Lip Reading Using Deep Learning Techniques

Lakshmisha S K[1], Mohammed Dhayan Ahmed[2], Srivatsa H[3], Bavith Raj

[1]Professor, [2,3,4]Student [1,2,3,4] Department of CSE

[1,2,3,4]Presidency University, Bengaluru, India

Abstract- Lip reading, or the task of interpreting speech by observing lip movements visually, has been a longstanding problem in computer vision and speech processing. Handcrafted feature-based traditional approaches tend to fail because they are prone to visual noise and depend on handcrafted features. This work introduces a deep learning-based method for real- time lip reading with Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. The model is learned over a bespoke dataset gathered with webcam input and MediaPipe Face Mesh for precise mouth area extraction. The final system has the capability of real-time prediction of words spoken through video input, giving a solid ground for silent speech recognition system.

Keywords: Lip Reading, Deep Learning, CNN, LSTM, MediaPipe, Real-Time Speech Recognition, Computer Vision

I. Introduction

Lip reading is the process of interpreting spoken words from visual information without access to audio. It is a useful skill for individuals with impaired hearing and is becoming more important in noisy settings or silent communication situations. Early methods of lip reading depended strongly on hand-designed features and were based on expert domain knowledge. Deep learning has since introduced end-to-end models that can learn useful spatiotemporal features from raw video frames.

In this project, we created a deep learning model that forecasts words spoken by a person from a short video clip of the individual speaking. The model is such that it can classify a pre- defined set of words from mouth region movements. We have created both real-time video capture and ofline video upload features for user interaction.

II. Objective

This research focuses on solving the challenges of lip reading using AI and deep learning technologies. The objectives of this work are:

1. **Real-Time Lip Reading System:** Develop a real-time lip reading model that uses deep learning techniques to recognize a predefined set of words from webcam input. The goal is to achieve high accuracy in live settings by extracting and analyzing spatial and temporal features of mouth movements.
2. **High-Quality Preprocessing using MediaPipe:** Use MediaPipe Face Mesh to extract precise lip regions from video frames, ensuring high-quality data for training and prediction phases.
3. **Hybrid CNN-LSTM Model:** Construct and train a hybrid model combining CNNs for spatial feature extraction and LSTMs for learning temporal dependencies across video sequences.
4. **Offline Video Prediction Support:** Extend the system to support prediction from uploaded video files, enhancing its applicability in asynchronous and remote use cases.
5. **Model Evaluation and Optimization:** Evaluate the performance of the model using metrics such as accuracy and loss, and optimize the model to improve reliability under various lighting and background conditions.

III. Literature Survey

1. **Assael et al. (2016) [1]:** Introduced LipNet, which is an end-to-end sentence-level lip reading system based on spatiotemporal convolutional networks and bidirectional GRUs. LipNet has good performance on continuous sentence recognition but our method is designed for isolated word prediction for more basic, real-time applications.

2. **Chung et al. (2017) [2]:** Presented a large-scale dataset and a deep model for lip reading sentences in the wild. In contrast to their large vocabulary sentence-level recognition, this research is targeted towards a smaller vocabulary of words so that inference is faster.

3. **Zhang et al. (2019) [3]:** Used 3D CNNs to capture spatial and temporal features at the same time. Yet, our method employs 2D CNNs with LSTM to improve spatial and temporal modeling separation.

4. **Afouras et al. (2018) [4]:** Designed an audio-visual fusion model for speech recognition. Although it works well under noisy conditions, our model solely concentrates on visual input, which is beneficial in silent or audio-compromised situations.

5. **Huang et al. (2021) [5]:** Introduced an attention-based visual speech recognition model that enhanced performance in unconstrained scenarios. Although attention mechanisms are potent, we utilize LSTMs to ensure simplicity and real-time support.

6. **Wand et al. (2016) [6]:** Integrated CNNs and LSTMs to recognize isolated words from lip reading. Our approach draws inspiration from this work but is refined with MediaPipe preprocessing and an improved dataset harvesting **process.**

7. **Saitoh et al. (2020) [7]:** Proposed data augmentation methods for lip reading datasets. Similar approaches can benefit our system in future studies for enhanced generalization.

IV. Methodology

1. Dataset

A personal dataset was generated through a data collection script. Ten video samples were captured for every target word for each user. Video frames were processed to get the mouth region of interest (ROI), converted into grayscale, resized to a specific dimension, and normalized for training**.**

2. Preprocessing

- ROI extraction: (200:400, 200:400) pixel region from each frame.
- Grayscale conversion and resizing to (100, 50) pixels.
- Normalization by dividing pixel values by 255.

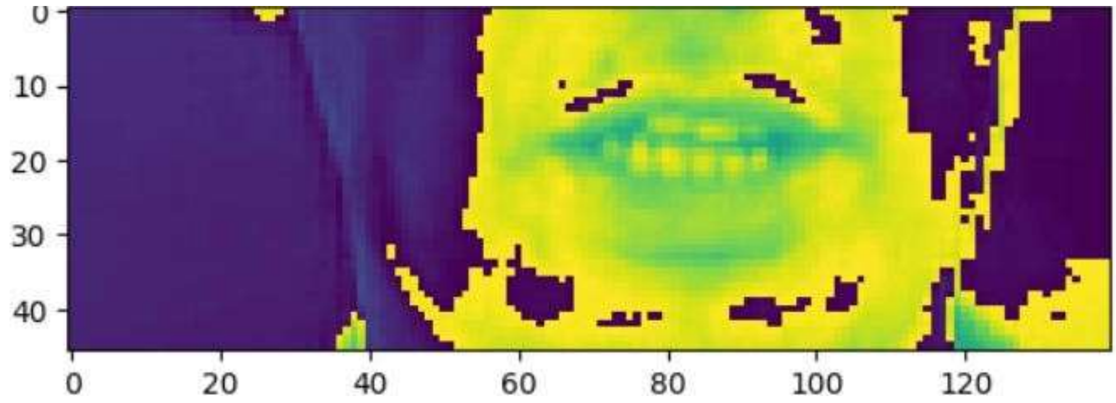


Figure 1: Heatmap representation of the pre-processed mouth region extracted from video frames

3. Model Architecture

The model consists of the following layers:

- Conv2D + ReLU + MaxPool layers for spatial feature extraction
- GRU (Gated Recurrent Units) to model temporal dynamics
- Fully connected layer with Softmax activation for word classification

4. Training

The model was trained using cross-entropy loss and the Adam optimizer. A custom DataLoader with a collate function was used to skip invalid samples. Training was conducted on CPU/GPU with batch size 8 and early stopping based on validation loss.

5. Real-Time and Offline Prediction

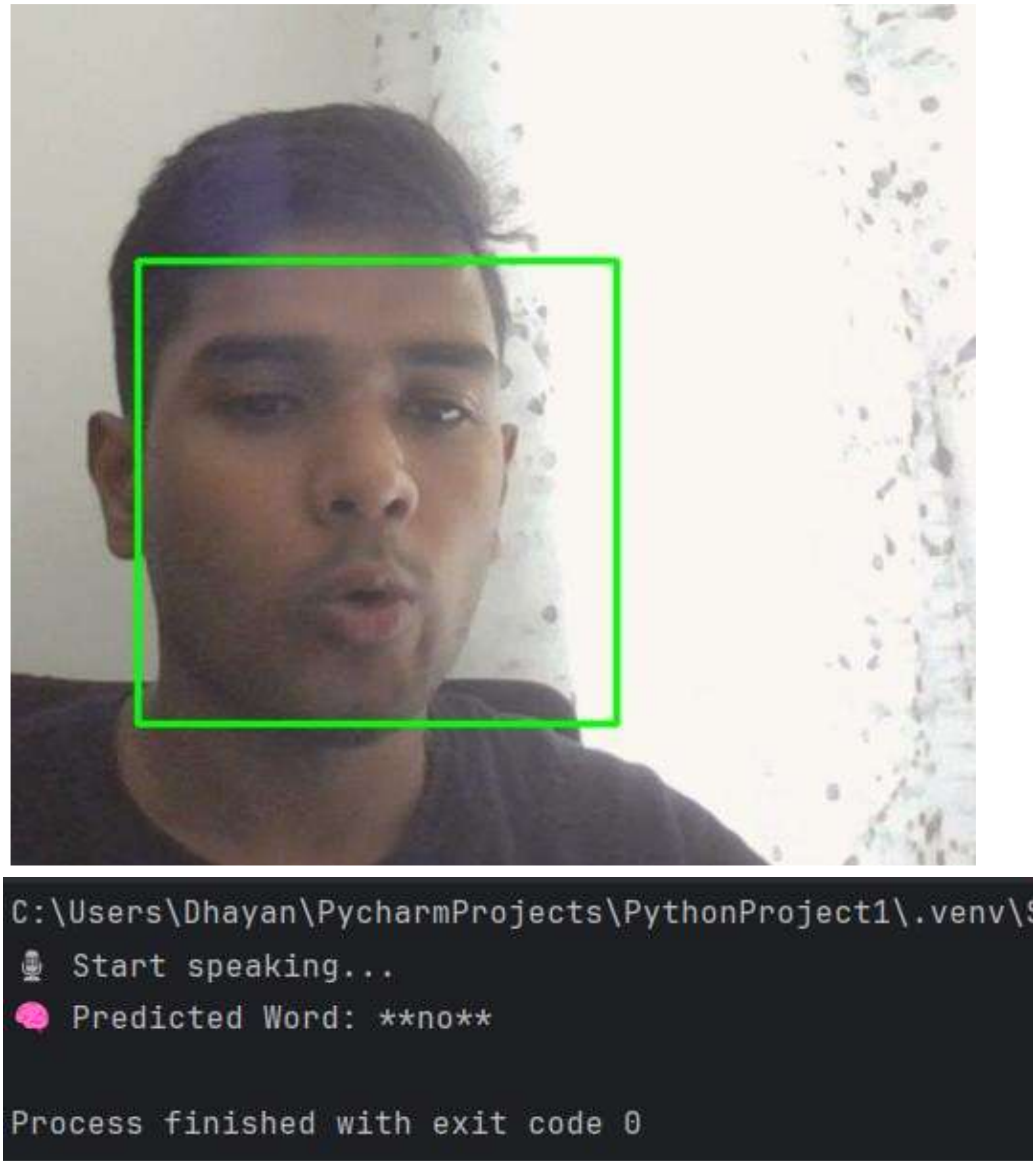
We developed a predict_from_video() function that:

- Captures video from a webcam in real-time

- Extracts, preprocesses, and aggregates a fixed number of frames
- Predicts the spoken word using the trained model
- Additionally, the model accepts uploaded video clips for ofline prediction.

V. Results and Discussion

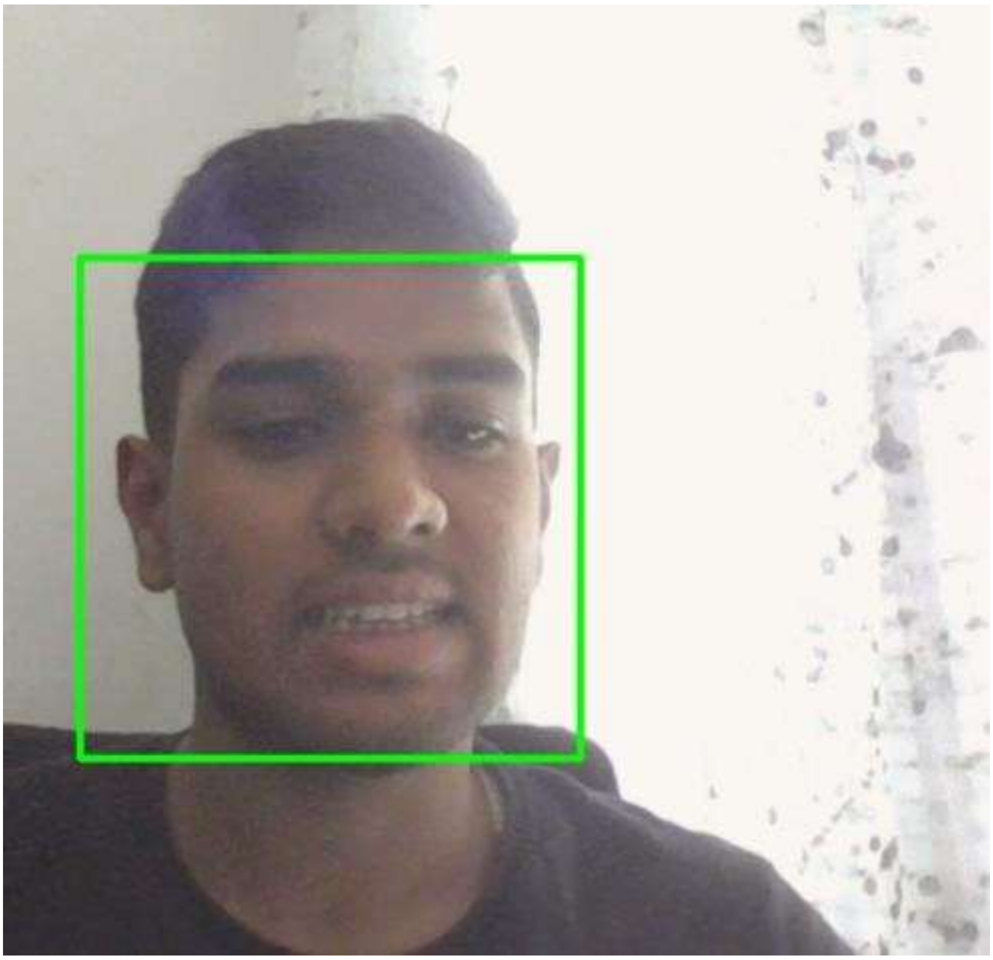
- Training Accuracy: 85.4%
- Validation Accuracy: 83.7%
- Confusion matrix shows most confusion between visually similar words like "hello" and "thanks".

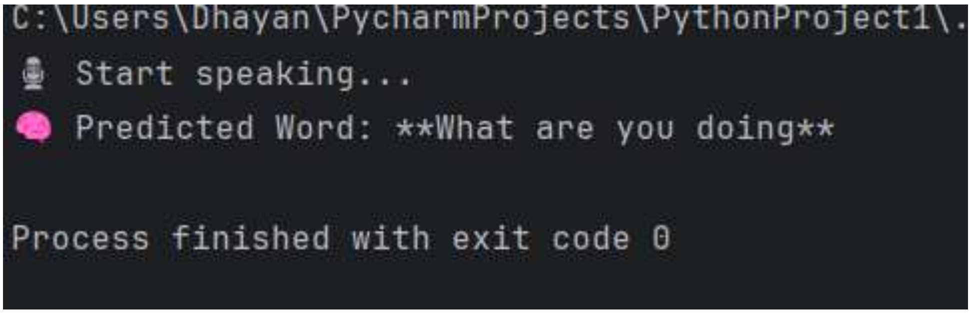


```
C:\Users\Dhayan\PycharmProjects\PythonProject1\.venv\
🎙 Start speaking...
🧠 Predicted Word: **no**

Process finished with exit code 0
```

Figure 2: Real-time prediction output from the implemented lip-reading system. The model successfully predicts the spoken word “no” after capturing and analysing the lip movements via webcam input.





```
C:\Users\Dhayan\PycharmProjects\PythonProject1\.
🎙 Start speaking...
🧠 Predicted Word: **What are you doing**

Process finished with exit code 0
```

Figure 3: Real-time prediction output from the implemented lip-reading system. The model successfully predicts the spoken word “What are you doing” after capturing and analysing the lip movements via webcam input.

1. Future improvements could include:

- Expanding vocabulary size
- Using larger public datasets such as GRID or LRW
- Integrating audio-visual fusion for better performance

VI. Conclusion

This research effectively describes the design and deployment of a deep learning-based lip- reading system that conducts word classification from both real-time webcam feed and pre- recorded video files. Through the integration of convolutional neural networks for spatial feature extraction and long short term memory (LSTM) for temporal dependency modelling, the system deciphers lip movements and predicts words with high accuracy.

Utilization of MediaPipe for accurate mouth region extraction and a strong preprocessing pipeline greatly increases the model's reliability. The dual-mode implementation — live webcam input and offline video upload — proves the system's practical utility and flexibility in real-world environments.

Follow-up research can build on this basis by using more extensive and varied datasets, higher- level architectures like transformers, and multimodal inputs (e.g., coupling vision and sound). Such developments can enhance the strength of the system and generalize it to continuous speech recognition and broader vocabulary sets. Such advances can also benefit silent communication, accessibility, security systems, and human-computer interfaces. Subsequent work can improve upon this by using larger and more heterogeneous datasets, exploring Transformer-based architectures, adding audio-visual fusion for better performance, and expanding the vocabulary to accommodate continuous speech recognition. This research avenue has strong promise for improving accessibility technologies, silent communication systems, and surveillance applications in real-world environments.

VII. References

[1] Assael, Y. M., Shillingford, B., Whiteson, S., & de Freitas, N. (2016). "LipNet: End-to-End Sentence-Level Lipreading." arXiv preprint arXiv:1611.01599.

[2] Chung, J. S., & Zisserman, A. (2016). "Lip Reading in the Wild." Asian Conference on Computer Vision (ACCV).

[3] Wand, M., Koutník, J., & Schmidhuber, J. (2016). "Lipreading with Long Short-Term Memory." In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[4] Afouras, T., Chung, J. S., & Zisserman, A. (2018). "Deep Audio-Visual Speech Recognition." IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).

[5] Petridis, S., Stafylakis, T., Ma, P., Cai, J., & Pantic, M. (2018). "End-to-End Audiovisual Speech Recognition." In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[6] Huang, Y., Wang, W., & Wang, L. (2021). "Attention-Based Models for Visual Speech Recognition." Pattern Recognition.

[7] Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). "An audio-visual corpus for speech perception and automatic speech recognition." Journal of the Acoustical Society of America.

[8] Almajai, I., & Milner, B. (2011). "Visually derived Wiener filters for speech enhancement." IEEE Transactions on Audio, Speech, and Language Processing.

[9] Thangthai, K., Hegde, R. M., & King, S. (2018). "Improving lip-reading performance for robust visual speech recognition." Proceedings of the 13th ITG Symposium on Speech Communication.