LIP READING - A SURVEY
ECE 209AS PROJECT

Amulya Shruthi (505626283)
Shweta Katti (505604846)

INTRODUCTION

Lip reading refers to understanding speech or words based on analysing the speakers lip movements to interpret the content. It is also known as Visual Speech Recognition.

Lip reading is used in a lot of real world scenarios. This project can primarily help in the following use cases :

- Lip reading mainly helps those people with hearing loss and aids them with communication.
- Helps people work in situations where the background is noisy and assists them in interpreting speech.
- Lip reading softwares used in kiosks and ATMs can help people have contactless interactions and ensure privacy without speaking out loud.
- Biometric identification
- Can help in law enforcement - A few cases being it’s usage in forensic videos and CCTV footages in areas of crime.

OVERALL PROJECT GOALS

- **The overall goal of this project is to provide a comparative analysis on the various lip-reading techniques available and explore the various architectures.**

SPECIFIC AIMS

- Our aim was to test several popular models and architectures available for lip reading and conduct a survey to analyze their performance on a given test dataset.

- We selected a combination of popular and relatively new architectures based on our literature survey and obtained their pre-trained weights for evaluation.

- Utilized various metrics to understand which of the implemented models perform the best given a particular dataset.

STATE OF THE ART & ITS LIMITATIONS

- There has been a lot of work and progress on lip reading in the recent years.
- The following excel sheet link consists of our entire literature review conducted for ~20 research papers of the various models available:

  https://docs.google.com/spreadsheets/d/137hSvHNE2dYTGoN9PBO9bHsO_sFK0LwsScMN_iWpnRs/edit?usp=sharing

LIMITATIONS:

- Each of the proposed works available provides results of training and testing on a specific dataset.
- No survey available till date which tests on a specific dataset and compares the models.

NOVELTY OF THE PROJECT

- We test our choice of models on the same test dataset, in our case the LRW dataset.
- We find the Error Rates for evaluating and comparing the models.

TECHNICAL APPROACH

STEP 1	STEP 2	STEP 3	STEP 4
Pre-Processing (Any data augmentation/ data cleaning)	Extracting the required ROI (our region of interest being the lip)	Training the Model with the chosen dataset	Testing the Model and evaluating the metrics

DATASET REVIEW

01	**GRID Corpus**	• Consists of 34 english speakers • Videos are ~5 seconds long • Each subject speaks 1000 sentences each
02	**LRS2 - BBC**	• Collected from BBC television • Over 1000 sentences. Each sentence is upto 100 characters in length. • Divided into pre-train, train, test and validation sets.
03	**LRS**	• Collected from BBC television. Is an outdated version of LRS2 • Consists of fewer sets of data • Is not available and has been upgraded as LRS2
04	**LRW**	• Consists of 1000 utterances of 500 different words • Over 100 subjects/speakers • Train, validation, test sets

EVALUATION METRICS

- The evaluation of our models will be based on the Error Rates of the model prediction. There are multiple levels of calculating these error rates.
- The most widely used is the WER (Word Error Rate)

Any Error Rate has the following mathematical definition: ER = (S+D+I)/N

Here,
S = number of substitutions
D = number of deletions
I = number of insertions
N = Total number of words/characters

Multiple open-source packages available: asr_evaluation, fastwer, jiwer, STCK etc.

LIPNET

PAPER LINK:
https://arxiv.org/abs/1611.01599

- Architecture: Spatiotemporal CNNs + GRU + CTC
- Pre-processing: D-lib for face detection, data augmentation
- Dataset : GRID corpus
- Pre-trained weights available
- WER = 26.3%
- Dependencies: Keras, Tensorflow

TRANSFORMER MODEL

PAPER LINK:
https://arxiv.org/abs/1809.02108

- Architecture: TM - CTC model - Self-attention-based encoder architecture.
- Pre-processing: SSD face detection, pre-training and data augmentation
- Dataset : LRS2-BBC
- Pre-trained weights available online
- WER = 57.2%
- Dependencies: PyTorch

TEMPORAL CNNs

PAPER LINK:
https://ieeexplore.ieee.org/document/9053841

- Architecture used : Has 3D convolution + ResNet18 + MS TCN
- Preprocessing done : Face detection, face alignment, crop mouth ROI, convert to grayscale
- Dataset used - LRW BBC dataset
- Pretrained weights - obtained online
- WER = 12%
- Dependencies : PyTorch

RESNET WITH LSTM MODEL

PAPER LINK:
https://arxiv.org/pdf/1703.04105.pdf

- Architecture - Spatiotemporal convolution + ResNet + Bidirectional LSTM + Softmax layer
- Preprocessing - Detections of landmarks + cropping of ROI + conversion to grayscale
- Dataset: LRW BBC dataset
- WER = 24.3%
- Dependencies: PyTorch
- Weights -Trained on the lab GPU

WLAS

PAPER LINK:
https://arxiv.org/abs/1611.05358

- Dataset used for this is LRS2
- We were unable to obtain pre-trained weights or train the model from scratch
- Outdated code/dependencies
- Chose to skip it as a few papers use WLAS as a baseline model for comparison and other newer models perform better

RESULTS AND CHALLENGES

- Successfully implemented 4 models and evaluated them.
- Evaluation was done based on a common metric (WER) and test dataset and the models were compared.
- Temporal CNNs provided the best results.

- WLAS could not be implemented.
- Could not extend the testing with videos of our choice.

FUTURE WORK

- We can incorporate more personalized videos into the dataset for testing the models.
- We can check how the models performance varies when multiple datasets are combined and used for training.
- We can incorporate other evaluation metrics such as CER or SER.

INDIVIDUAL CONTRIBUTIONS:

- Literature Survey was split equally between the two of us.
- The testing of models was also equally split between the two of us with each one of us testing 2 models each.
- Slides and Github Repo updation was done based on individual work.

THANK YOU!