Neural Network Approach for Visual Speech Recognition

Navya Kamepalli, Sree Surya Sreemanth Yedidhi, Vivek Chandra, Yagna Praseeda Atmuri

Introduction:

This paper presents a Visual Speech Recognition (VSR) system that directly converts video frames of mouth movements into text for sentence-level predictions. By utilizing neural networks (CNNs), recurrent neural networks (RNNs), and connectionist temporal classification loss, the system achieves remarkable accuracy on sentence-level tasks in the GRID corpus, surpassing human lip readers and previous methods. Automated lip reading, merging computer vision and natural language processing, holds promise in enhancing communication accessibility and security. This project aims to develop an advanced lipreading system capable of transcribing entire sentences from videos of lip movements, benefiting individuals with hearing impairments and improving speech recognition in noisy environments. Despite challenges like speech variations and background noise, automated lip reading presents an exciting frontier for innovation, leveraging deep learning techniques to decipher spoken content from visual cues and adapt to diverse speakers, languages, and speech styles.

Beyond technical advancements, this endeavor has profound societal impacts, empowering those with hearing challenges to participate more fully in conversations. Furthermore, in security domains like surveillance or forensic analysis, accurate lipreading can offer valuable insights and complement existing audio-based systems.

Problem statement:

The primary objective of this project is to develop an automated lipreading system that can transcribe full sentences directly from video input of a speaker's mouth movements. Lipreading is a complex task due to its requirement to capture spatiotemporal features from video frames. While recent deep learning-based models have improved visual feature extraction, they mostly perform word classification without addressing sentence-level sequence prediction. The ambiguity in lipreading also poses challenges due to similar mouth movements for different phonemes, making context essential for accurate predictions. The need for an end-to-end model capable of extracting spatiotemporal features and making accurate sentence-level predictions across various speakers is crucial for advancing lipreading applications.

Data Description:

The performance of this Visual Speech Recognition (VSR) system is assessed using the GRID corpus, which includes both audio and video recordings from 34 speakers, each generating 1000 sentences for a total of 34,000 sentences. The corpus follows a structured grammar pattern involving **command + color + preposition + letter + digit + adverb**, with multiple word choices for each category. This diversity results in a vast array of 64,000 potential sentence combinations, making it a robust benchmark for evaluating lipreading models. The primary focus

of the corpus is on sentence-level recognition tasks, and the VSR system's capability to predict entire sentences allows it to leverage the temporal context provided by this dataset. Consequently, it achieves higher accuracy compared to models that concentrate solely on individual word recognition.

Methodology

The methodology for developing the visual speech recognition system involves several key steps.

1. Data Acquisition and Preprocessing:

The initial phase focuses on obtaining a diverse dataset for training and evaluation. This includes acquiring audio and video recordings of speakers demonstrating a range of sentences as per the structured grammar defined by the GRID corpus. Preprocessing steps involve aligning audio with corresponding video frames, extracting relevant mouth regions, and augmenting the dataset to enhance model generalization.

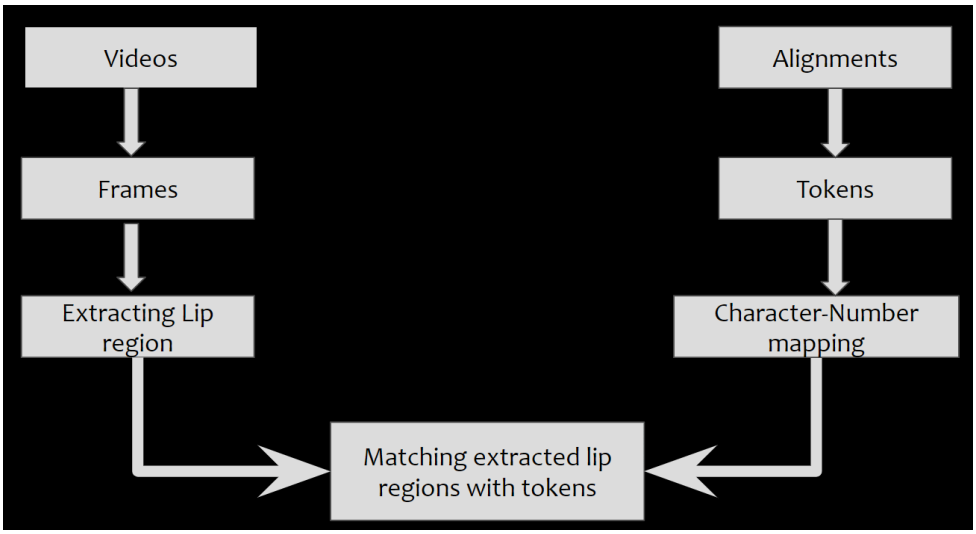

Figure 1: Workflow of Data Preprocessing

2. Feature Extraction:

Feature extraction is critical for capturing spatiotemporal patterns from video frames. Techniques such as facial landmark detection, and 3D convolutional representations are explored to extract meaningful visual features that encode lip articulations and dynamics over time. Facial landmark detection is used to localize and extract the lip region from each video frame, focusing the analysis on the most relevant area for lip reading. The extracted lip regions are then processed using 3D convolutional neural networks (CNNs), which can effectively capture both spatial and

temporal information. The 3D CNNs apply convolutional operations across both the spatial dimensions (height and width) and the temporal dimension (time) of the video frames, allowing the model to learn discriminative features that represent the lip movements and their evolution over time.

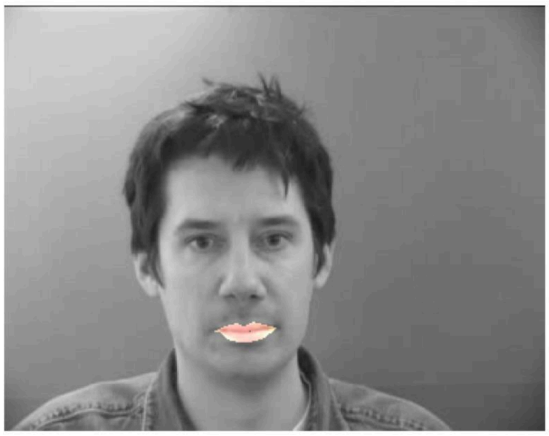

Figure 3.1: Processed frame

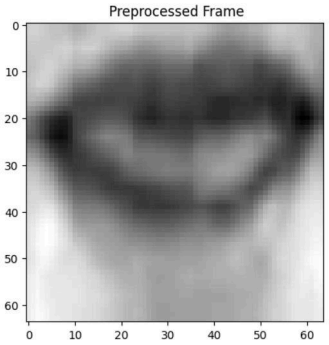

Figure 3.2: Lip region extracted from frame

3. Model Architecture Design:

The core of the project lies in designing a sophisticated neural network architecture capable of handling spatiotemporal data for visual speech recognition. This involves a hybrid approach combining Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). CNNs extract spatial features and RNNs capture temporal dependencies that may handle long-range contextual information.

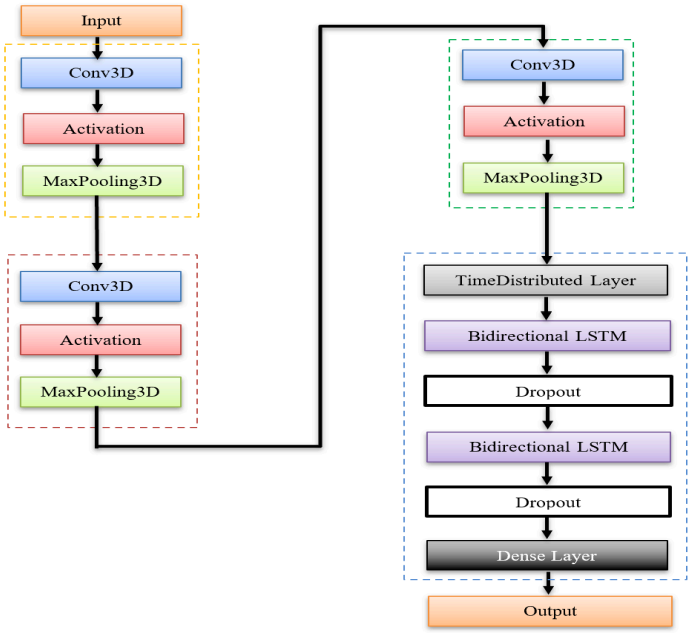

Figure 3: Architecture of the visual speech recognition system

4. Training and Optimization:

The model is trained using the Adam optimizer with a learning rate of 0.0001. The learning rate is adjusted using a learning rate scheduler, which reduces the learning rate exponentially by a factor of 0.1 after 30 epochs. This helps the model converge to a better solution by gradually decreasing the learning rate as training progresses.

The model is trained for a total of 60 epochs. Initially, the model is trained for 20 epochs, and the weights are saved using the ModelCheckpoint callback. The saved model is then loaded, and training is resumed for an additional 20 epochs. This process is repeated one more time, resulting in a total of 60 epochs of training.

During training, the CTCLoss function is used as the loss function. CTC Loss is specifically designed for sequence prediction tasks with variable-length inputs and outputs, such as speech recognition and lip reading. It efficiently handles the alignment between the input sequences and the target sequences.

5. Evaluation Metrics:

The primary evaluation metric used is accuracy. Accuracy measures the percentage of correctly predicted characters by comparing the predicted text with the original text. Accuracy is calculated by dividing the total number of correctly predicted characters by the total number of characters in the original texts. This metric provides a quantitative assessment of the model's performance in accurately transcribing lip movements into text.

Results:
The overall accuracy of the system is reported as 57.36%. This metric indicates the proportion of correctly transcribed sentences compared to the total number of sentences evaluated.

```
Original Text: set blue with o one soon
Predicted Text: set blue in o one soon

Original Text: set blue with o two please
Predicted Text: set blue ith f twor please

Original Text: set blue with o three again
Predicted Text: set blue in f thre again

Original Text: set blue with o zero now
Predicted Text: set blue with u tewo now

Original Text: set blue with u four now
Predicted Text: set blue it t four now

Original Text: set blue with u five soon
Predicted Text: set blue in five soon

Original Text: set blue with u six please
Predicted Text: set blue with t thr please

Original Text: set blue with u seven again
Predicted Text: set blue in z seven again
```

The attached Colab notebook file contains complete code implementation, detailed explanations, and results of our lip reading project Link

Unsuccessful Architectural Attempts:

1. We attempted to implement a complex architecture that incorporated multiple convolutional layers, dense layers, transition layers, and parallel processing layers. However, during the execution of this architecture, we encountered an error where the tensor size exceeded the available GPU memory. The Colab notebook file containing the code for this architecture can be accessed through the following Link.
2. **Another architecture Link**

Appendix:

I. Navya was responsible for conceptualizing the project idea and conducting a thorough literature review. This involved identifying the significance of automated lip reading, exploring existing research and advancements in the field, and formulating the project's overarching goals and objectives based on the insights gained from the literature review.

II. Surya played a key role in collecting and preprocessing the data required for the project. This included sourcing audio and video recordings from the GRID corpus, aligning audio with video frames, extracting relevant mouth regions, and augmenting the dataset to enhance model generalization. Surya also ensured the quality and integrity of the data throughout the preprocessing phase.

III. Vivek took charge of designing the model architecture for the project. This involved selecting appropriate deep learning techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Vivek's role also included fine-tuning

hyperparameters, optimizing the model's performance, and ensuring its scalability and adaptability to diverse speakers, languages, and speech styles.

IV. Praseeda was responsible for documenting the entire project process and writing the final project report. This included detailing the methodology, experimental setup, results analysis, and conclusions drawn from the project. Praseeda ensured that the project documentation was comprehensive, well-structured, and effectively communicated the project's objectives, methodologies, and outcomes.

References:

1. Davis, S. B., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4), 357-366.

2. Summerfield, Q. (1979). Use of visual information for phonetic perception. Phonetica, 36(4-5), 314-331.

3. Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. Science, 313(5786), 504-507.

4. Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 18(5-6), 602-610.

5. Chung, J., et al. (2017). Lip reading sentences in the wild. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3444-3453.

6. Wand, M., et al. (2020). A comprehensive study of deep lip reading: Models, benchmarks, and resources. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(8), 1827-1843.

7. Chandrasekaran, C., et al. (2009). The GRID corpus: A multimodal corpus for lipreading, speech recognition, and person identification. Image and Vision Computing, 27(3), 345-352.

8. Chung, J. S., et al. (2016). Lip reading in the wild using LSTM networks. In Advances in Neural Information Processing Systems (NeurIPS), 2853-2861.

9. Koller, O., & Ney, H. (2015). Continuous speech recognition by phoneme-based models with joint-sequence training. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(3), 429-441.

10. Potamianos, A., et al. (2003). Recent advances in the automatic recognition of audiovisual speech. Proceedings of the IEEE, 91(9), 1306-1326.