PURDUE
UNIVERSITY

Lip-to-Text for the Hearing Impaired: Multi-Modal Approach Using Vision-and-Language Transformer

CNIT 581-AST Project, Fall 2022

Nadine, Srushti, Yi

October 4, 2022

Outline

- Our Team
- Introduction
  - Motivation
  - Objective
- Literature Review
- Targeted Gap
- Our Proposed Approach
- Equipment List
- Time Plan

Our Team

Introduction

Motivation

Lip-to-Text for the Hearing Impaired: Multi-Modal Approach Using Vision-and-Language Transformer

1.5 billion people with hearing loss, 25% of people over 60 years (WHO, 2021)

heavy dependence on lip reading & multi-tasking can be impractical

considering several modalities boosts performance

efficiency and speed are crucial

Objective

Evaluate the performance of the Vision-and-Language Transformer (ViLT) in a visual-linguistic speech recognition system.

Our focus is on training and testing the ViLT model using publicly available datasets; we are not concerned with how the input data is obtained or output is displayed in real-time.

Literature Review

Literature Review

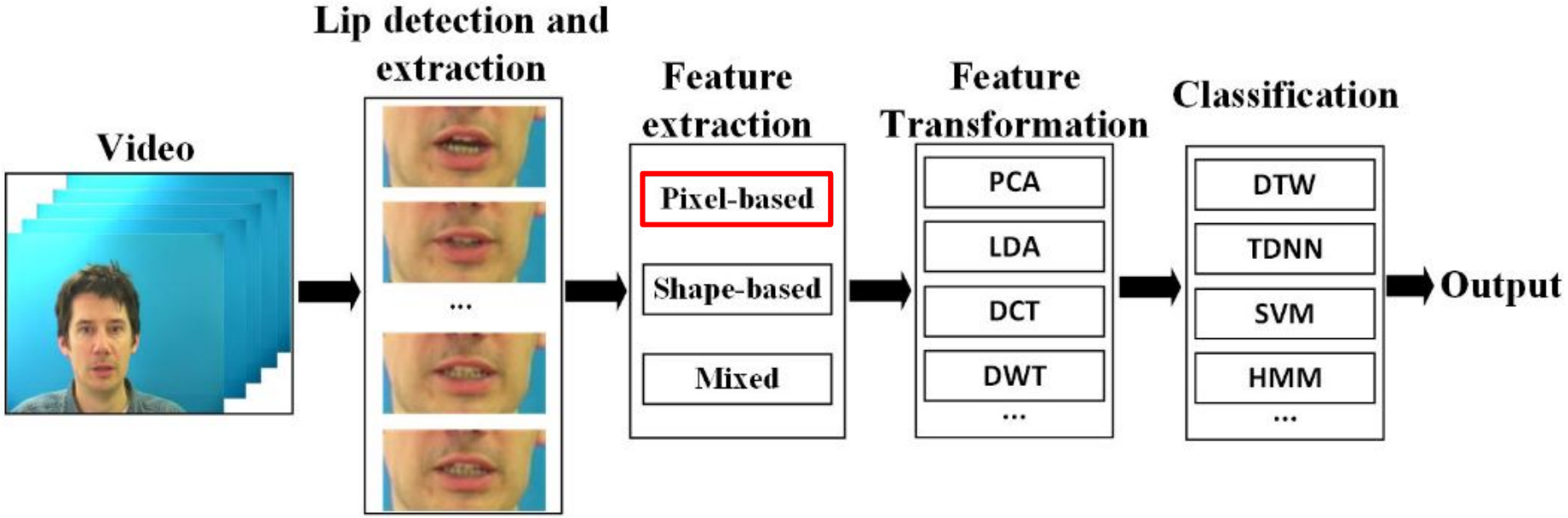
- **Lip Reading**

Traditional Lip Reading

Deep Lip Reading

- **Multi-Modal Approach**

Traditional Lip Reading

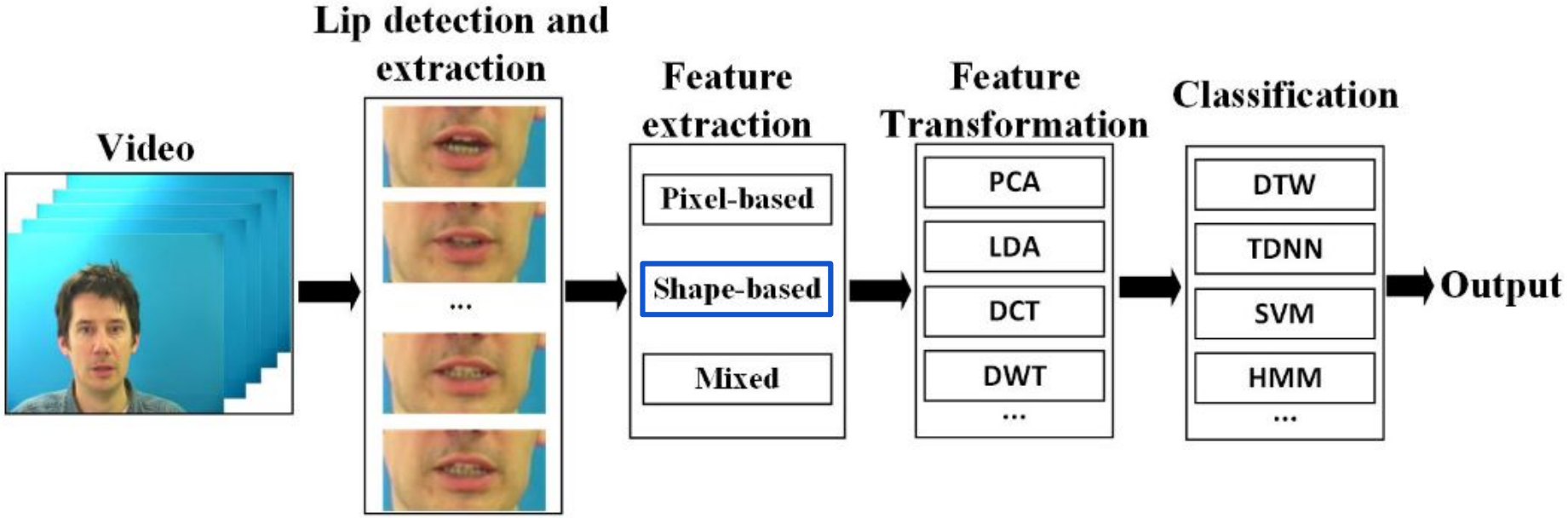


Morade & Patnaik, 2014

Sterpu & Harte, 2017

(Hao et al., 2020)

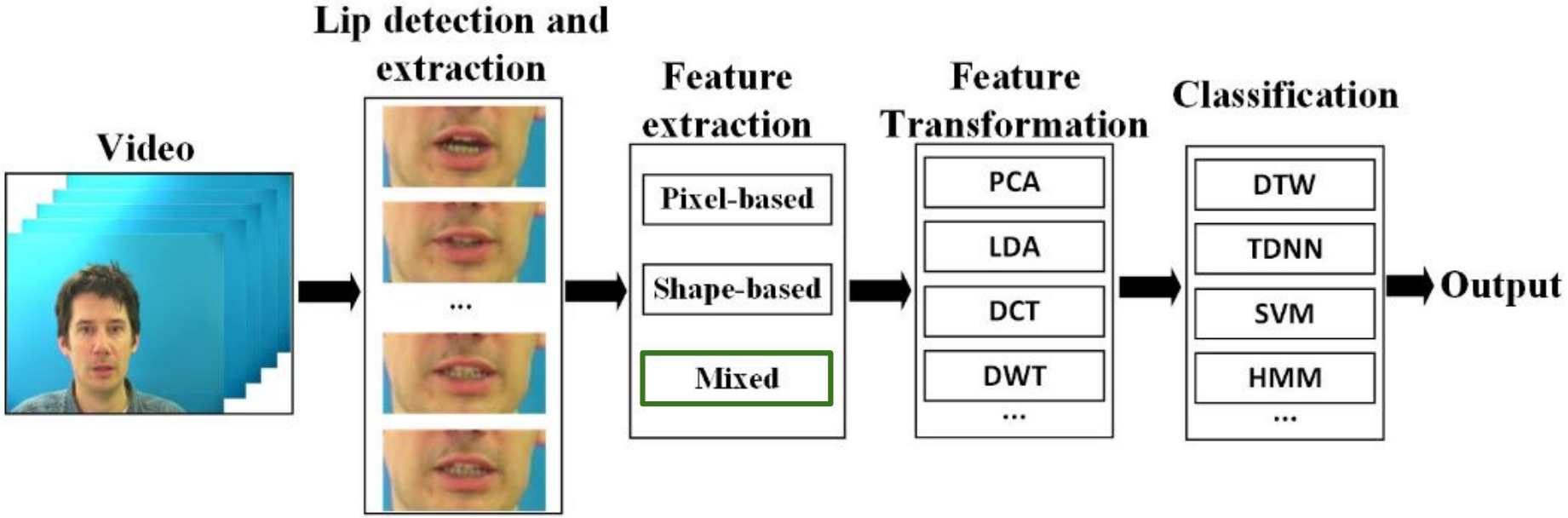
Traditional Lip Reading



Luettin and N. A. Thacker, 1997
Ma et al., 2016

(Hao et al., 2020)

Traditional Lip Reading

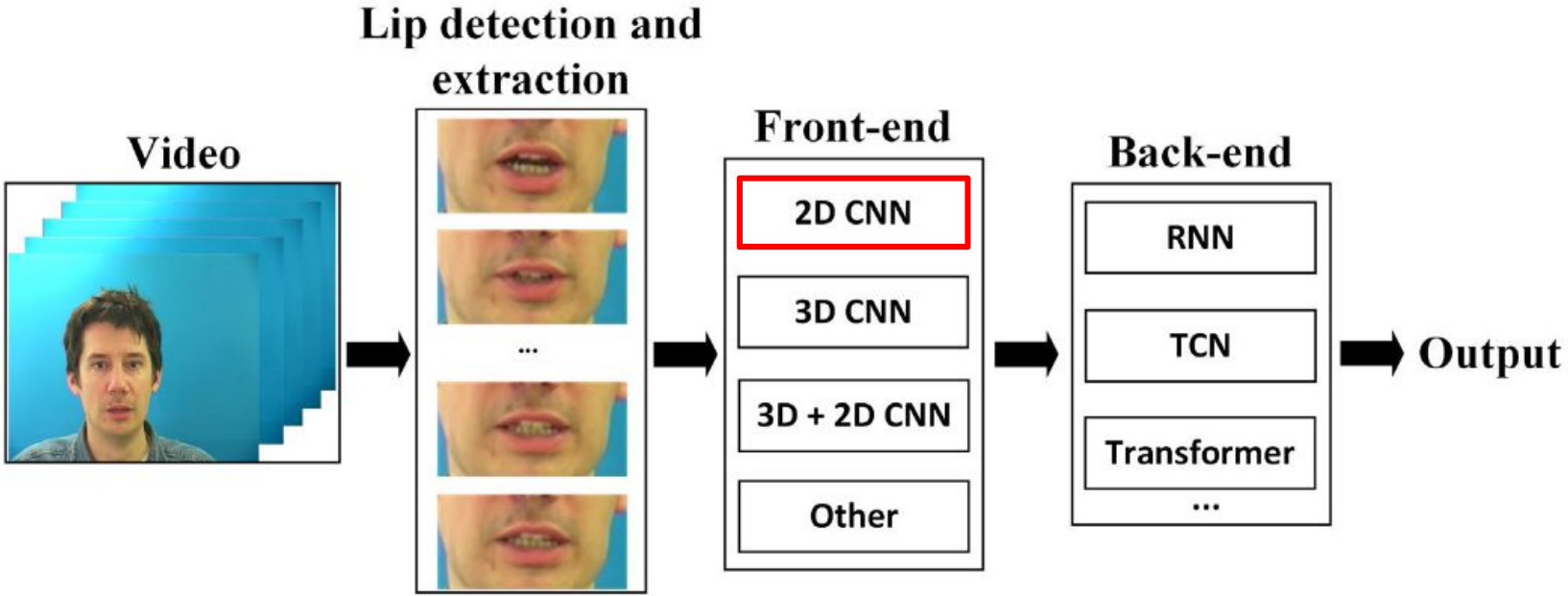


Bear et al., 2017
Watanabe et al., 2016

Howell et al., 2016

(Hao et al., 2020)

Deep Lip Reading



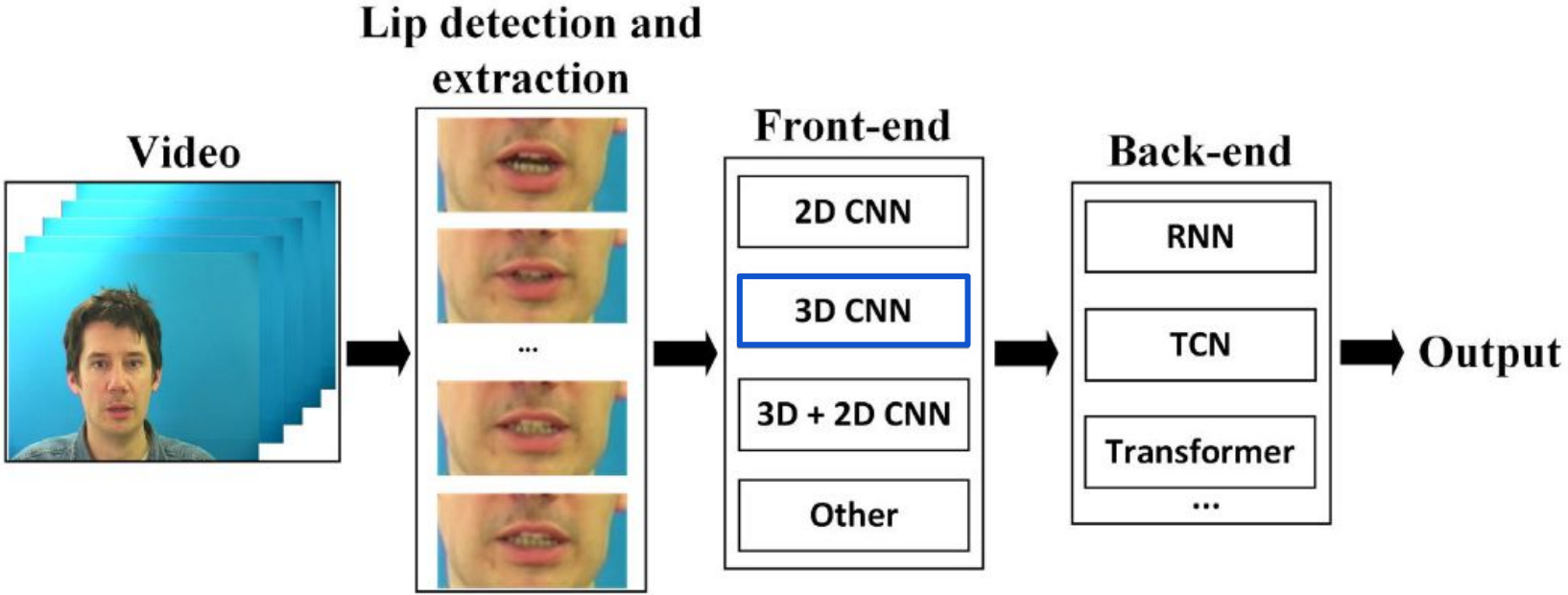
Garg et al., 2016
Noda et al., 2014

Li et al., 2016
Saitoh et al., 2016

Mesbah et al., 2019
Zhang et al., 2019

(Hao et al., 2020)

Deep Lip Reading



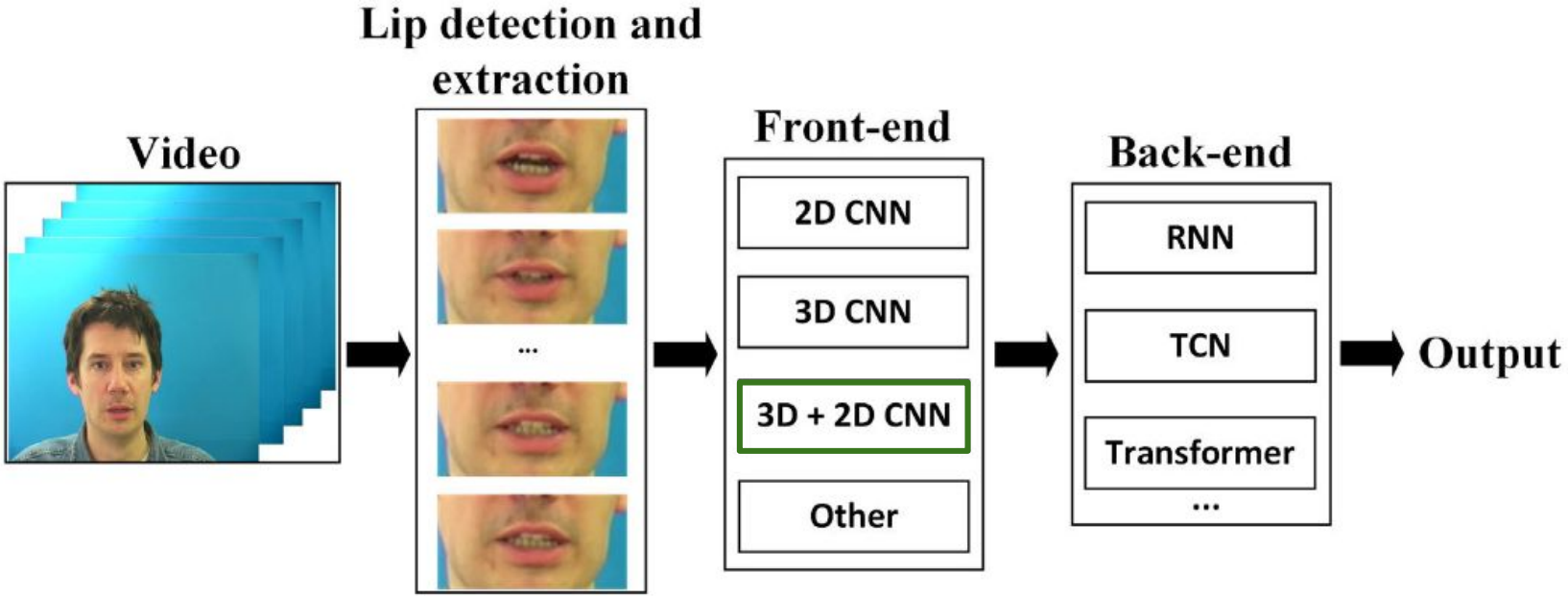
Assael et al., 2016
Torfi et al., 2017

Fung & Mak, 2018
Tran et al., 2017

Qiu et al., 2017
Yang et al., 2019

(Hao et al., 2020)

Deep Lip Reading

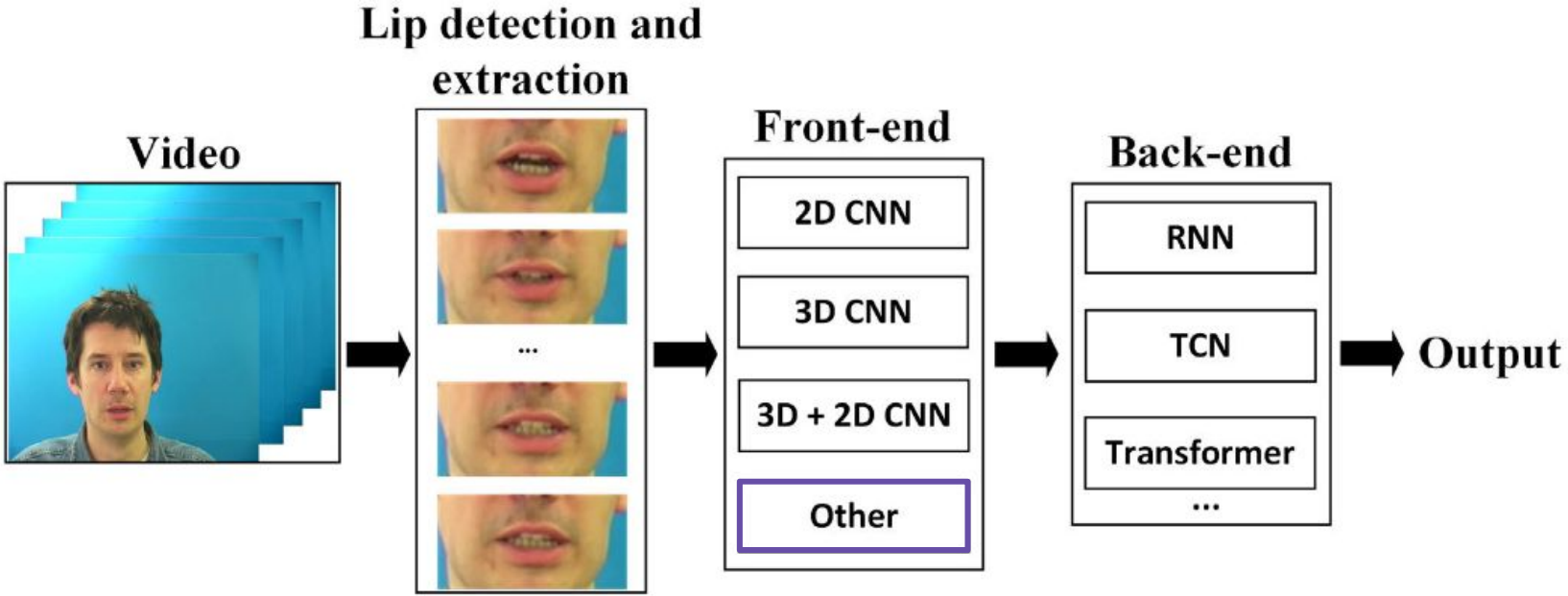


Margam et al., 2019
Stafylakis & Tzimiropoulos, 2017

Petridis et al., 2018

(Hao et al., 2020)

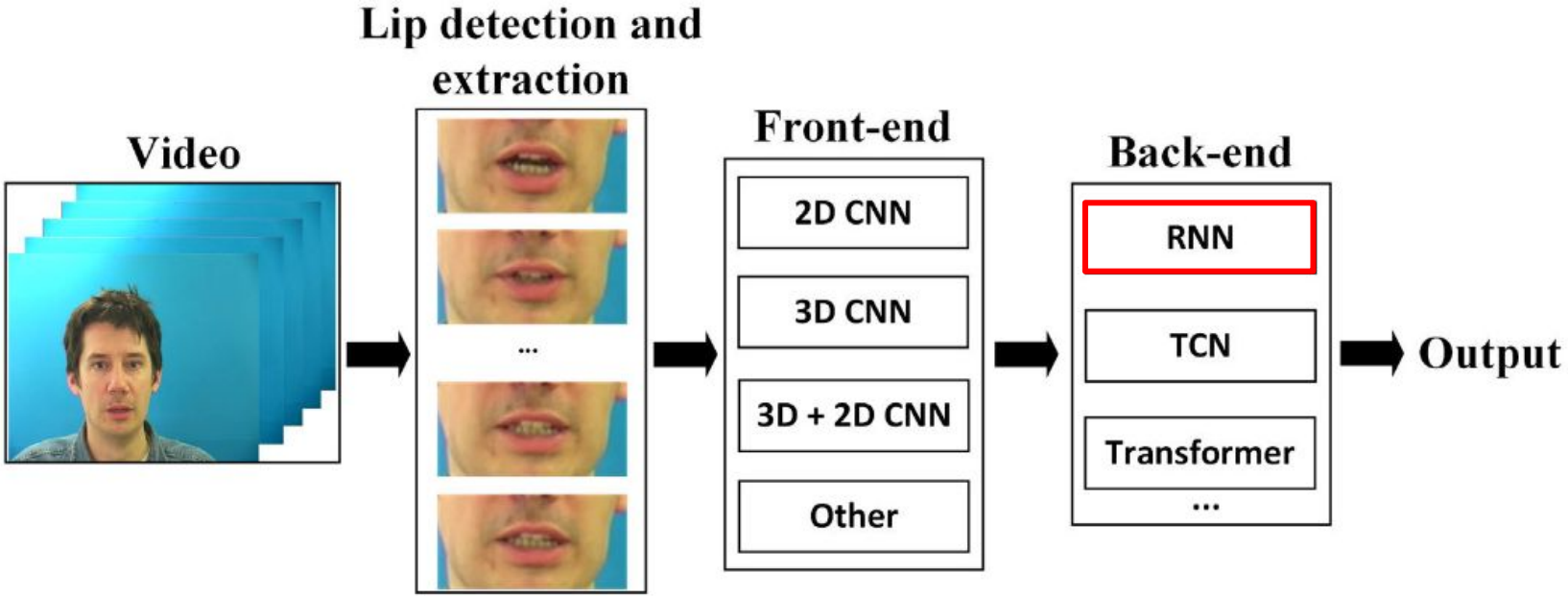
Deep Lip Reading



FNN: Wand et al., 2016, 2017 & 2018
Autoencoder: Petridis et al., 2017 & 2018

(Hao et al., 2020)

Deep Lip Reading

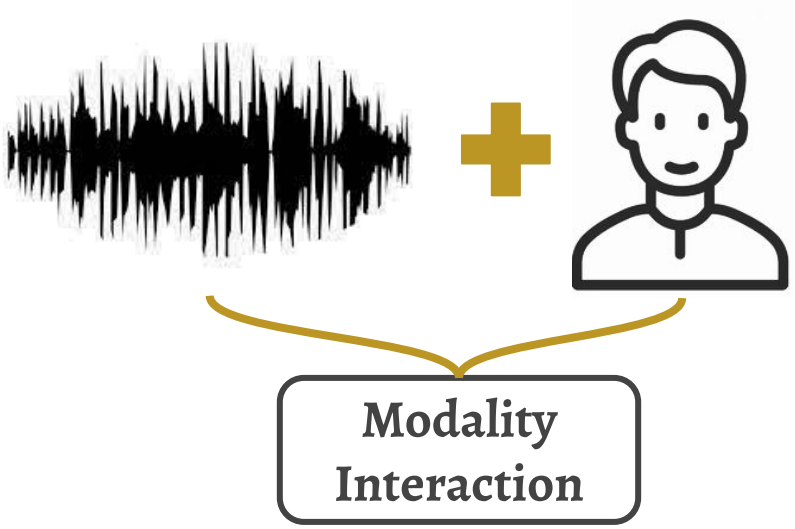


Bi-LSTM: Stafylakisa et al., 2018; Weng & Kitani, 2019

Bi-GRU: Luo et al., 2020; Xiao et al., 2020; Zhao et al., 2020; Zhang et al., 2020

(Hao et al., 2020)

Multi-Modal Approach



Audio-Visual System

Isobe et al., 2021
Ma el al., 2021

Kumar et al., 2022
Yu et al., 2021

Liu et al., 2021

Targeted Gap

Targeted Gap

The system heavily relies on computationally complex feature extraction from visual input.

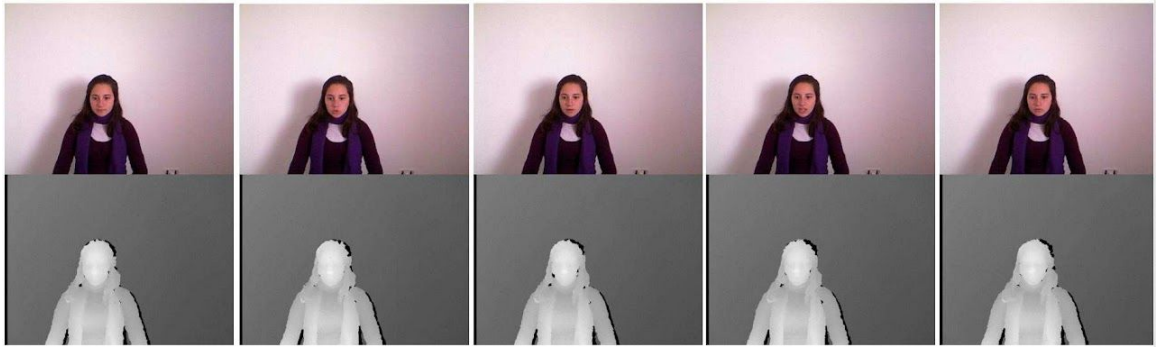
affects efficiency & speed of overall system

(Kim et al., 2021)

Our Proposed Approach

Dataset

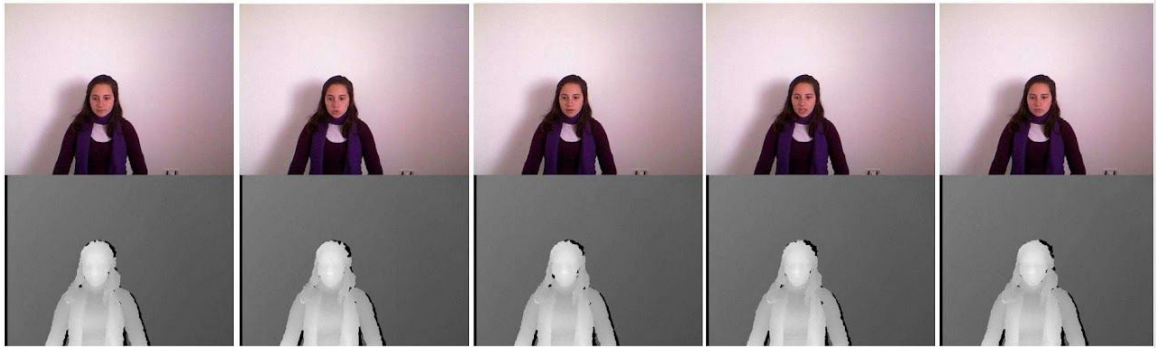
- **MIRACL-VC1 (Rekik et al., 2014):** a lip-reading dataset including both depth and color images
  - Captured by Microsoft Kinect sensor, 640x480 pixels
  - 15 speakers (five men and ten women)
  - Each speaker read 10 times for a set of ten words and ten phrases
  - A total number of 3000 instances (15 x 20 x 10)



ID	Words	ID	Phrases
1	*Begin*	1	*Stop navigation.*
2	*Choose*	2	*Excuse me.*
3	*Connection*	3	*I am sorry.*
4	*Navigation*	4	*Thank you.*
5	*Next*	5	*Good bye.*
6	*Previous*	6	*I love this game.*
7	*Start*	7	*Nice to meet you.*
8	*Stop*	8	*You are welcome.*
9	*Hello*	9	*How are you?*
10	*Web*	10	*Have a good time.*

Dataset

- MIRACL-VC1 (Rekik et al., 2014): a lip-reading dataset including both depth and color images
  - Captured by Microsoft Kinect sensor, 640x480 pixels
  - 15 speakers (five men and ten women)
  - Each speaker read 10 times for a set of ten words and ten phrases
  - A total number of 3000 instances (15 x 20 x 10)



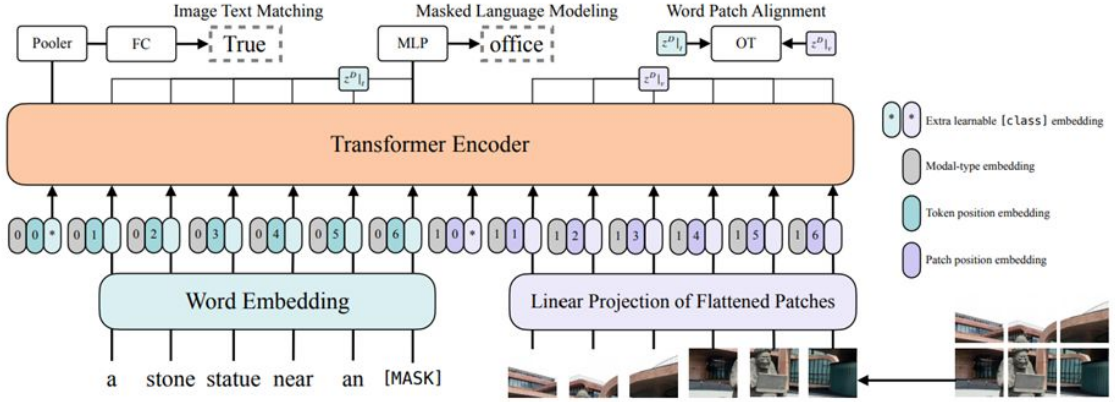
https://link.springer.com/chapter/10.1007/978-3-319-11755-3_3

ID	Words		ID	Phrases
1	*Begin*		1	*Stop navigation.*
2	*Choose*		2	*Excuse me.*
3	*Connection*		3	*I am sorry.*
4	*Navigation*		4	*Thank you.*
5	*Next*		5	*Good bye.*
6	*Previous*		6	*I love this game.*
7	*Start*		7	*Nice to meet you.*
8	*Stop*		8	*You are welcome.*
9	*Hello*		9	*How are you?*
10	*Web*		10	*Have a good time.*

Model Selection

- **Vision and Language Transformer (ViLT)** (Kim et al., 2021)
  - Four image-text datasets: COCO, Visual Genome, Conceptual Captions, & SBU Captions
  - Two pre-training tasks:
    - Image Text Matching
    - Masked Language Modeling

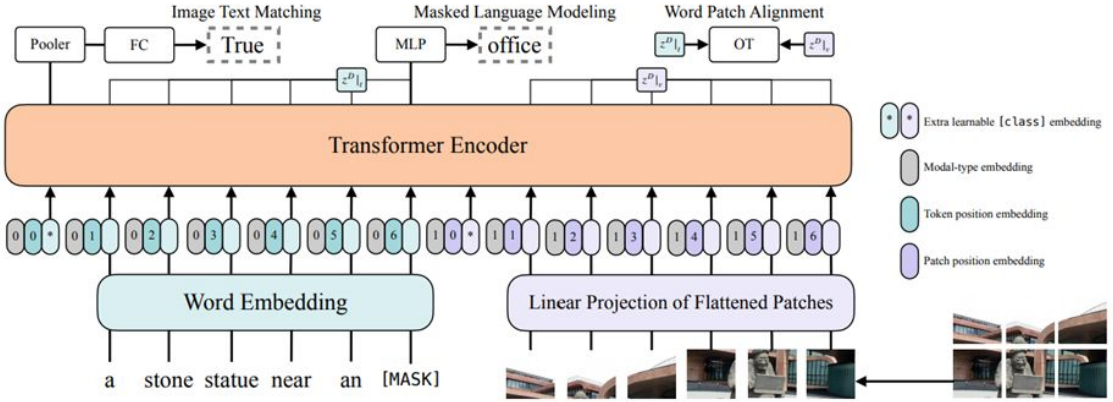
Visual Embed	Model	#Params (M)	#FLOPS (G)	Time (ms)
Region	ViLBERT	274.3	958.1	~900
	UNITER	154.7	949.9	~900
Linear	ViLT	87.4	55.9	~15



Model Selection

- **Vision and Language Transformer (ViLT) (Kim et al., 2021)**
  - Four image-text datasets: COCO, Visual Genome, Conceptual Captions, and SBU Captions
  - Two pre-training tasks:
    - Image Text Matching
    - Masked Language Modeling

Visual Embed	Model	#Params (M)	#FLOPS (G)	Time (ms)
Region	ViLBERT	274.3	958.1	~900
	UNITER	154.7	949.9	~900
Linear	ViLT	87.4	55.9	~15



https://proceedings.mlr.press/v139/kim21k

Proposed Method

- **Dataset**
  - MIRACL-VC1 (Rekik et al., 2014)
- **Model**
  - ViLT (Kim et al., 2021)
- **Evaluation**
  - For each word/phrase per speaker, 8 for training 2 for testing
  - Baseline: encoder decoder approach (CNN + LSTM)
  - New video instances captured by a smartphone

Proposed Method

- Dataset
  - MIRACL-VC1 (Rekik et al., 2014)
- Model
  - ViLT (Kim et al., 2021)
- Evaluation
  - For each word/phrase per speaker, 8 for training 2 for testing
  - Baseline: encoder decoder approach (CNN + LSTM)
  - New video instances captured by a smartphone

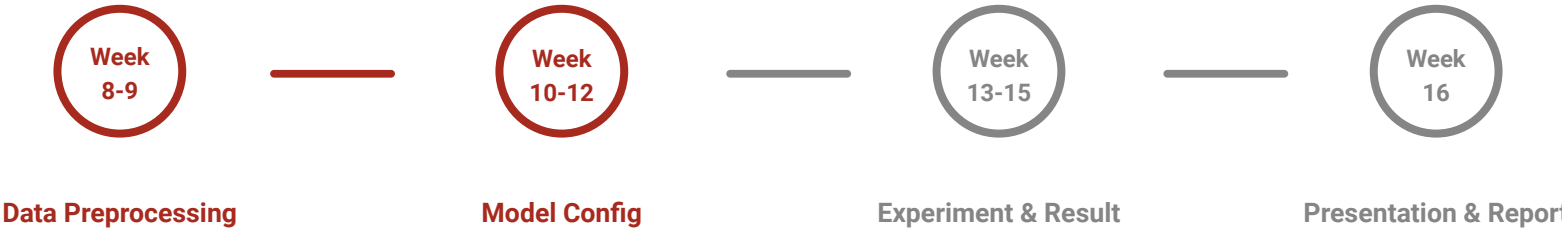
Equipment List

Equipment List

- PC
- Smartphone
- Google Colab

Time Plan

Time Plan


Week 8-9
Week 10-12
Week 13-15
Week 16
Data Preprocessing
Model Config
Experiment & Result
Presentation & Report

Thank You! Questions?

References

- Ahmed Rekik, Achraf Ben-Hamadou, and Walid Mahdi. 2014. A New Visual Speech Recognition Approach for RGB-D Cameras. In *Image Analysis and Recognition*, Aurélio Campilho and Mohamed Kamel (eds.). Springer International Publishing, Cham, 21–28. DOI:https://doi.org/10.1007/978-3-319-11755-3_3
- Hong Liu, Wenhao Li, and Bing Yang. 2021. Robust Audio-Visual Speech Recognition Based on Hybrid Fusion. In *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, Milan, Italy, 7580–7586. DOI:https://doi.org/10.1109/ICPR48806.2021.9412817
- L Ashok Kumar, D Karthika Renuka, S Lovelyn Rose, M C Shunmuga priya, and I Made Wartana. 2022. Deep learning based assistive technology on audio visual speech recognition for hearing impaired. *International Journal of Cognitive Computing in Engineering* 3, (June 2022), 24–30. DOI:https://doi.org/10.1016/j.ijcce.2022.01.003
- Mingfeng Hao, Mutallip Mamut, Nurbiya Yadikar, Alimjan Aysa, and Kurban Ubul. 2020. A Survey of Research on Lipreading Technology. *IEEE Access* 8, (2020), 204518–204544. DOI:https://doi.org/10.1109/ACCESS.2020.3036865
- Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2021. End-To-End Audio-Visual Speech Recognition with Conformers. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Toronto, ON, Canada, 7613–7617. DOI:https://doi.org/10.1109/ICASSP39728.2021.9414567
- Shinnosuke Isobe, Satoshi Tamura, Satoru Hayamizu, Yuuto Gotoh, and Masaki Nose. 2021. Multi-Angle Lipreading with Angle Classification-Based Feature Extraction and Its Application to Audio-Visual Speech Recognition. *Future Internet* 13, 7 (July 2021), 182. DOI:https://doi.org/10.3390/fi13070182
- Wentao Yu, Steffen Zeiler, and Dorothea Kolossa. 2021. Fusing Information Streams in End-to-End Audio-Visual Speech Recognition. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Toronto, ON, Canada, 3430–3434. DOI:https://doi.org/10.1109/ICASSP39728.2021.9414553
- Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *Proceedings of the 38th International Conference on Machine Learning* (Proceedings of Machine Learning Research), PMLR, 5583–5594. Retrieved from https://proceedings.mlr.press/v139/kim21k.html