UNIVERSITY of WASHINGTON

# Lip-Reading Using Transformer

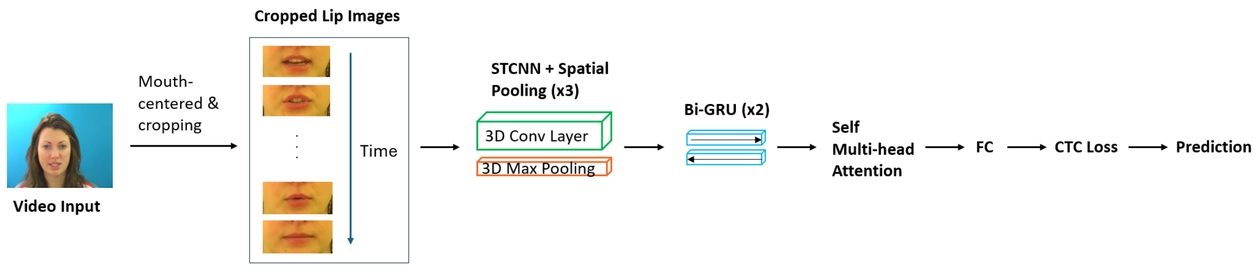
Audrey Sein Aye, Diana Kim, Jaehyeon Ahn

## Motivation

- Lip-reading has emerged as a significant field in humancomputer interaction and artificial intelligence, potentially improving accessibility in multimedia content and communication systems.
- We will develop an advanced LipNetWithAttention architecture including self-attention mechanisms in existing LipNet architectures based on CNN and Multihead Attention Layer.

## Architectures

- Figure 1. LipNetWithAttention architecture.




1. The model receives video data from the GRID Corpus as input.

2. The video is labeled with cropped lip images and annotations through mouth-centered and cropping.

3. The video's spatial and temporal features are extracted using 3D convolutional layers + 3D max pooling (x3).

4. The extracted features are processed sequentially using the Gated Recurrent Unit (GRU).

5. Self-attentiveness is applied to the output of the GRU to emphasize important information at each step of the sequence. Multi-head attention is used to compute multiple attentions simultaneously.

6. The text of the speech is predicted in the fully connected layer (FC). The output dimension is 28 classes, including alphabets(27) and blanks(1).

7. The connectionist temporal classification loss (CTC) is calculated during training and testing.

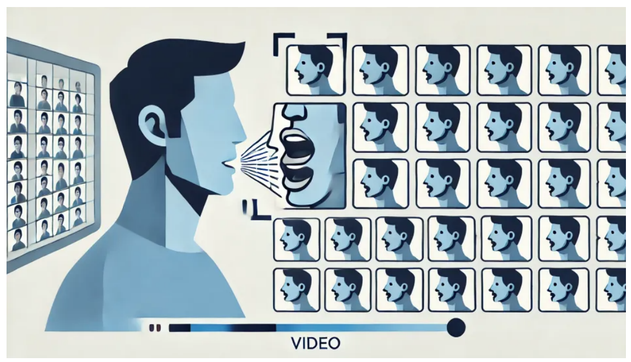
## Results

- Table 1. Original LipNet Model VS. Our LipNetWithAttention Model

| Scenario | Image Size (W x H) | CER | WER |
|---|---|---|---|
| Unseen speakers (Origin) | 128 x 64 | 6.7% | 13.3% |
| Overlapped speakers (Origin) | 128 x 64 | 1.9% | 4.6% |
| **Unseen speakers (Ours)** | 128 x 64 | **12.5%** | **17.9%** |
| **Overlapped speakers (Ours)** | 128 x 64 | **8.0%** | **10.8%** |

## Dataset

- Data Sources: GRID corpus, consisting of 34 speakers and 34,000 labeled sentences.
- Labeling Methodology: Use pre-aligned word and sentence labels. Preprocessing includes face detection and mouth-centered cropping.




## Discussion

- Attention is expected to improve model performance by emphasizing important information, helping the model learn more precisely. However, the results showed a deterioration in performance in terms of WER (Word Error Rate) and CER (Character Error Rate).
- We consider the following as possible causes for the performance degradation:

1. Misalignment Between Attention and CTC Mechanisms

-Problem: Attention focuses on global dependencies, while CTC requires strict monotonic alignment, leading to conflicts during optimization.

-Impact: The model struggles to align input frames with output sequences, increasing CER and WER, particularly on unseen datasets.

2. Overfitting Due to Increased Model Complexity

- Problem: The attention mechanism introduces additional parameters, making the model prone to memorizing patterns in overlapping datasets.

- Impact: This results in poor generalization, significantly increasing CER and WER on both overlapping and unseen datasets.

## Future Work

- Introduce monotonic attention mechanisms
- Add regularization techniques like Dropout in the attention layers
- Use techniques like Label Smoothing
- Train with data augmentation
- Apply RNN-ish skill into Attention layer
- Add the highway network before the bidirectional GRU layer.

## References

[1] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Deep audio-visual speech recognition. In arXiv:1809.02108, 2018.
[2] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. Lipnet: End-to-end sentence-level lipreading. arXiv preprint arXiv:1611.01599, 2016.
[3] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America, 120(5):2421–2424, 2006.