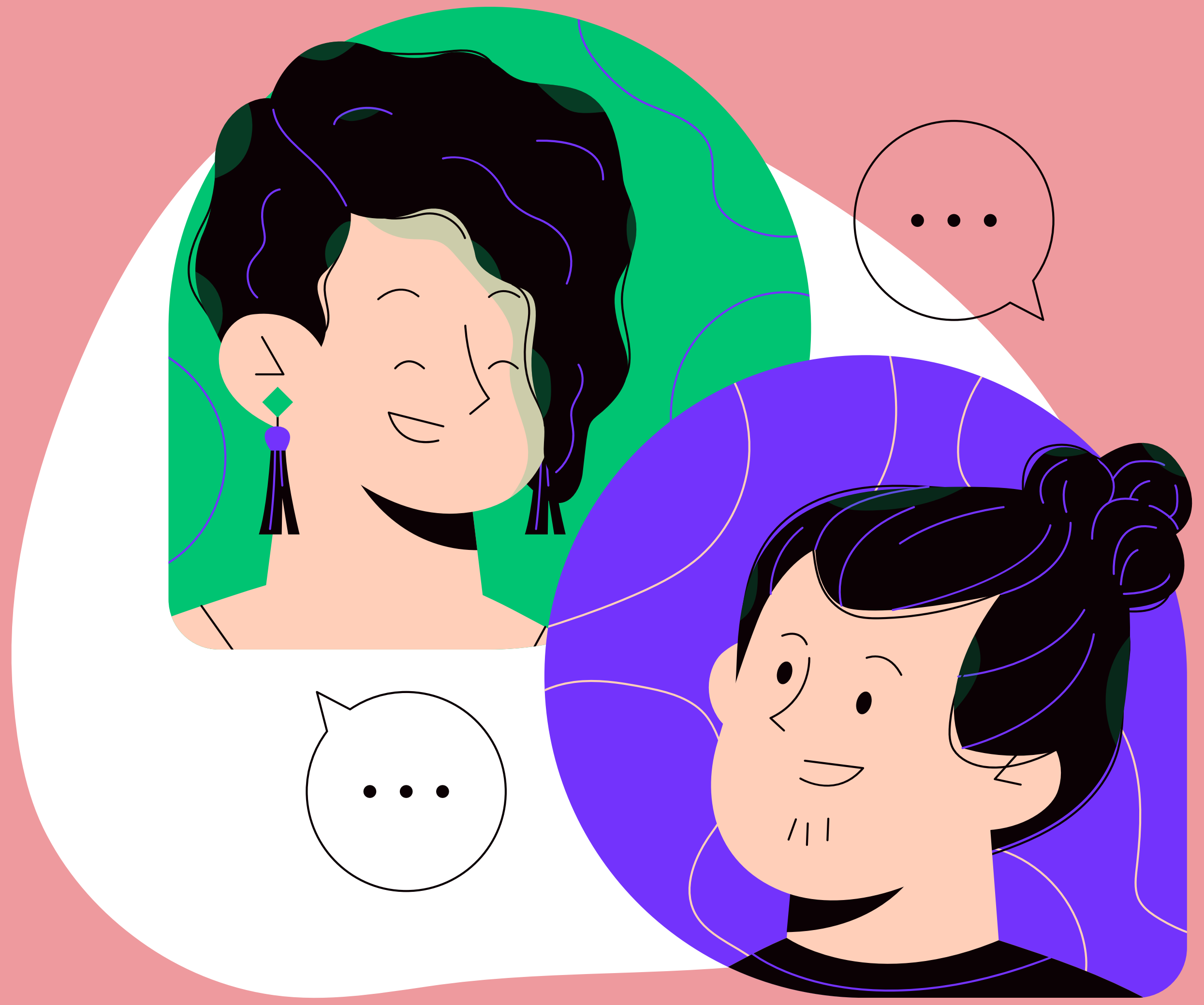
CONTENTS



1 INTRODUCTION

2 RELATED WORKS

3 METHODS

4 EXPERIMENTS

5 CONCLUSION

1.INTRODUCTION

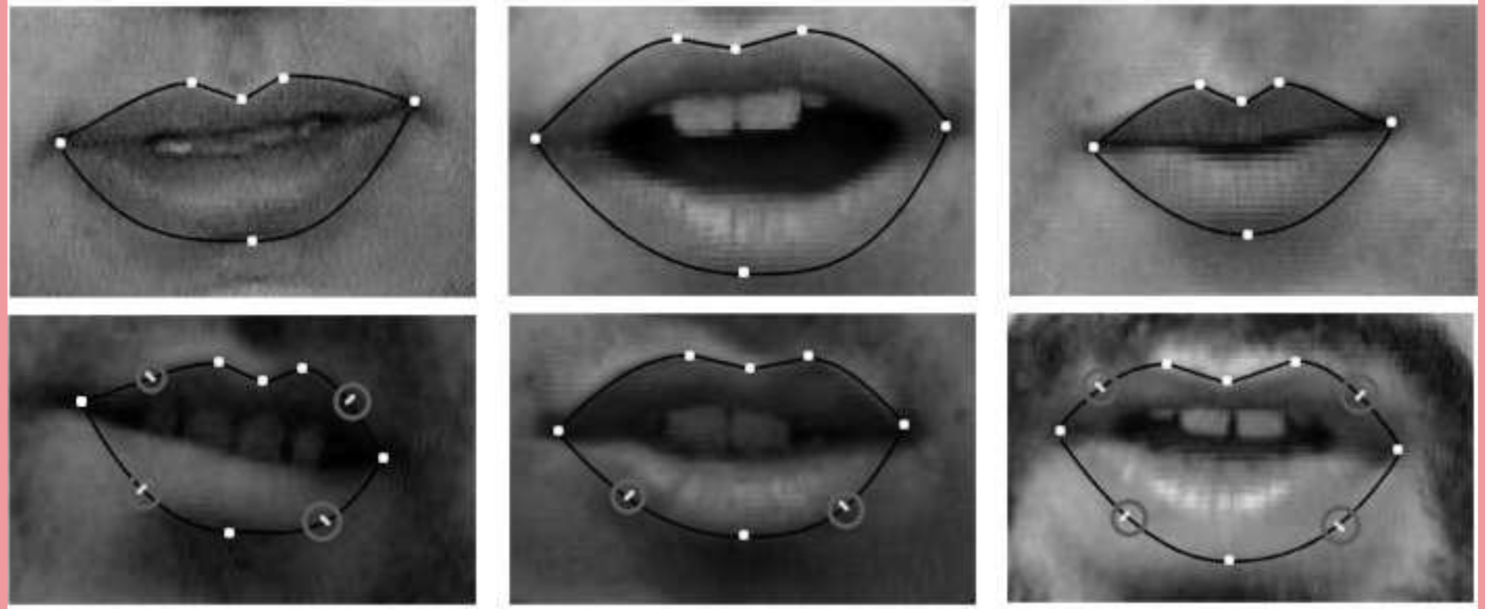
What is Lip reading?

The ability that recognizes what is being said only with visual information around the lip

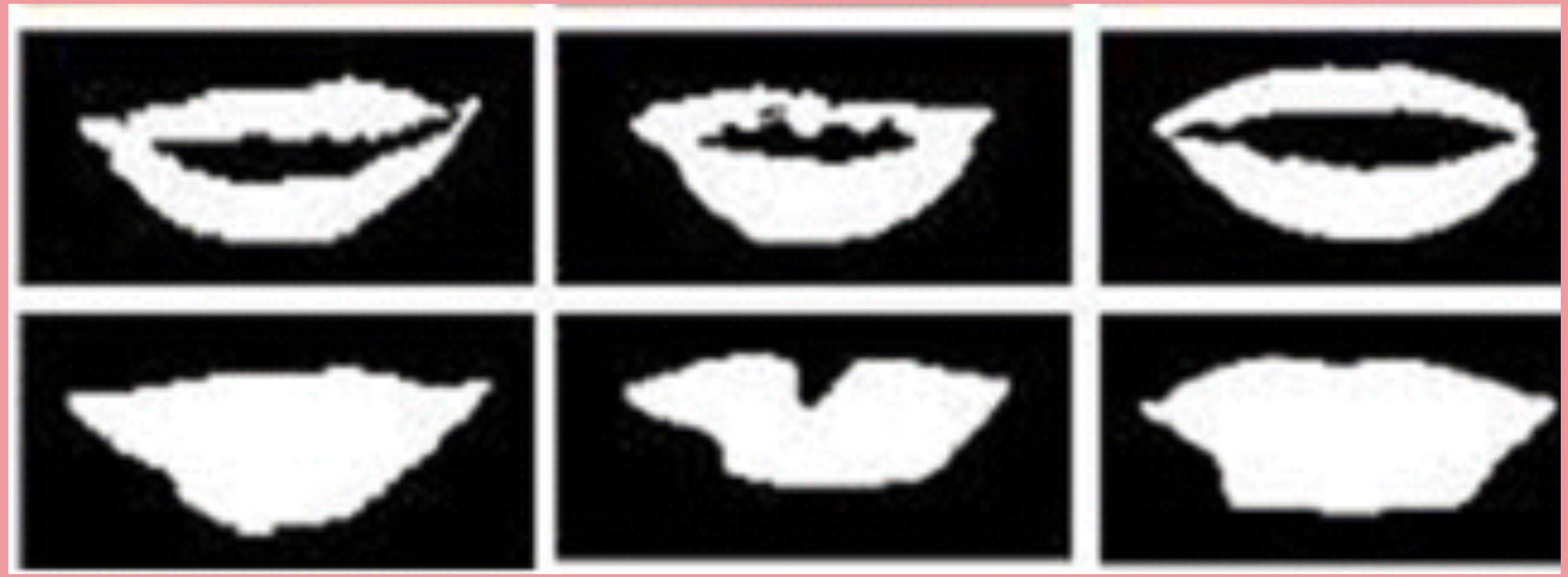
1.INTRODUCTION

What we aim ?

Improving Lip reading performance by utilizing methodologies such as Lip contour detection and Lip region detection in pre-processing stage



Lip contour



Lip Region

1.INTRODUCTION

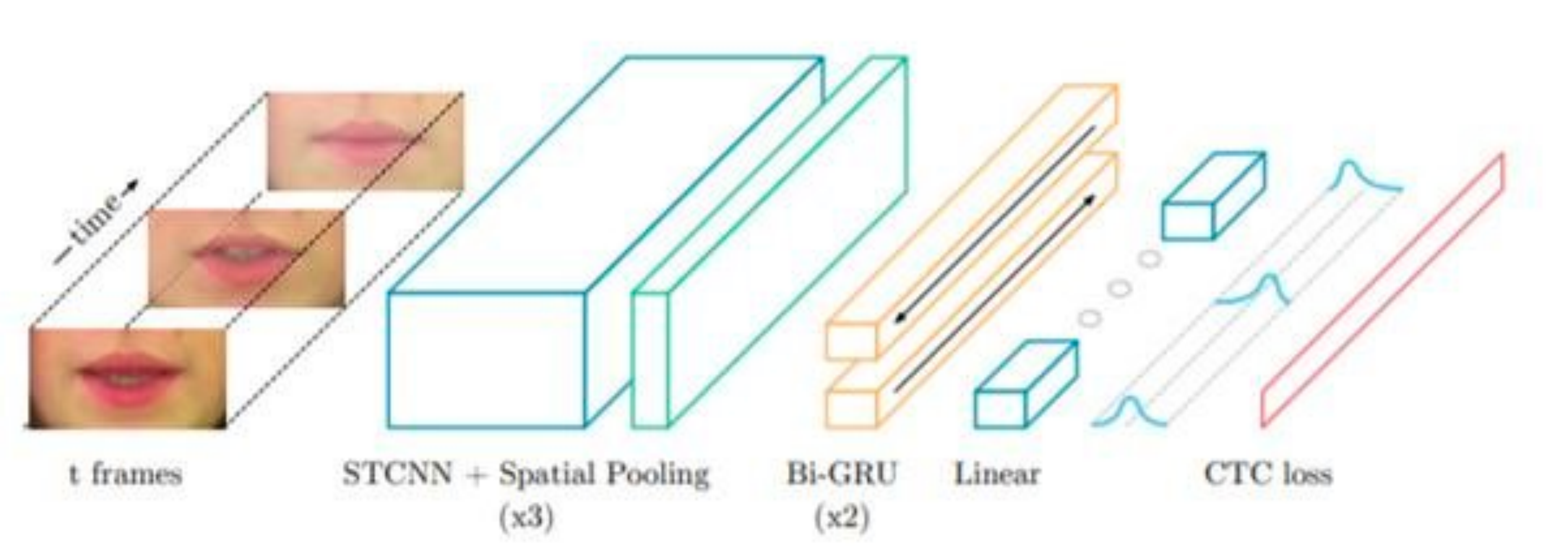
What we aim ?

- In order to compare performance, we will use the LipNet deep neural network model commonly used in LipReading.
- Lip Reading have been conducted with English datasets, we apply our methodology with the best performance to Korean dataset

2.RELATED WORKS

❶ LipReading

- Machine learning approaches
  - Random forest manifold alignment
  - HMMs
- Deep learning approaches
  - CNN + LSTM
  - LipNet



2.RELATED WORKS

❷ Lip Segmentation

- In lip reading systems, Lip segmentation holds a fundamental role, as the accuracy of the segmentation directly influences the overall recognition rate.

- Two types of lip segmentation we consider
  - Lip Contour : BlazeFace (Mediapipe Face Mesh)
  - Lip Region : BiseNet V2

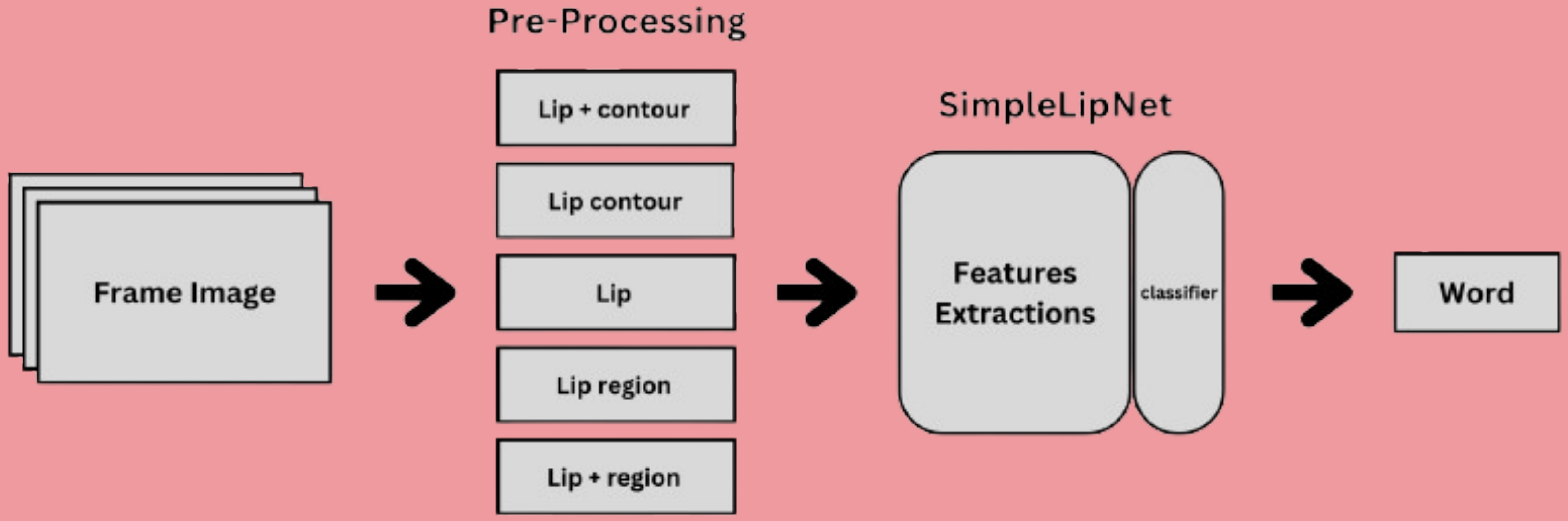
3.METHODS

Improving Lip reading performance by highlighting lip segment in pre-processing stage

❶ Lip Segmentation: Extracting and segmenting the lip region and solve them as a Computer Vision problem

- Lip contour: BlazeFace
- Lip region: BiseNet V2

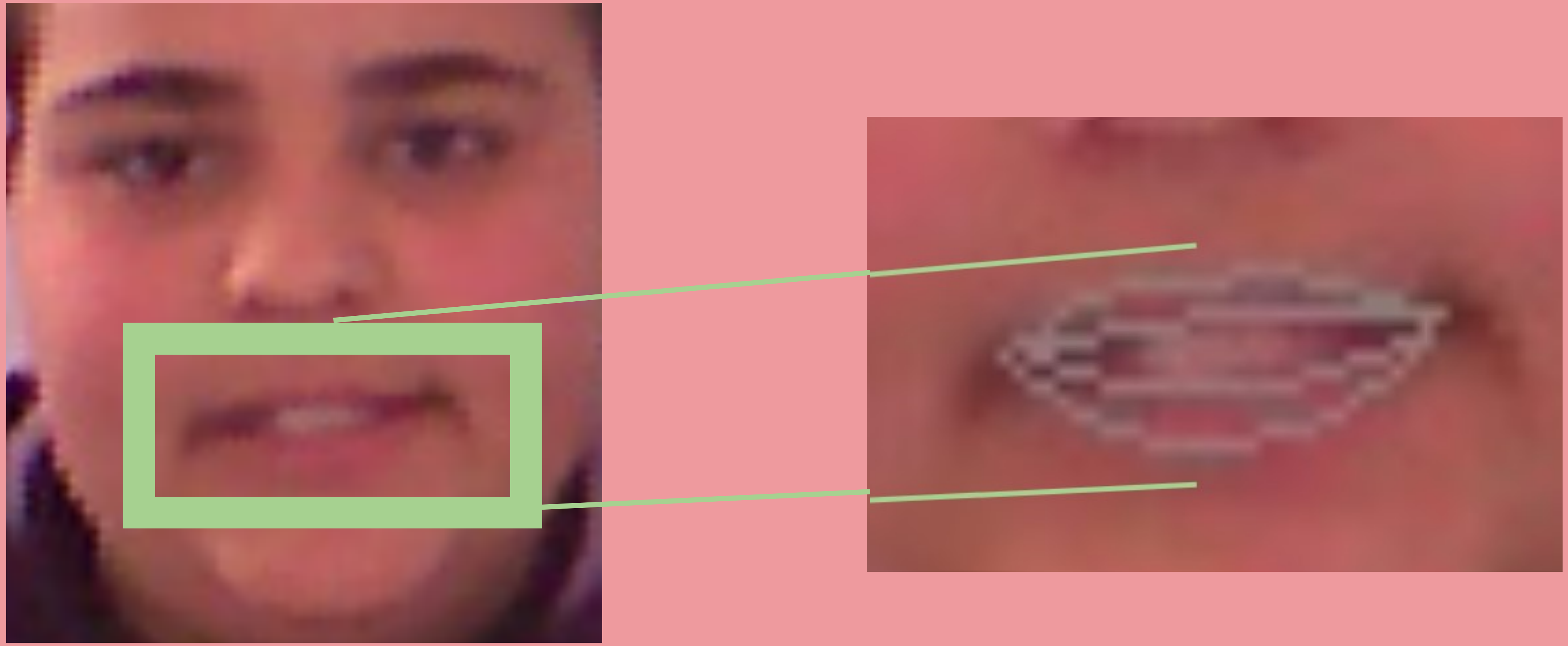
❷ LIPNET: The resultant lip segment serves as input and are used for training the LipNet model



3.METHODS

❶ Lip Segmentation

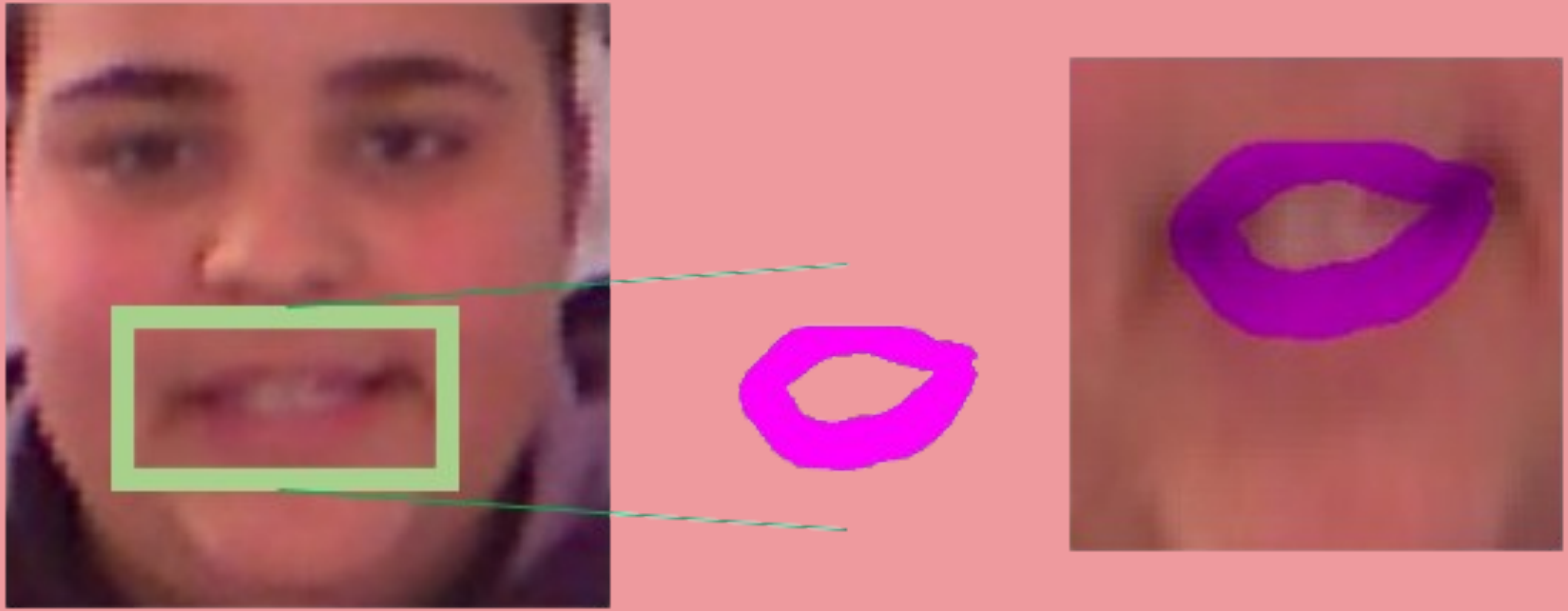
- Lip contour : BlazeFace (MediaPipe)
  - Designed for high speed predictions on mobile GPUs
  - well-suited for lip reading that demands real-time resolution



3.METHODS

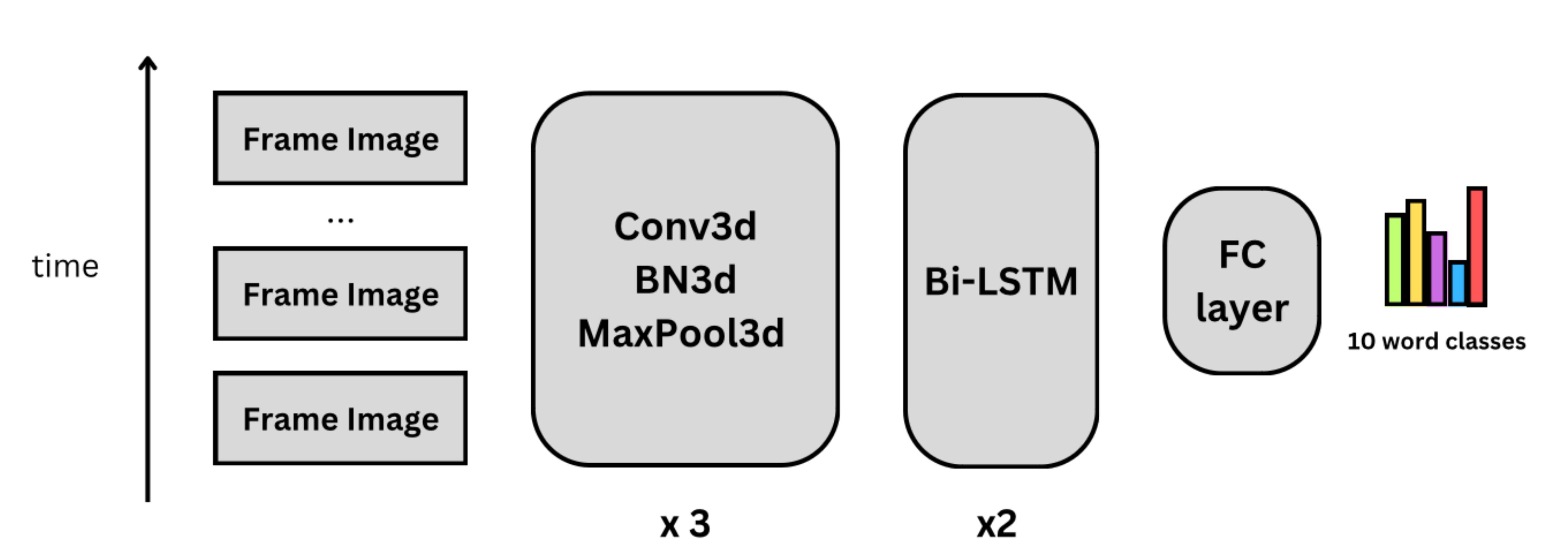
❶ Lip Segmentation

- Lip region : BiseNet V2
  - Represents a real-time semantic segmentation model that performs both high accuracy and preservation of low-level details
  - The architecture comprises the Detail Branch and the Semantic Branch



3.METHODS

❷ LIPNET



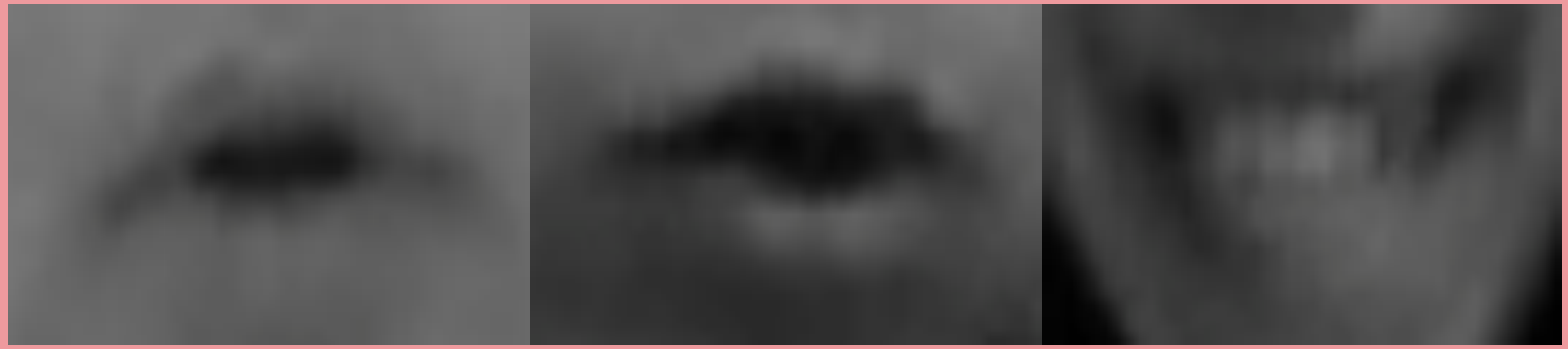
- Simplified Lipnet Architecture
  - 3x Convolution, Batch Normalization, MaxPooling layers
  - 2x Bi-LSTM
  - FC layer
  - Total number of parameters: 733K

3.METHODS

❷ LIPNET

- Baseline
  - The LipNet paper processed the image using the DLib face detector and the iBug face landmark predictor with 68 landmarks.
- For a fair comparison, a lip image was obtained with 30×45 by applying the preprocessing method of the original LipNet

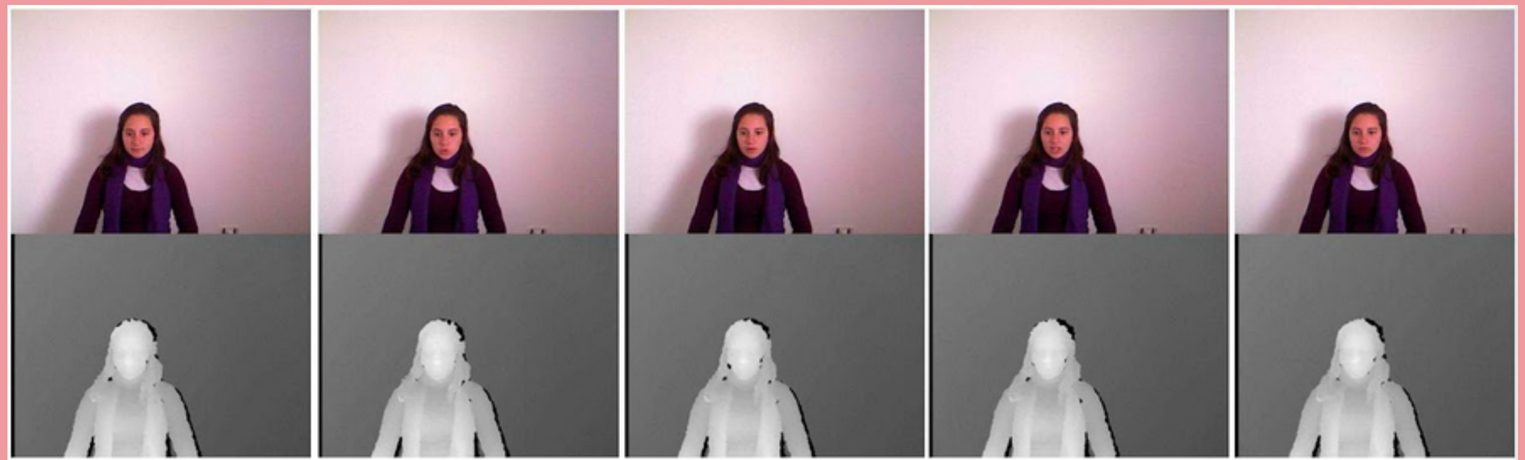
We trained simple LipNet with the two lip segmentation methods and baseline methods described earlier, and compared the test performance



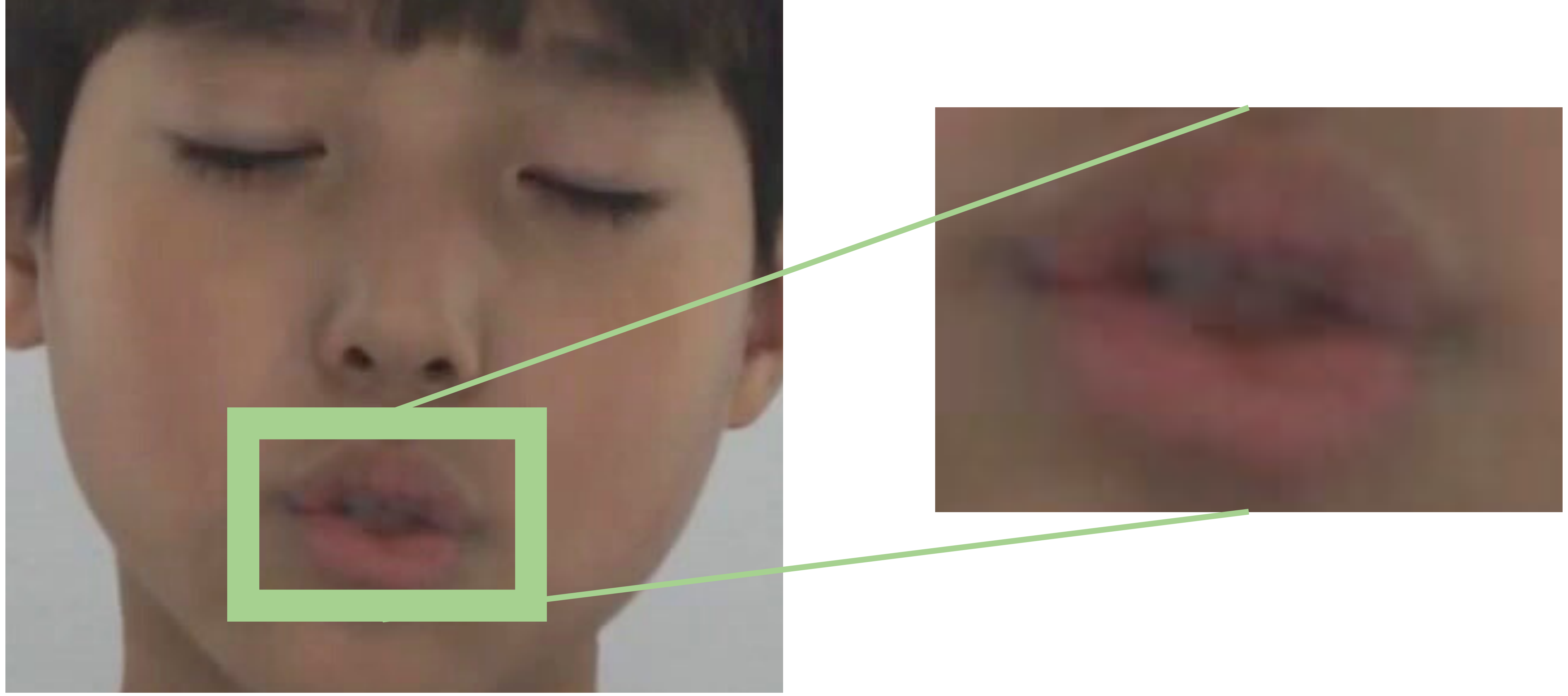
4. EXPERIMENTS

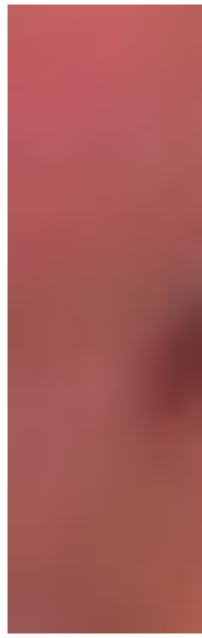
❶ Datasets

- Miracle-VC
  - Used for performance comparison between techniques
  - Contains depth and color images of 15 speakers uttering 10 words and 10 phrases, 10 times each
  - We use only 10 words data and color images.
    - words : begin, choose, connection, navigation, next,previous, start, stop, hello, web
  - 3 speakers had difficulty in lip segmentation process and were excluded
  - remaining 12 speakers : 9 train, 1 validation, 2 test



❶ Datasets





- AIHU

4. EXPERIMENTS

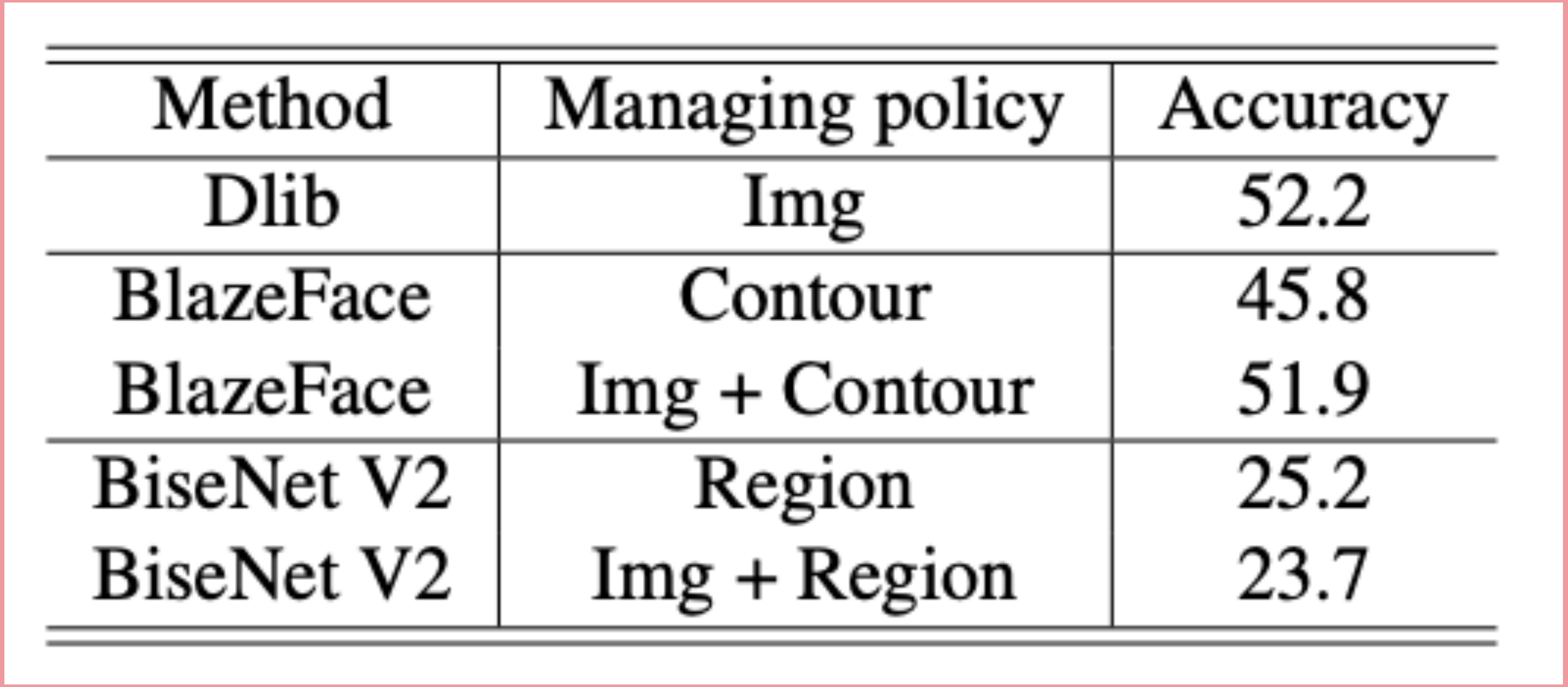
❷ Experimental Settings

- Data augmentation
    - randomly rotates the image by a 15 degrees
    - randomly mirrors the image across a vertical axis with a probability of 0.5
    - randomly apply resized crop on the image

- Hyperparameter setting
    - epoch: 50
    - optimizer: Adam, learning rate=1e-4, and weight decay=5e-3
    - batch size: 32
    - number of frames per each words: 16
    - loss: cross entropy

4. EXPERIMENTS

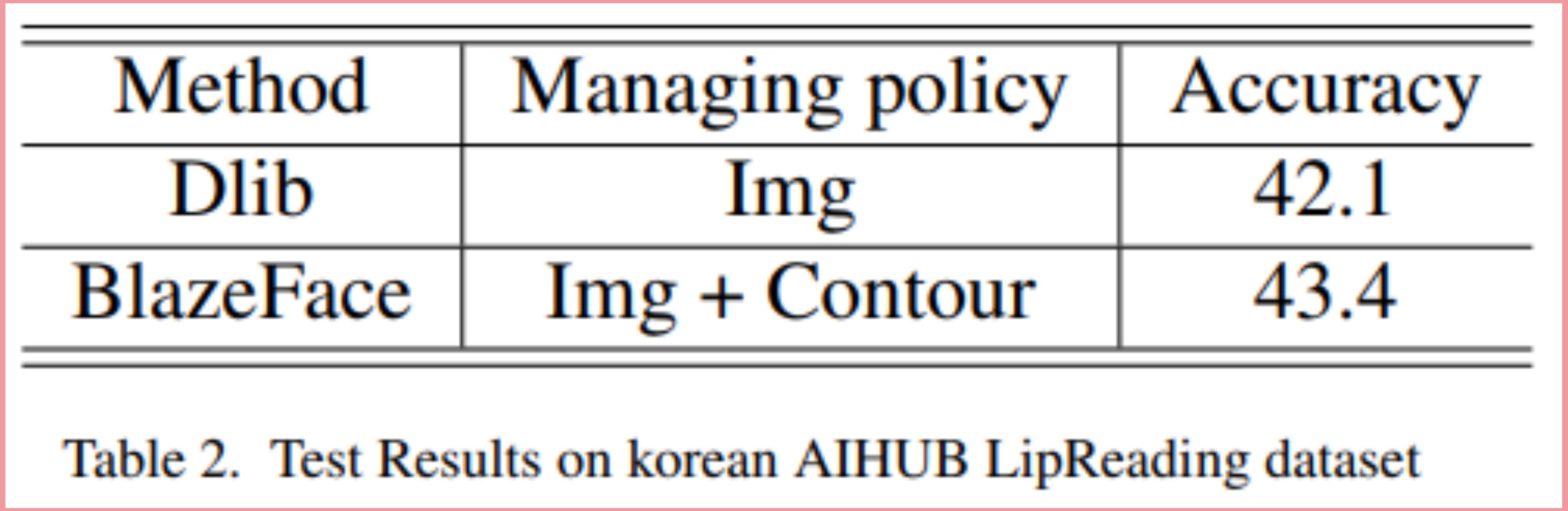
❸ Miracle-VC Experiments Results



Method	Managing policy	Accuracy
Dlib	Img	52.2
BlazeFace	Contour	45.8
BlazeFace	Img + Contour	51.9
BiseNet V2	Region	25.2
BiseNet V2	Img + Region	23.7

4. EXPERIMENTS

❹ AIHUB Korean LipReading Experiments Results



Method	Managing policy	Accuracy
Dlib	Img	42.1
BlazeFace	Img + Contour	43.4

Table 2. Test Results on korean AIHUB LipReading dataset

- The preprocessing method using original image and lip contours showed better performance than baseline.
- However, it is difficult to say that this difference is significant.

4. EXPERIMENTS

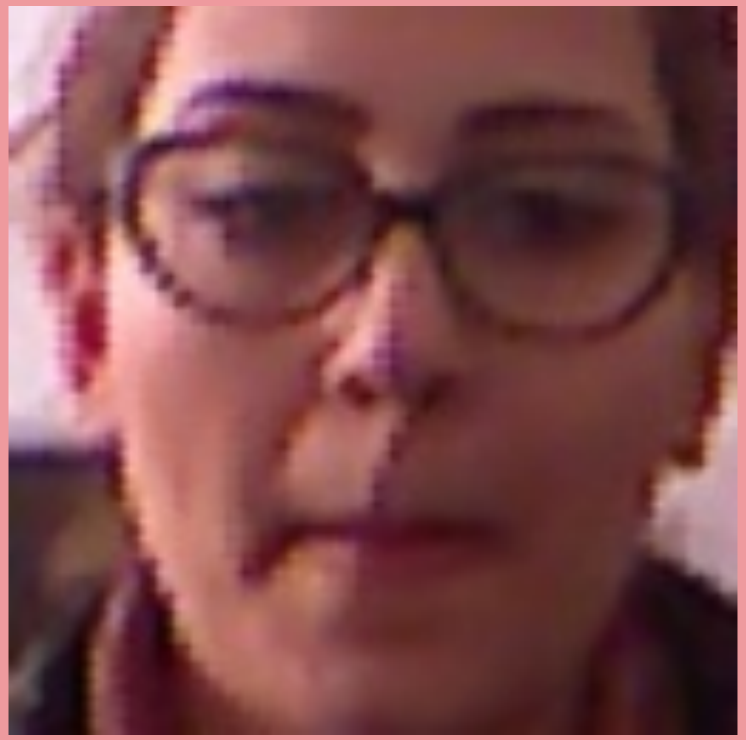
❹ Live Demo

Let's check out Live Demo for our best model!

5.CONCLUSION

❶ Discussion - Result 1

- **Possible reasons why our methods using Lip Region are poor?**
  - **Accuracy issue in preprocessing steps** : In particular, when the lip’s shape was unclear or when the speaker's background color is similar to face or lip color, the performance of finding lips using BiseNet V2 was very poor.



5.CONCLUSION

❶ Discussion - Result 2

- **Possible reasons why don’t our methods outperform baseline?**
  - **Accuracy issue in preprocessing steps** : Incorrect identification of lip boundaries leads to the transmission of misleading information.
  - **Overfitting due to lip contour** : While LipNet could benefit from utilizing Lip features in a “wild” environment, our dataset used in this project was generated in a controlled lab environment and was difficult to obtain its benefits.
  - **Insufficient information of lip contour**: In lip reading, other visual features may be more important than Lip Contour such as Lip Movement.

5.CONCLUSION

❷ Future Works

- **Improve Preprocessing Accuracy**
  - Lip segmentation was conducted with a relatively good performance model, but there was still a limitation in accuracy.
  - We will be able to try different preprocessing models that show better performance in environment of the data we used, or try different datasets.

- **Use another model**
  - LipNet is a relatively outdated lip-reading model.
  - It is necessary to increase the overall word classification performance by applying a relatively latest model.

THANKYOU