Visual Speech Recognition summary

V. Tosounidis, E. Bikas, Prof. Ioannis Pitas
Aristotle University of Thessaloniki
pitas@csd.auth.gr
www.aiia.csd.auth.gr

Version 4.0
Date: 18/01/2021

Artificial Intelligence & Information Analysis Lab

Visual Speech Recognition

- Visual Speech Recognition
  - Visemes and Phonemes
  - Face detection
  - Landmark Localization
  - Lip reading
  - Speech reading beyond the lips
- Audio-Visual Speech Recognition
  - Deep Audio-Visual Speech Recognition
    - Convolutional Neural Networks
    - Recurrent Neural Networks
  - Overlapped speech
  - Speaker targeted AVSR models
- Visual Speech Recognition for mobile devices
- Visual Speech Recognition DataSets
- Experiments on each data set

Visual Speech Recognition

Visual Speech Recognition **(VSR)** is a method of recognizing spoken language by exploiting only visual signals that are observed during speech.

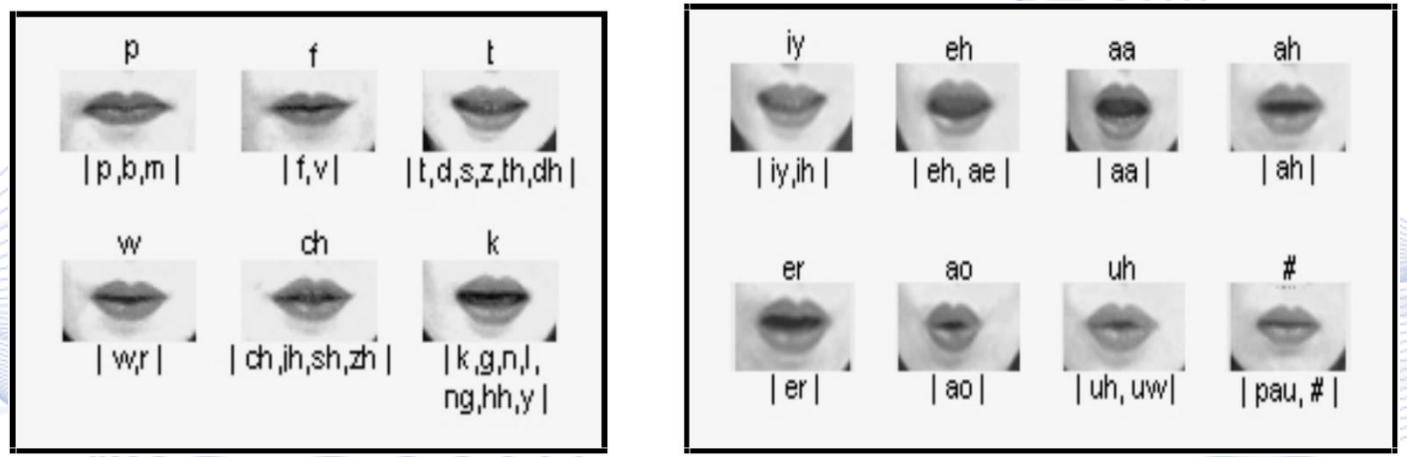
Methods involved:

- Object detection (face detection, lip localization)
- Image processing (feature extraction)
- Pattern Recognition (lip reading)

Visemes and Phonemes Classification

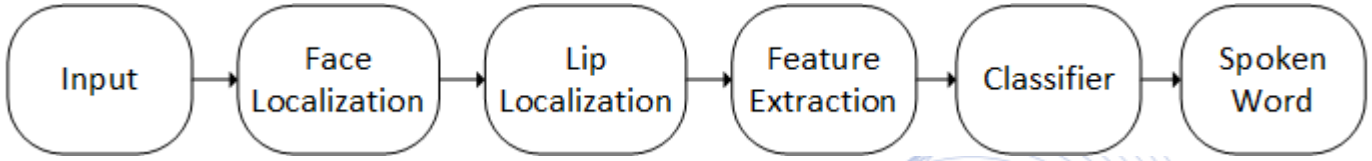
Phonemes are the smallest distinguishable units of sound that can carry meaning. Collectively phonemes make up words.

Visemes are the visual representation of phonemes.



Example of viseme map used in [LEE2002]

Typical VSR System



Flow of a typical Visual Speech Recognition System

Face Detection

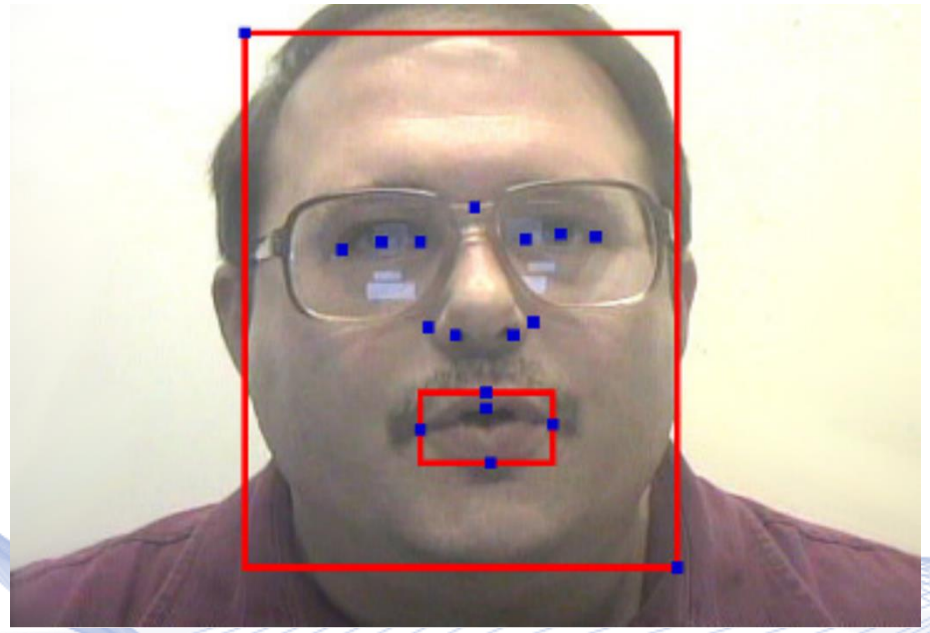
Definition:

Given photos or video sequences is intended to find or identify one or more people using an existing database of persons.

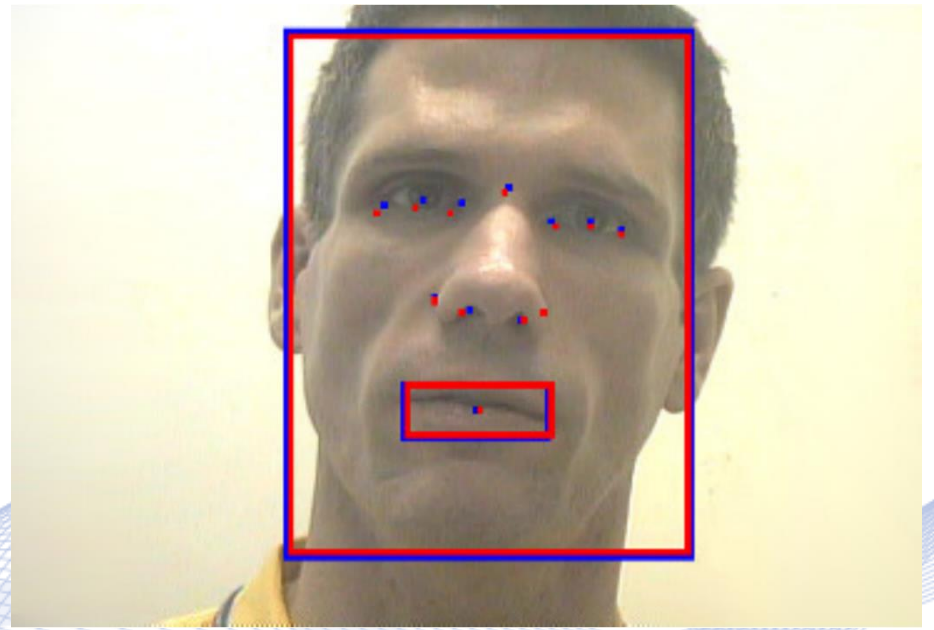
Procedure in VSR:

Isolating the face from the environment, finding the mouth Region of Interest (RoI) and working on it.

Face detection and landmark localization



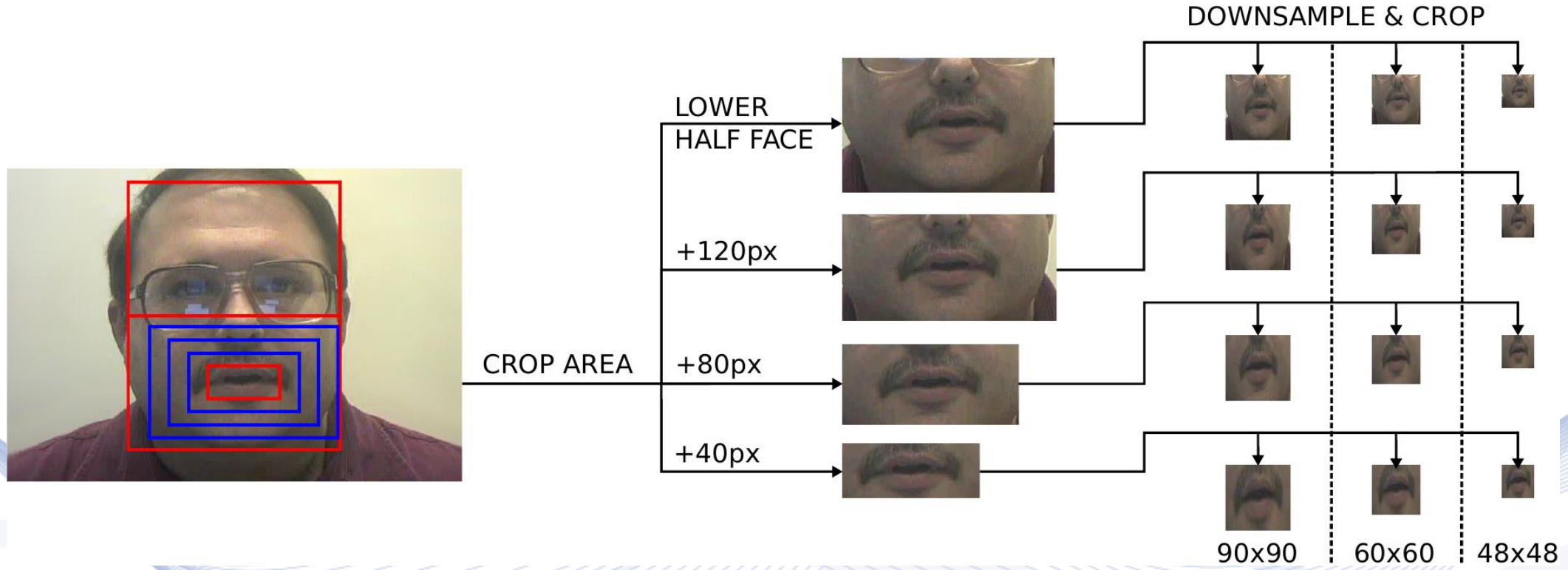
(a)



(b)

(a) Blue shows the points predicted and with red are the bounding boxes (b) Red points and bounding boxes are the predicted values and blue are the ground truth [KOU2017].

ROI Extraction



ROI extraction process following face and mouth detection, and split-ting of different ROIs with variant size [KOU2017].

Lip Localization

Definition:

Localization of the external contours of the lips on the given image or on the first image of the video sequence (the mouth is detected and localized by similar technique to eye detection).

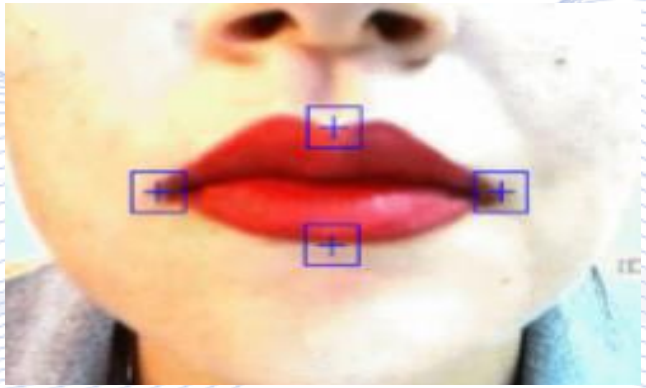


Image: The Different point of interest for the extraction stapes (from [LL]).

Lip Localization

The active contour based method:

Consists of placing an initial line around a shape of contour. This line deforms itself progressively according to the action of several strengths that push it toward the shape.

Image: contour method (from [LL]).

Lip Localization

Final result:

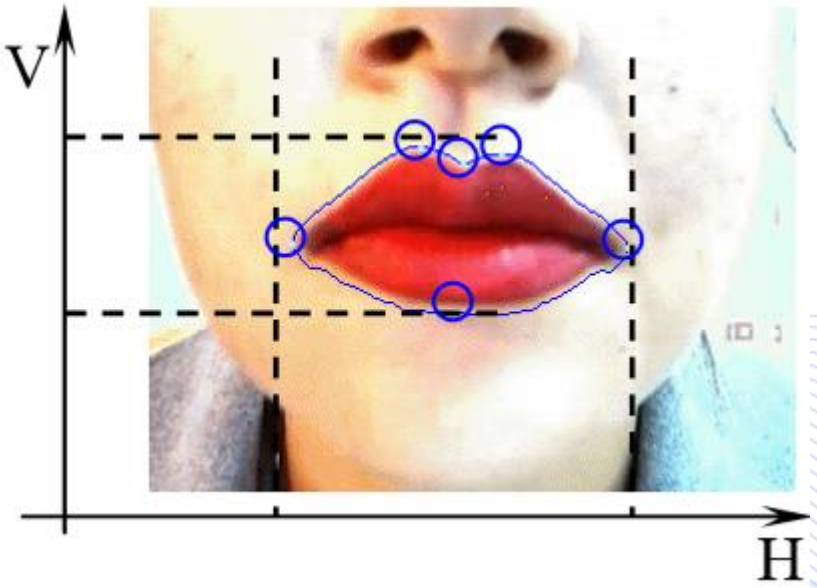


Image: Points of interest detection by the projection of final contour on horizontal and vertical axis (from [LL]).

Lip Reading

Definition:

Recognizing the content of speech based on observing the speaker's lip movements.

Approach 1

Problem statement:

Detecting video frames where a person speaks using only visual information.

- Input: A mouth RoI produced by mouth detection.
- Output: yes/no visual speech indication.

Lip Reading

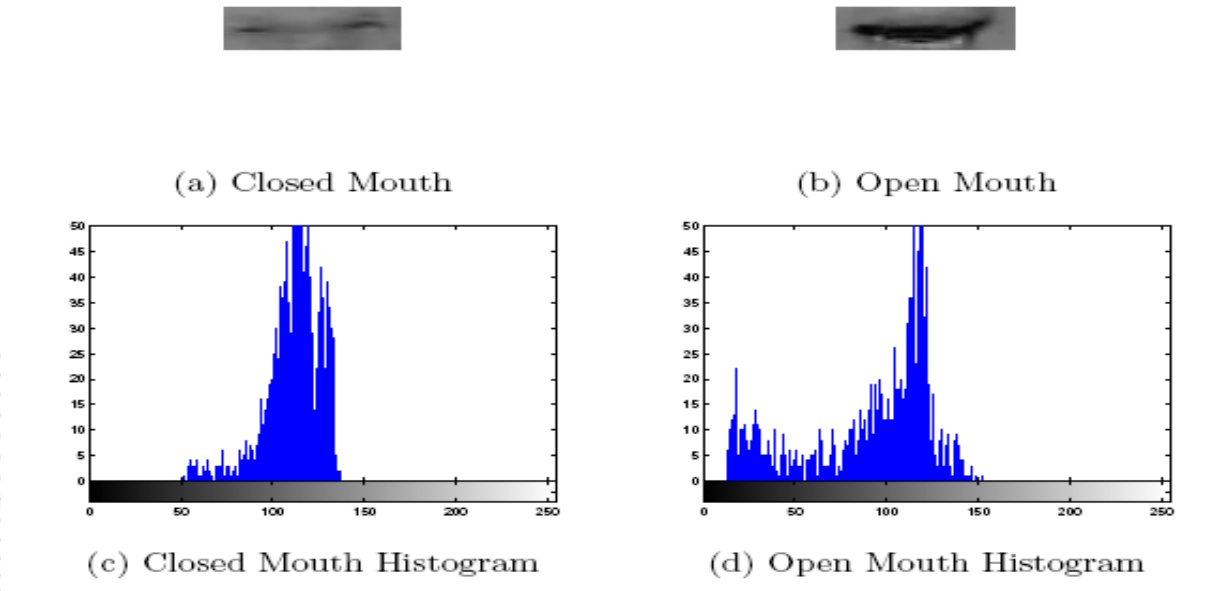


Image: Increase in the number of low intensity pixels, in the mouth region, when mouth is open (from [IP]).

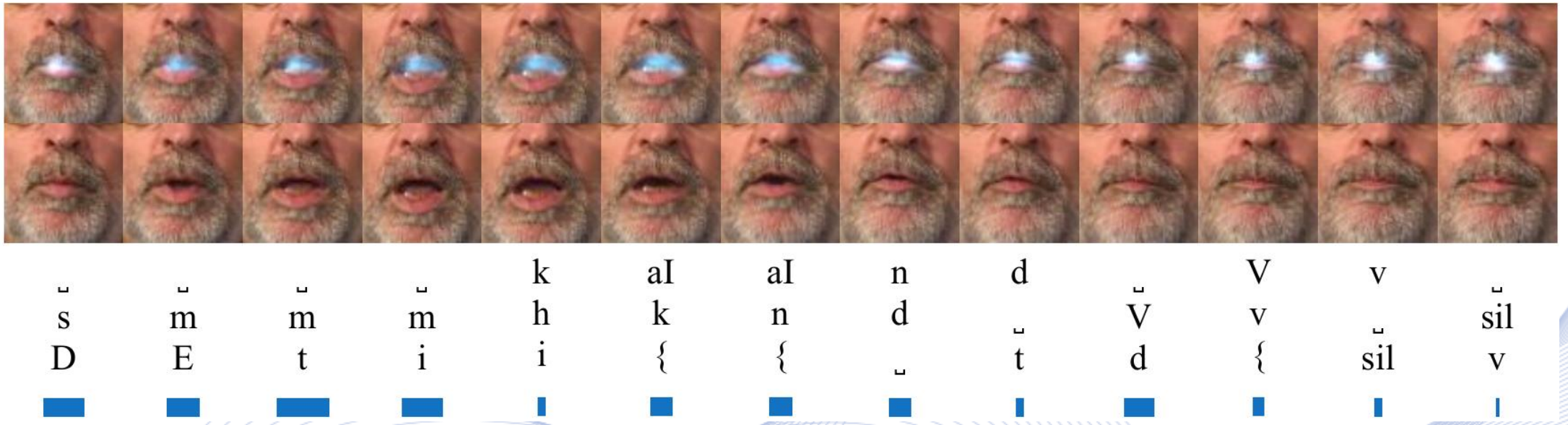
Lip Reading

Approach 2

Action recognition approach (VSD in the wild):

- Input: random videos.
- Step 1: facial ROI determination by face detection.
- Step 2:
  - Extraction of spatiotemporal locations of interest using STIPs and Dense Trajectories methods,
  - description of the detected interest points,
  - facial video representation using Bag of Visual Words model (BoW).
- Step 3: KSVM, KELM classification
- Output: video-based visual speech/silence indication (can also be frame-based after input video preprocessing)

Lip Reading



Saliency map for “kind of” and the top-3 predictions of each frame. [SHILL2018]

Overlapped Speech

Definition:

Speech recognition when more than one speaker are talking.

Problems:

- Overlapped Speech Separation
- Overlapped Speech Recognition.

Speech Reading Beyond the Lips

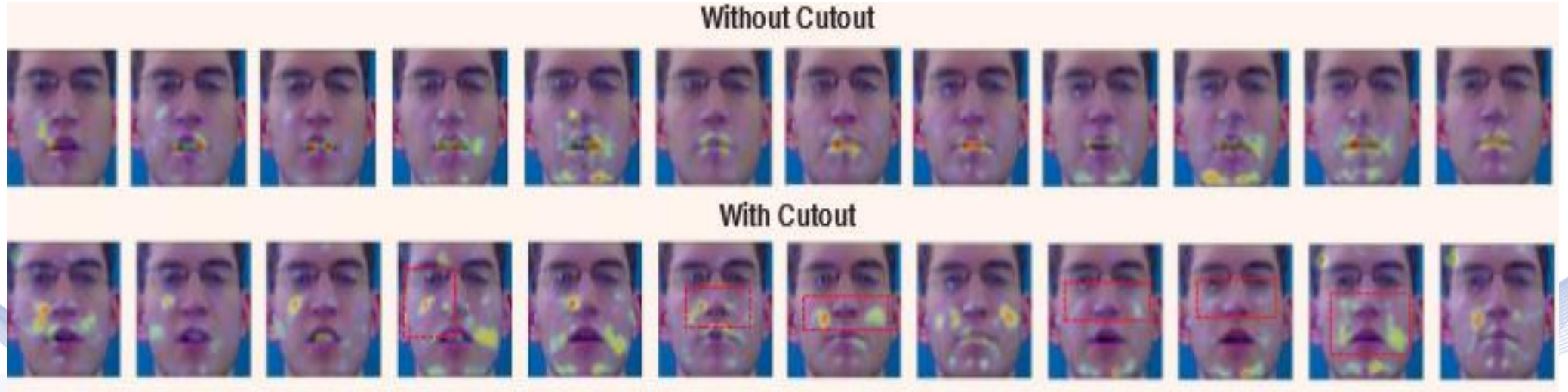


Image: Visualizations of saliency map technique behavior with Cutout applied (from [SBTL]).

Audio-Visual Speech Recognition

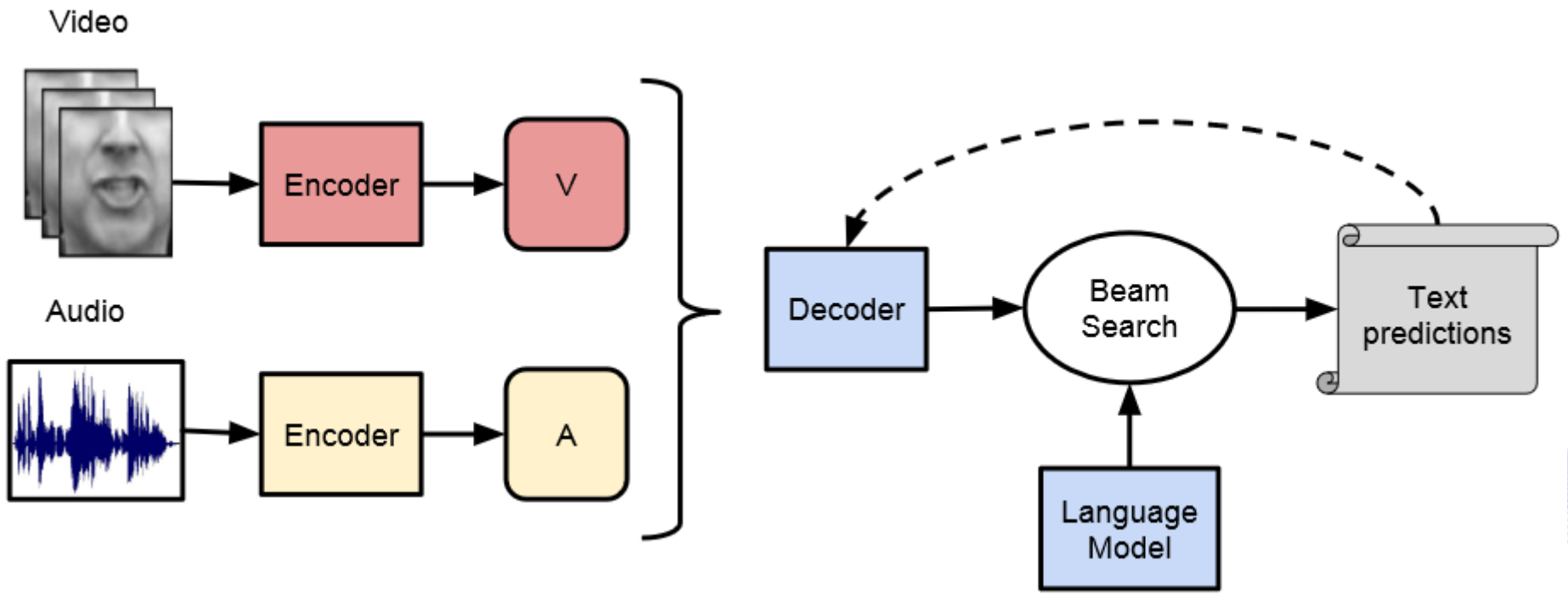
Definition:

AVSR is a process of recognizing the speech, using both audio signal and visual information.

Why do we need the visual information?

Sometimes the acoustic signal alone can confuse us, especially in noisy environments (phenomena such as the illusions of McGurk or the “Cocktail party” effect show the importance of visual information in speech perception).

Basic A-VSR pipeline



Audio-visual speech recognition pipeline [AFO2018]

Deep AVSR

VSR neural network training models:

- Sequence to sequence (seq2seq)

A Seq2Seq model is a model that takes a sequence of items (words, letters, time series, etc.) and outputs another sequence of items.

- Connectionist Temporal Classification (CTC)

it is a type of neural network output and associated scoring function, for training Recurrent Neural Networks (RNNs) to tackle sequence problems where the timing is variable.

Convolutional Neural Networks

VSR system overview:

For visual feature extraction, a seven-layered CNN is utilized to recognize phonemes from the mouth area image sequences.

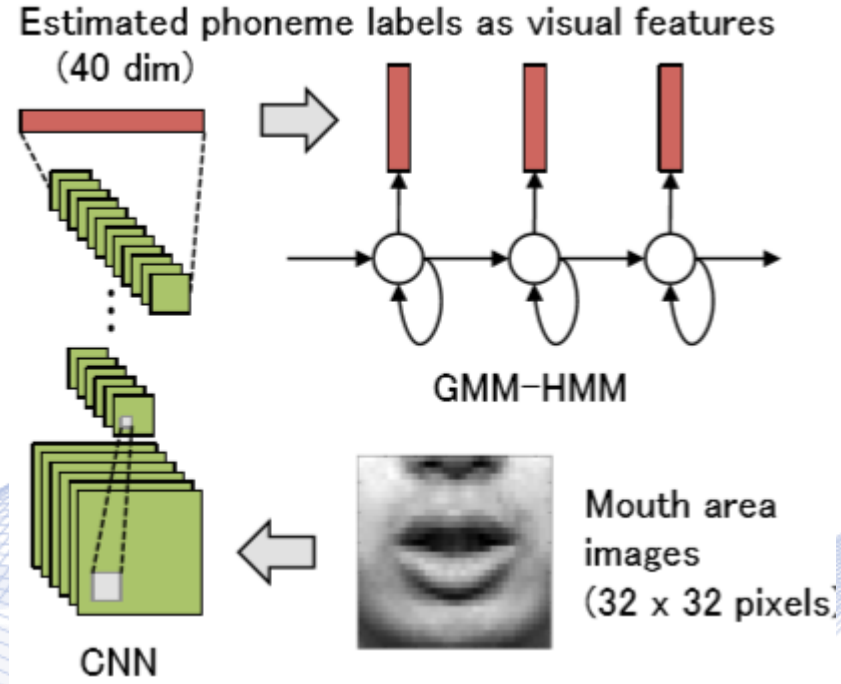


Image: Architecture of VSR CNN system (from [CNN]).

Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a class of neural networks used to map sequences to sequences.

They have feedback loops at each cell so they have temporal memory.

RNN types:

- Vanilla RNN,
- Long Short-Term Memory (LSTM),
- Gated Recurrent Units (GRU).

Recurrent Neural Networks

Fusion Models:

(a) Feature Fusion:

In feature fusion technique, a single RNN is trained by concatenating the audio and visual features using the CTC objective function.

(b) Decision Fusion:

In decision fusion technique the audio and visual modalities are modeled by separate networks.

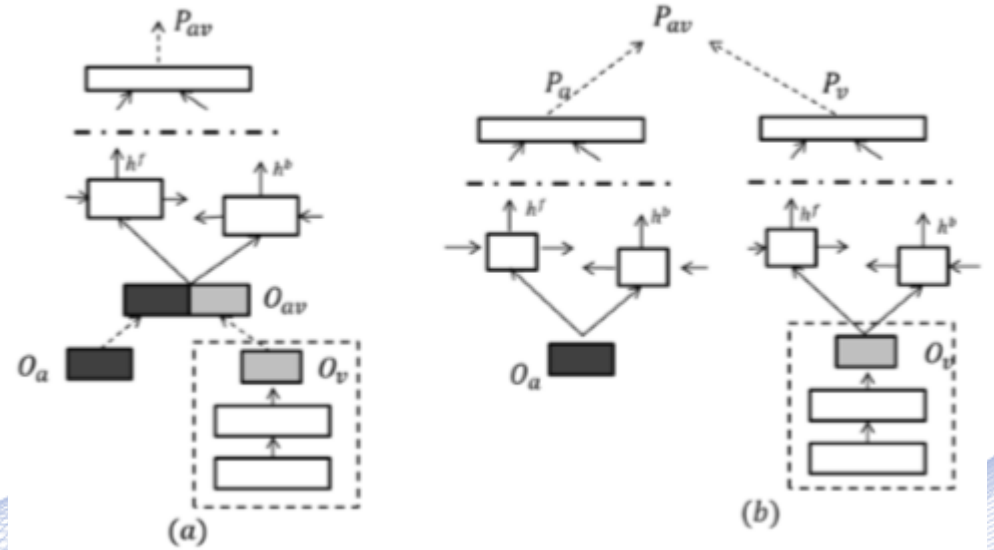


Image: AVSR using Deep RNN (from [RNN]).

Overlapped Speech

AVSR systems for overlapped speech:

- Pipelined
- Integrated.

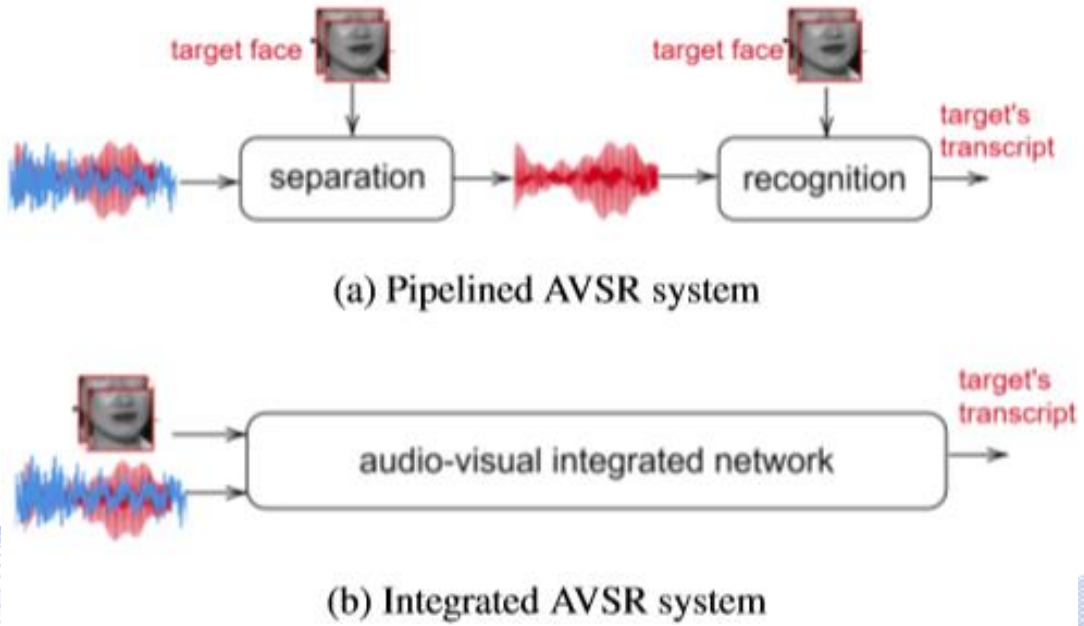


Image: Illustration of pipelined and integrated AVSR systems (from [AVOS]).

Pipelined AVSR system

Input:

An overlapped audio signal, the target speaker's mouth RoI.

Steps:

- The system uses the visual information to directly estimate the TF mask of the target speaker.
- The spectrogram of the separated audio is obtained by multiplying the TF mask with the overlapped spectrogram.

Output:

Target's transcript.

Integrated AVSR system

Input:

An overlapped audio signal, the target speaker's mouth RoI.

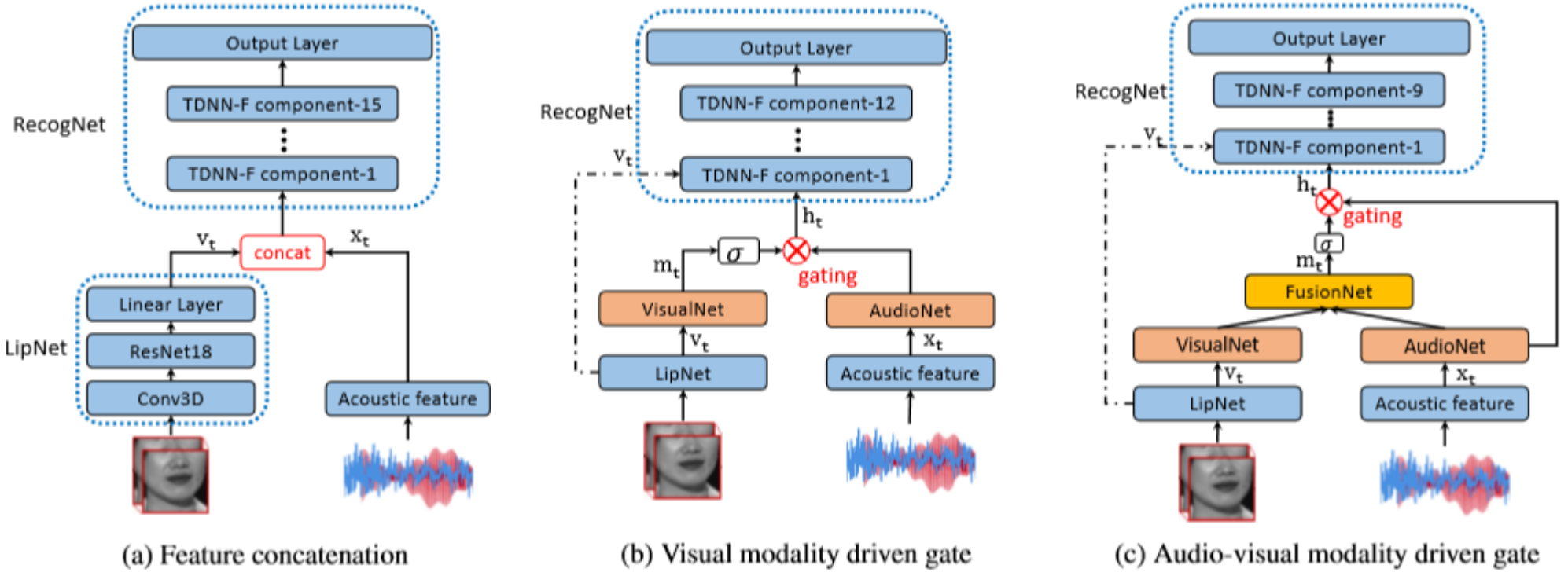
Then:

Integrated tries to implicitly do both separation and recognition in a compact model architecture using a single recognition cost function.

Output:

Target's transcript.

Integrated AVSR system



(a) Feature concatenation (b) Visual modality driven gate (c) Audio-visual modality driven gate

Image: Illustration of audio-visual fusion methods for AVSR systems (from [AVOS]).

Speaker-Targeted AVSR Models

Speaker-Independent Models:

They use the visual information in conjunction with the acoustic features.

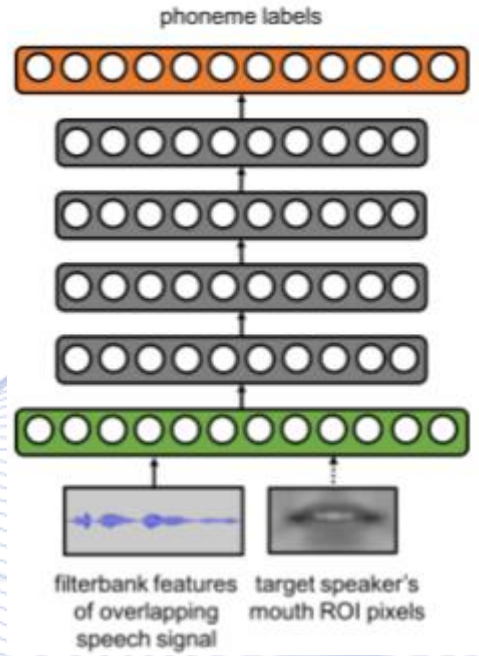


Image: Speaker-Independent Models (from [AVST]).

Speaker-Targeted AVSR Models

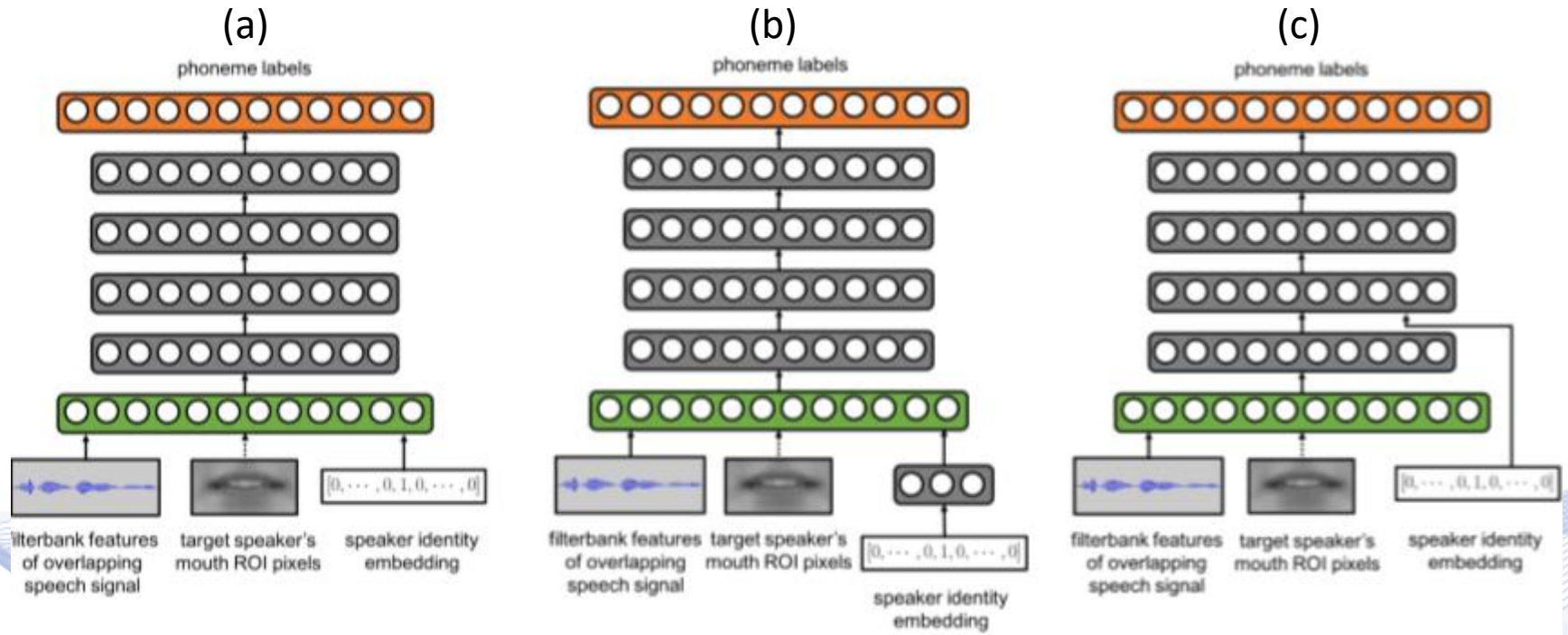


Image: Three variants of speaker-targeted models (from [AVST]).

VSR for mobile devices

The MobiVSR architecture addresses the problem of deploying visual speech recognition models on resource constrained devices, it has 6× fewer parameters and 20× smaller model size than the state-of-the-art(*) model.

Architecture:

The MobiVSR model essentially maps visemes (basic units of visual speech) to textual units (i.e., characters/words).

(*) State-of-the-art deep learning stands for any deep neural network that performs as good as human performance for specific task.

VSR for mobile devices

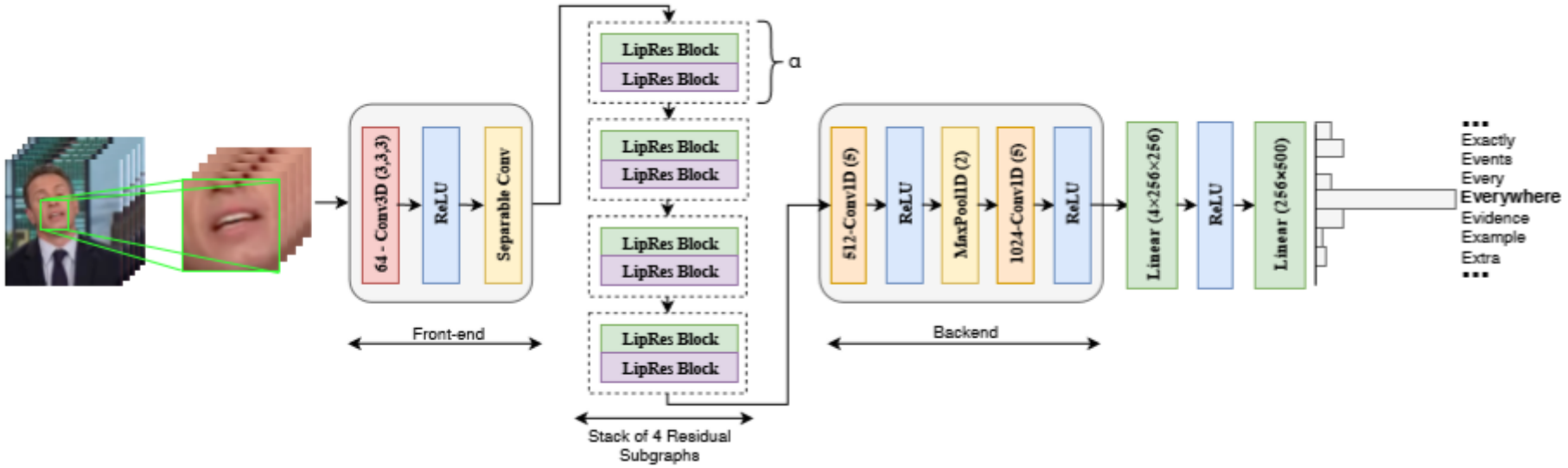


Image: MobiVSR architecture (from [MobiVSR]).

Popular VSR Datasets

Name	Comments	Speakers
LRW (BBC)	Up to 1000 utterances of 500 different words	
LRS2 (BBC)	1000s of natural sentences from the British television	
LRS3 (TED)	1000s of natural sentences from TED and TEDx videos	
DAVID	Design issues for a digital A-V integrated database.	124
XM2VTS	The extended M2VTS database.	295
CUVAE	Moving-Talker, speaker-independent data set.	36
GRID	Short and simple phrase utterances.	34
OuluVS	Lipreading with local spatiotemporal descriptors.	20
OuluVS2	Multi-view A-V database for non-rigid mouth motion analysis.	53

LRW Dataset Experiments

Method	Accuracy
Multi-Scale TCN [MART2020]	85.3%
Two-Stream Deep 3d [WENG2019]	84.1%
ResNet + BGRU [PET2018]	83.4%

Bibliography

[PIT2021] I. Pitas, “Computer vision”, Createspace/Amazon, in press.
[PIT2017] I. Pitas, “Digital video processing and analysis” , China Machine Press, 2017 (in Chinese).
[PIT2013] I. Pitas, “Digital Video and Television” , Createspace/Amazon, 2013.
[NIK2000] N. Nikolaidis and I. Pitas, “3D Image Processing Algorithms”, J. Wiley, 2000.
[PIT2000] I. Pitas, “Digital Image Processing Algorithms and Applications”, J. Wiley, 2000.

Q & A

Thank you very much for your attention!

More material in
http://icarus.csd.auth.gr/cvml-web-lecture-series/

Contact: Prof. I. Pitas
pitas@csd.auth.gr