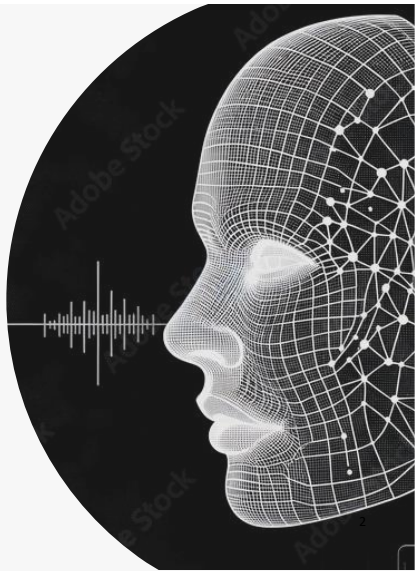
LIPNET: END-TO-END SENTENCE-LEVEL LIPREADING

- Harshith Deshalli Ravi – A-20552830
- Sachin Shivanna – A-20552795

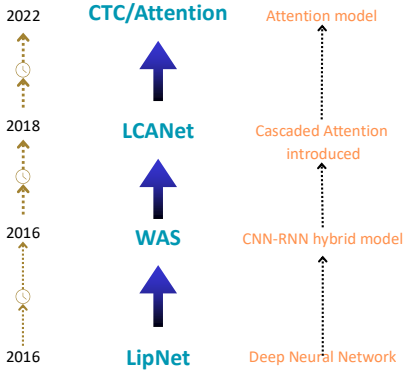
PROBLEM STATEMENT

The fine distinction is the major obstacle for humans to read from lips and as reported previously only around 20% reading accuracy can be achieved.

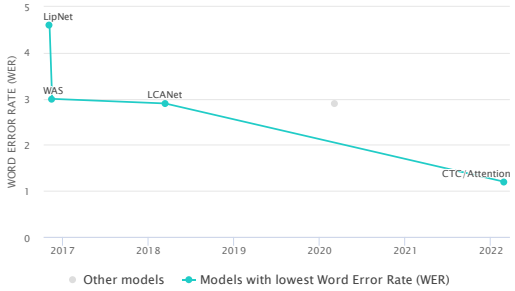
Digital editing tools enable audio manipulation in videos, leading to misinformation. Automated systems are needed to verify if lip movements match audio to detect such manipulations. Current methods often assess words individually, missing context and making it difficult to check alignment across entire sentences.



HISTORY



Source: GRID Corpus Benchmark (LipReading)

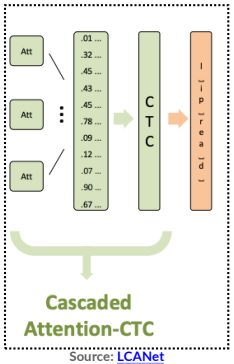


- Older approaches used word classification rather than classifying, rather than full sentence-level sequence prediction.

PROPOSAL

To enhance the LipNet model, we propose integrating techniques from the LCANet paper, where we are focusing on cascaded attention on Connectionist Temporal Classification (CTC) optimization. Using this architecture will significantly improve the decoder decision and more robust for continuous speech recognition.

The design has replaced GRU layer in the original LipNet and LCANet with Bi-Directional LSTM layer due to its accuracy in long-sequence understanding.



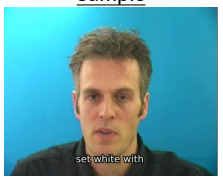
Source: LCANet paper

DATASET

Dataset Source: https://spandh.dcs.shef.ac.uk/gridcorpus/

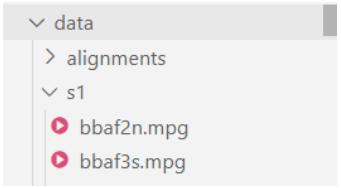
- **Train size:** 450 video clips
- **Test size:** 50 video clips

Sample

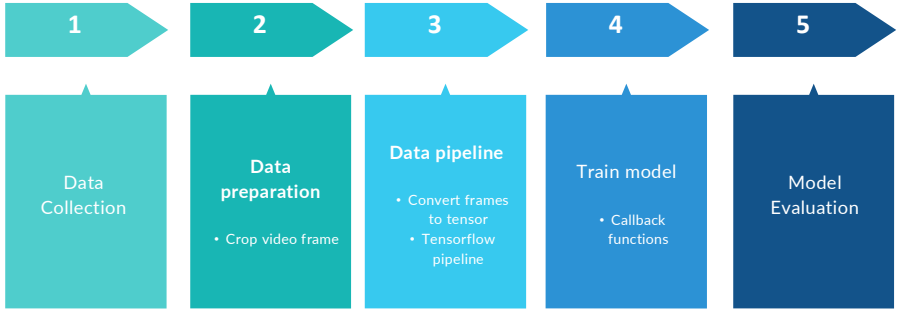


rizkiarm/LipNet git repo

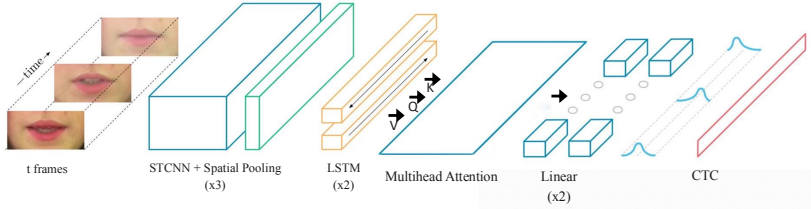
File structure



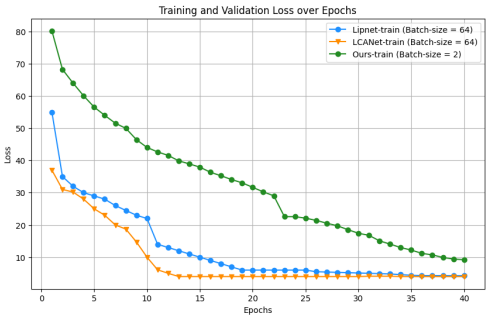
FLOW DIAGRAM



ARCHITECTURE



RESULTS OVERVIEW



Note : Batch size and dataset size is reduced due to memory constrains

Observations:

- By doubling the number of epochs with batch size 2 and dataset with 450 video clips, we were able to achieve better loss.

Model	Loss	Batch-size	Dataset size	Epochs
LipNet	4.27	64	25,330	50
LCANet	4.16	64	25,330	50
Our Model	3.90	2	450	100

```
Epoch 100/100
450/450 [==============================] - ETA: 0s - loss: 3.9059
```

DEMO

APPLICATIONS

- Speech understanding for people with hearing impairments
- Federal agencies to authenticate video evidence.
- Used for video forensics to track down lip-syncing deepfakes.
- Can be employed for security and surveillance to interpret conversations in situations where audio is not available or is of poor quality.
- The model can contribute to research in speech perception, language acquisition, and the relationship between visual and auditory processing in the brain.

REFERENCES

Research Paper-1 : LipNet: End-To-End Sentence-Level Lip Reading

Research Paper-2 : LCANet: End-to-End Lipreading with Cascaded Attention-CTC

Repository : https://github.com/rizkiarm/LipNet

Dataset : https://spandh.dcs.shef.ac.uk/gridcorpus/

Other Resources : https://paperswithcode.com/sota/lipreading-on-grid-corpus-mixed-speech

QUESTIONS?

THANK YOU!