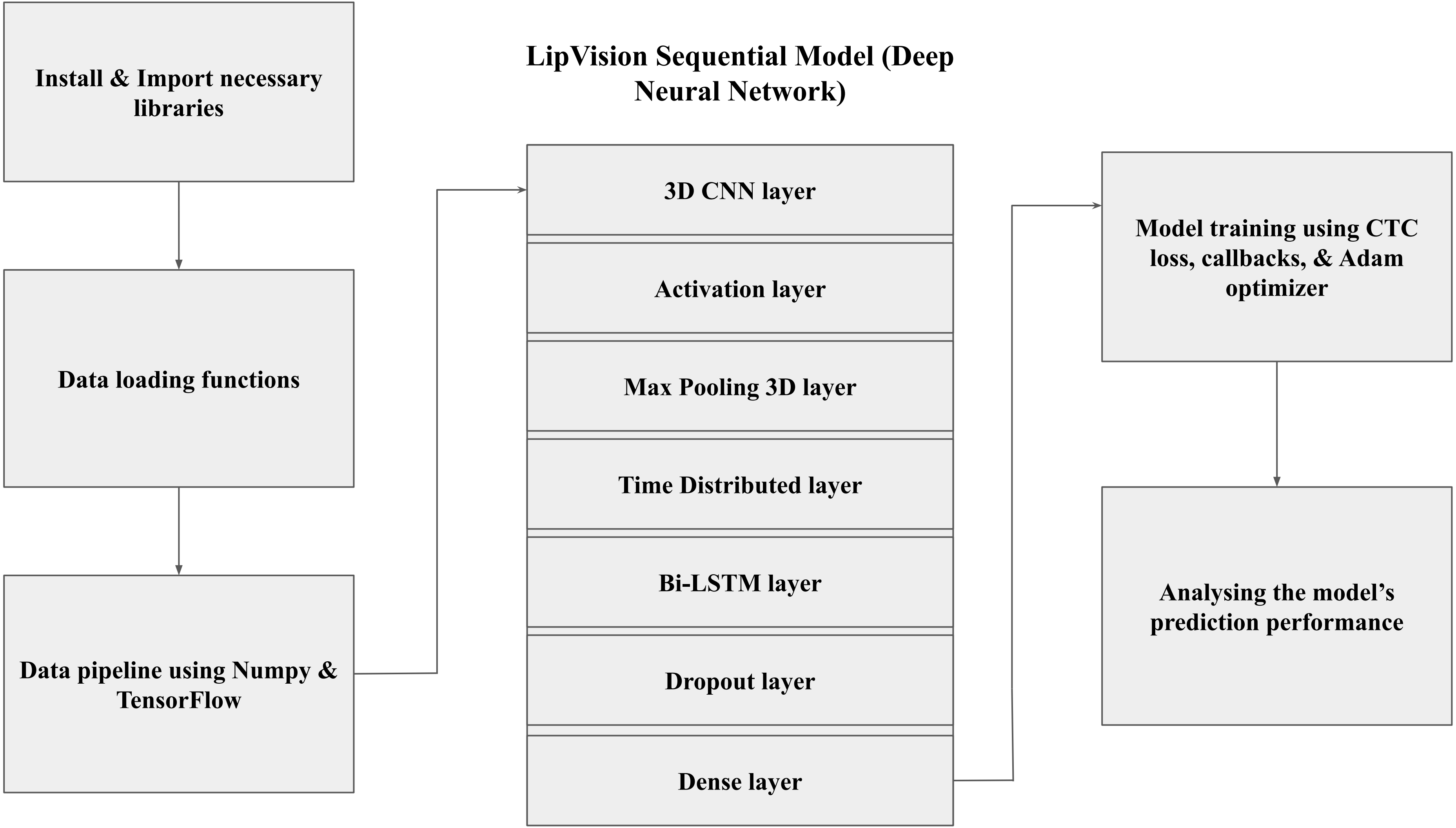


DRAWBACKS OF EXISTING SYSTEMS

The existing literature has either one or more of the following research gaps in the field of lip-reading:

- Lack of training on a sufficiently big and diverse dataset,
- Extension of word-level lip-reading to sentence-level lip-reading,
- Improving accuracy and robustness by using a more efficient and feasible model, and
- Development of practical applications and increasing accessibility of the technology for the general public.

Proposed System


Install & Import necessary libraries
Data loading functions
Data pipeline using Numpy & TensorFlow
LipVision Sequential Model (Deep Neural Network)
3D CNN layer
Activation layer
Max Pooling 3D layer
Time Distributed layer
Bi-LSTM layer
Dropout layer
Dense layer
Model training using CTC loss, callbacks, & Adam optimizer
Analysing the model's prediction performance

ADVANTAGES

This project overcomes the current research gaps in the field of lip-reading by:

1. Developing a better foundational deep learning model on the diverse GRID Audiovisual Sentence dataset using TensorFlow, 3DCNN, LSTM, and OpenCV.

2. Extending the current technology of word-level lip-reading to sentence-level lip-reading,

3. Increasing the robustness of the model by rigorously training the model to generate accurate predictions of the sentences spoken in the videos, and

4. Developing an interactive web application using Streamlit to increase the accessibility of the technology for the general public.

RESULTS

The following work has been implemented:

- Installation and importing of necessary libraries and dependencies.
- Building data loading functions to extract videos, frames, and alignments (transcriptions) from the dataset.
- Using Numpy and TensorFlow to create a data pipeline to feed the data into the model through a standardized array format.
- Designing the deep neural network of the model using TensorFlow’s sequential model having various deep learning layers such as 3DCNN, Activation, Max Pooling 3D, Time distributed, Bi-LSTM, Dropout, and Dense layers.
- Using the CTC loss function, callback functions and Adam optimizer to train the model on the alignments and videos of the dataset.
- Analysing the model’s performance in making accurate predictions about the sentences being spoken in the videos.

Jupyter Notebook PDF: https://drive.google.com/file/d/1BHmpN53dbl8F-e3fxEHM4bRgMFhzbTW1/view?usp=sharing

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.
- Finally, the study proposes a novel solution to overcome the research gaps, and showcases the model's proficiency in lip-reading tasks.

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)

Thank
you