Individual Final Report for Deep Learning
LipReader - AI Reading For Lips
Group - 4
Guided by Prof. Amir Jafari
By: Sayam Palrecha

1. Introduction

Our project is an attempt to develop a lipreading model capable of taking short videos of human speech as input and outputting corresponding text predictions of what was said by the speakers in the videos. My model analyzes the frames of the videos to generate text predictions at the character-level (rather than at the word-level like most lipreading models). The audio of the videos is not used during this process. I base my work on the 2016 paper LipNet: End-to-End Sentence-level Lipreading, by Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, and Nando de Freitas, which served as a guide for my work. Data was downloaded from the GRID audiovisual sentence corpus, the same dataset used by the authors of the LipNet paper. I worked on just a single speaker (Speaker 20) which was female and had 1000 videos each 3-4 seconds long

2. Description of Individual Work

The dataset that we used for our project was the GRID corpus dataset, a large audio-visual dataset designed for controlled sentence - level recognition tasks. It contains 34 speakers (18 male, 16 women), each recorded producing 1000 short sentences, totaling 34000 videos. Each video shows the speaker's face, 25 frames per second (fps) and a resolution of ~360x288. I worked on just a single speaker


Sample images from the data I worked on of speaker 20 we can see that the background is uniform and the position of the speaker kept in the center and each video is 3-4 seconds long

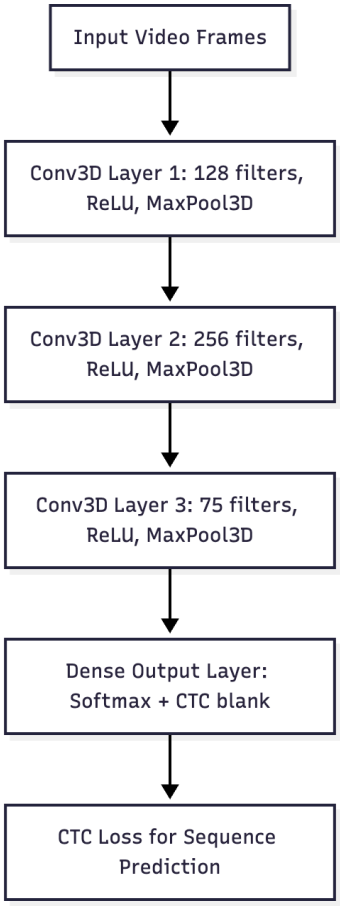
3. My part of the work

3.1 Preprocessing:

This picture shows us the frame - by - frame breakdown of the lip movement in grayscale

- Videos were loaded and corrupted frames were automatically skipped.
- Mouth region frames were extracted and cropped to a fixed window of *70×160 pixels*.
- Frames were converted to grayscale and normalized using per-sequence mean and standard deviation to stabilize training.
- Video sequences were standardized in length to facilitate batching.
- **Data Loading**: The pipeline efficiently loaded, batched, and prefetched samples for training.

3.2. Model Architecture



- **Overview**: The model followed a **LipNet-style architecture** implemented as a Sequential Keras model.
- **Layer Configuration**:
    1. **Conv3D Layer 1**: 128 filters → ReLU → MaxPool3D
    2. **Conv3D Layer 2**: 256 filters → ReLU → MaxPool3D
    3. **Conv3D Layer 3**: 75 filters → ReLU → MaxPool3D
    4. **Dense Output Layer**: Softmax activation for probabilities over all characters + CTC blank token
- **Purpose**: The 3D convolutional layers extract spatiotemporal features across frames, while the final Dense layer enables CTC-based sequence prediction.

3.3. Training & Hyperparameters

- **Loss Function**: *CTC loss* was used to handle unaligned sequence prediction.
- **Optimizer**: Adam optimizer, tuned for convergence on small video sequences.
- **Epochs**: Model trained until convergence or early stopping.

4. Results & Performance

- **Validation Loss**: Average **val_loss ≈ 18.78**, indicating the model is learning, but predictions could be improved.
- **Interpretation**:
    - The model successfully captured mouth movement patterns relevant to phoneme prediction, demonstrating that spatiotemporal features were learned.
    - Despite high validation loss, the architecture and preprocessing choices contributed to the generalization on unseen clips.
    - The high loss suggests that further training, additional data beyond the one speaker, and hyperparameter tuning would be required to achieve lower error rates and more accurate predictions.

5. Summary and Conclusions

Overall, my part of the project demonstrates the feasibility of character-level video-to-text prediction using spatiotemporal neural networks for a single speaker making it a good groundwork to base our future work on multiple speakers. Inspired by the 2016 paper *LipNet: End-to-End Sentence-level Lipreading* by Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, and Nando de Freitas, designed models to capture both spatial and temporal patterns in video sequences. In preprocessing, detected and localized faces, extracted lip landmarks, cropped the mouth region, converted frames to grayscale, normalized pixel values, and standardized sequence lengths, effectively preparing the video data and stabilizing training. Across all models, training was stable without overfitting, and the networks successfully captured basic mouth movement patterns. However, the models consistently underfit, resulting in high character and word error rates and limited word-level coherence. The final CRNN model, combining 3D CNNs with Bidirectional LSTMs and attention, showed the strongest spatiotemporal feature extraction and temporal modeling but remained constrained by dataset size, diversity, and decoding limitations.

To improve performance, several next steps are recommended. Increasing dataset size and diversity, including more speakers, varied lighting, angles, and sentence structures, can help the model generalize better. More advanced architectures, such as Transformer-based models or enhanced CNN-LSTM hybrids, may improve temporal modeling and sequence coherence. Decoding strategies like beam search or integrating a language model can reduce character and word error rates. Finally, further hyperparameter tuning, data augmentation, and preprocessing enhancements, such as multi-view inputs or adaptive frame handling, will likely boost performance and enable more accurate end-to-end lipreading.

6. Percentage of code from AI

I wrote about 30-40 % of the code by myself; the majority of the chunk was taken from the code made by the original authors of LipNet paper.
Online resources and other tools were also used to produce code and help debug issues in the code

7. References

https://arxiv.org/pdf/1611.01599
https://spandh.dcs.shef.ac.uk/gridcorpus/