Is Someone Speaking?

Exploring Long-term Temporal Features for Audio-visual Active Speaker Detection

(ACM MM 2021)

Authors:Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, Haizhou Li

陈仁苗

2021/10/20

Active Speaker Detection(ASD)

detect who is speaking in a visual scene of one or more speakers

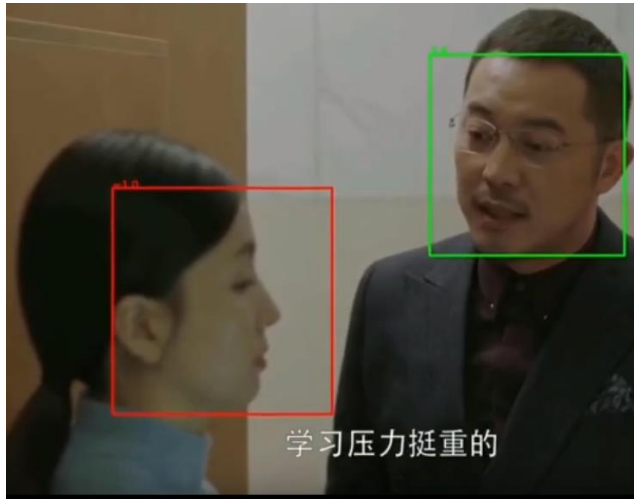
Input:

The cropped face video and corresponding audio

Output:

If the person is speaking in each video frame

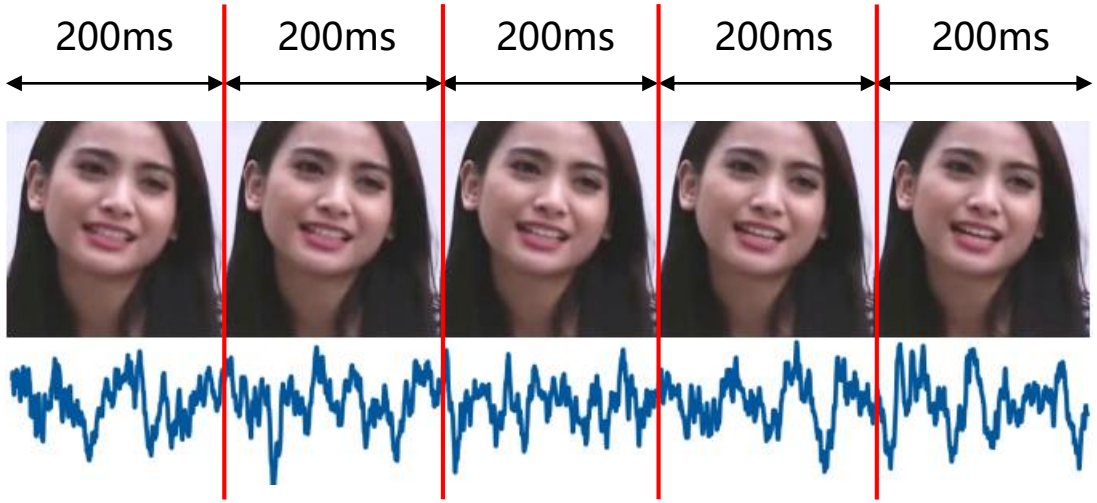
Audio-visual synchronization?



Green box: Active speaker

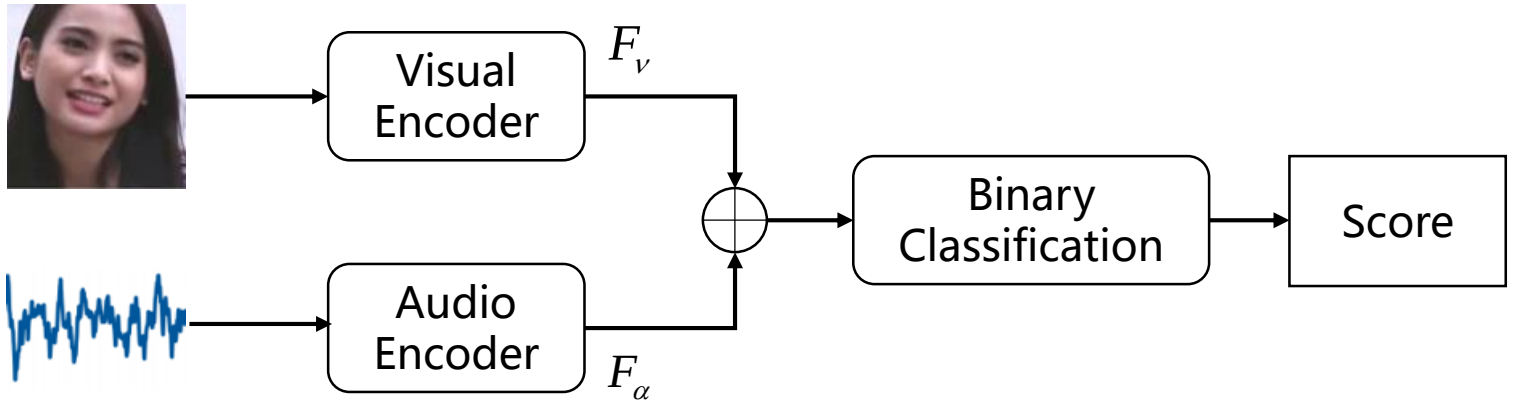
Red box: Inactive speaker

Traditional method



Steps:

1) Split video into short segements
2) Predict score
3) Smooth scores



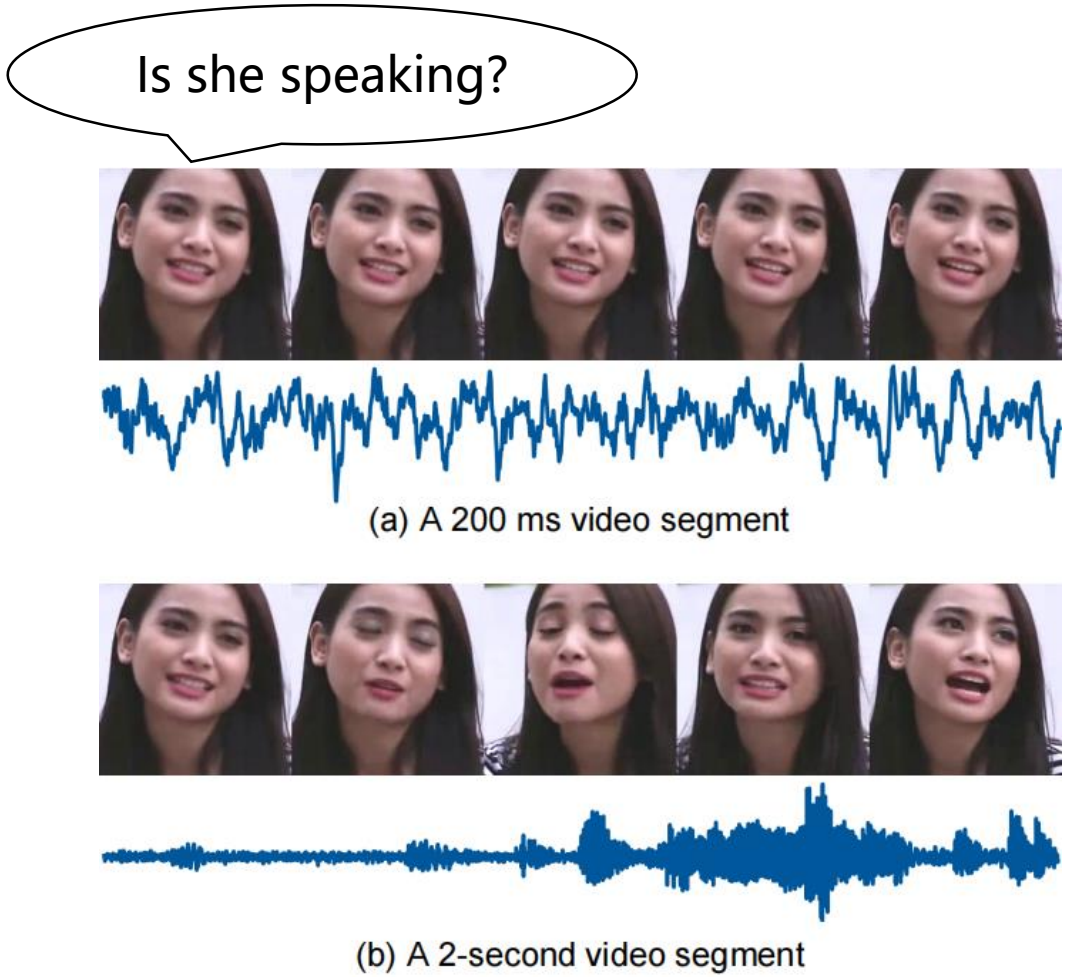
Problem in traditional method

Can we learn meaningful ASD information from such a short segememt?

a) A **200 ms** video segement, where the speaking activity is not evident.

b) A **2-second** video segement, where the speaking activity becomes evident in long-term temporal context and audio-visual synchronization.

Challenge: If we simply extend the length of the splitted segments, the extract feature will not be fine granularity, which will drop the performance.

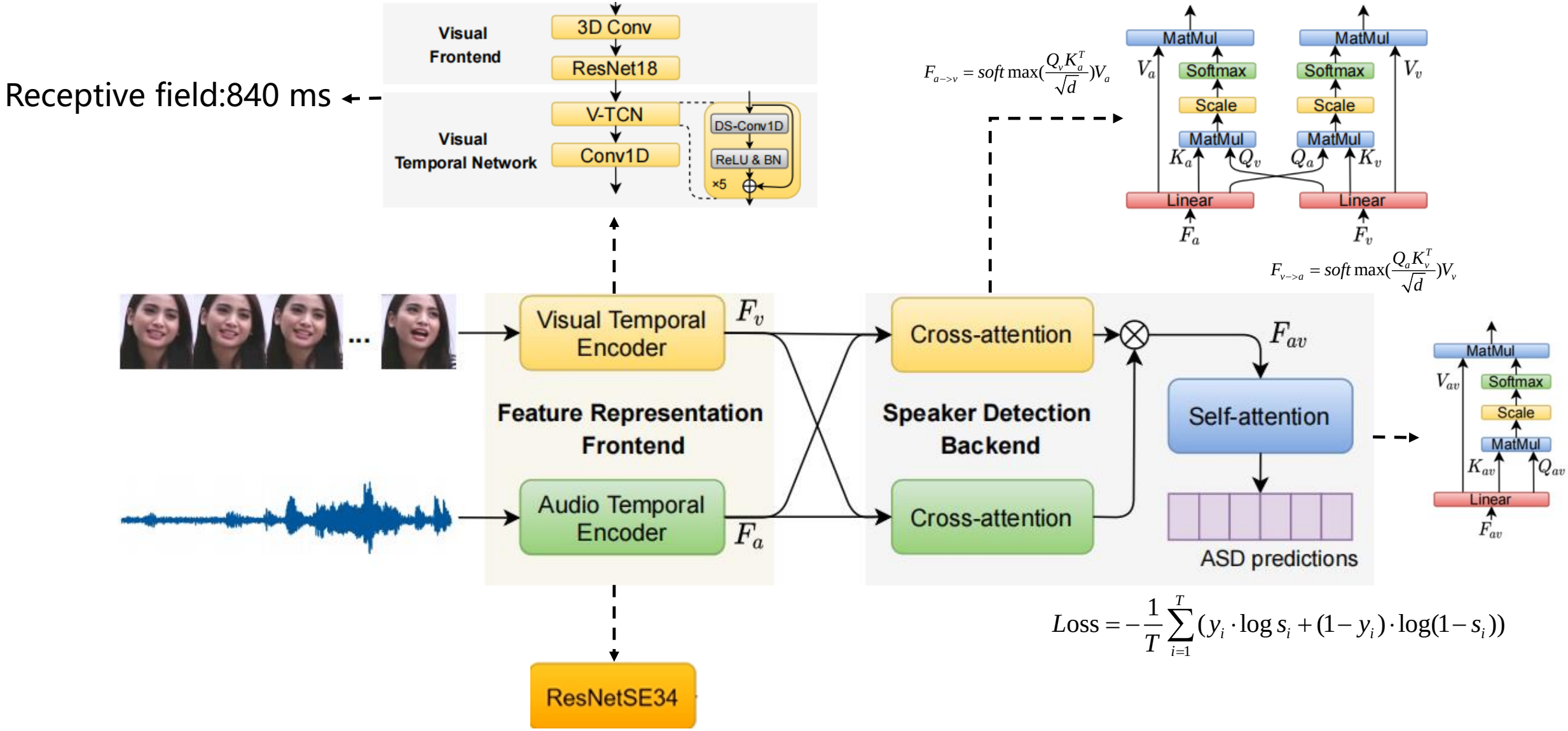


(a) A 200 ms video segment

(b) A 2-second video segment

TalkNet: Exploring long-term feature Active Speaker Detection

Receptive field:840 ms



$$Loss = -\frac{1}{T}\sum_{i=1}^{T}(y_i \cdot \log s_i + (1-y_i)\cdot \log(1-s_i))$$

Experimental Results: AVA-ActiveSpeaker Benchmark

Table 2: Comparison with the state-of-the-art on the AVA-ActiveSpeaker validation set in terms of mean average precision (mAP).

Method	mAP (%)
Roth et al. [26, 35]	79.2
Zhang et al. [50]	84.0
MAAS-LAN [26]	85.1
Alcazar et al. [5]	87.1
Chung et al [12]	87.8
MAAS-TAN [26]	88.8
TalkNet (proposed)	**92.3**

Table 4: Comparison with the state-of-the-art on the AVA-ActiveSpeaker test set in terms of mAP.

Method	mAP (%)
Roth et al. [35]	82.1
Zhang et al. [50]	83.5
Alcazar et al. [5]	86.7
Chung et al. [12]	87.8
TalkNet (proposed)	**90.8**

got **3.5%** and **3.0%** mAP improvement than SOTA in AVA-ActiveSpeaker validation set and test set, respectively.

Experimental Results: Columbia ASD Benchmark

Table 5: Comparison with the state-of-the-art on the Columbia ASD dataset in terms of F1 scores (%).

Method	Speaker					
	Bell	Boll	Lieb	Long	Sick	Avg.
Brox et al. [8, 37]	84.1	72.3	80.6	60.0	68.9	73.2
Chakravarty et al. [10]	82.9	65.8	73.6	86.9	81.8	78.2
Zach et al. [37, 48]	89.2	88.8	85.8	81.4	86.0	86.2
RGB-DI [37]	86.3	93.8	92.3	76.1	86.3	87.0
SyncNet [17]	93.7	83.4	86.8	**97.7**	86.1	89.5
LWTNet [4]	92.6	82.4	88.7	94.4	95.9	90.8
RealVAD [7]	92.0	**98.9**	94.1	89.1	92.8	93.4
S-VVAD [38]	92.4	97.2	92.3	95.5	92.5	94.0
TalkNet (proposed)	**97.1**	90.0	**99.1**	96.6	**98.1**	**96.2**

got **2.2%** F1 improvement in ColumbiaASD dataset than SOTA.

Ablation Study

Table 6: Performance evaluation by the length of the video on the AVA-ActiveSpeaker validation set. We use a fixed number of video frames during both training and testing.

# video frames	Length (APRX seconds)	mAP(%)
5	0.2	75.2
10	0.4	82.8
25	1	87.9
50	2	89.5
100	4	89.4
Variable	**1 - 10**	**92.3**

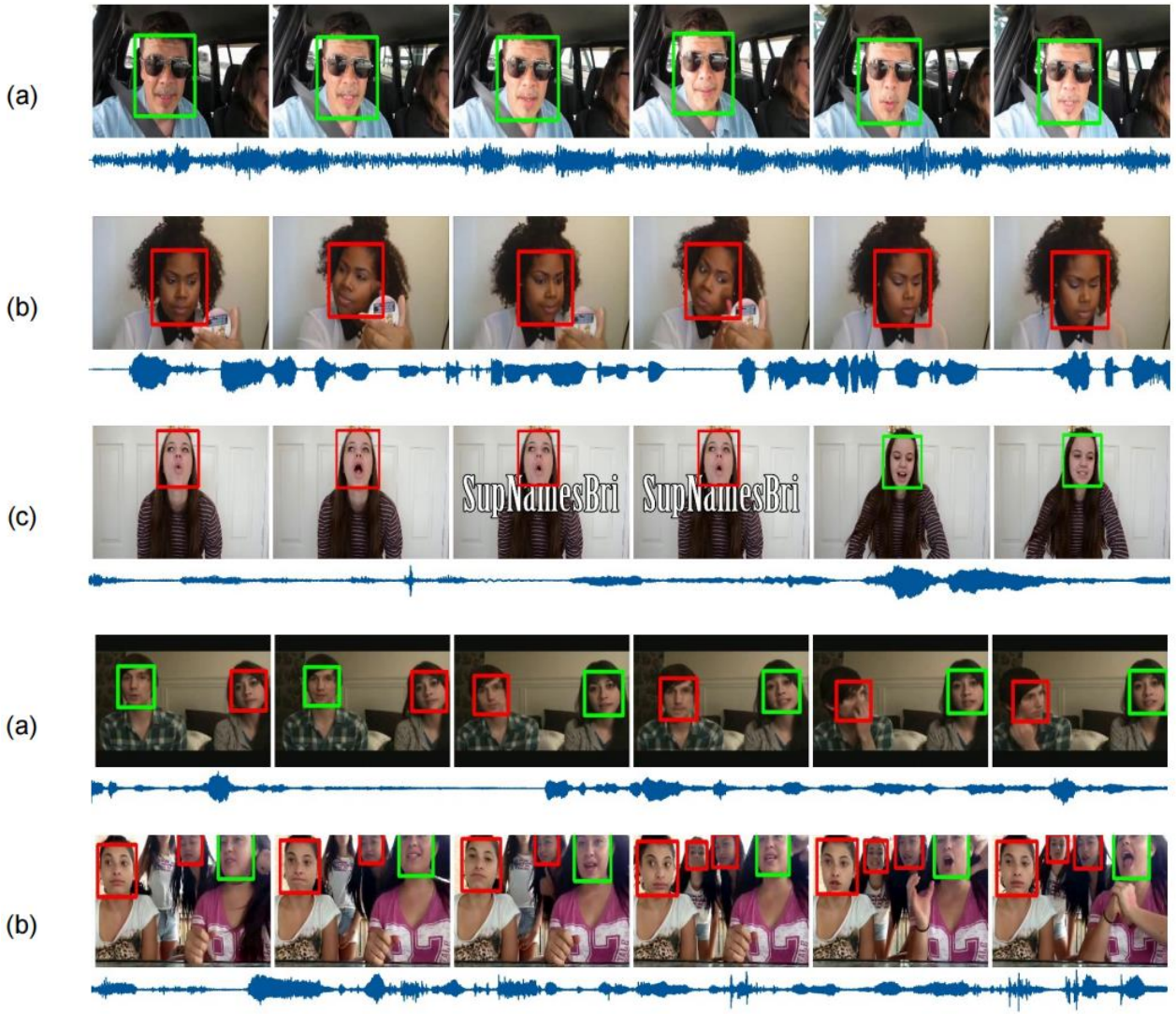
Table 7: A contrastive study between two systems on efficient use of long video segments on the AVA-ActiveSpeaker validation set in terms of mAP (%).

Method	# video frames		Change
	11	25	
Alcazar et al. [5]	78.2	76.1	−2.1
TalkNet (Proposed)	83.1	87.9	+4.8

Table 8: Ablation study of the cross-attention and self-attention mechanisms in TalkNet on the AVA-ActiveSpeaker validation set.

Model	mAP(%)
w/o Both	90.0
w/o Self-attention	90.9
w/o Cross-attention	91.6
TalkNet	**92.3**

Experimental Results: AVA-ActiveSpeaker Benchmark



The man is speaking in the noisy environment.

The woman is introducing the makeup process through dubbing. So the speech is not synchronized with her lip movement.

The woman is eating candy. Although her lips are always moving, she is not speaking in the beginning.

Two speakers take turns speaking, and the man’ s lips are concealed sometimes.

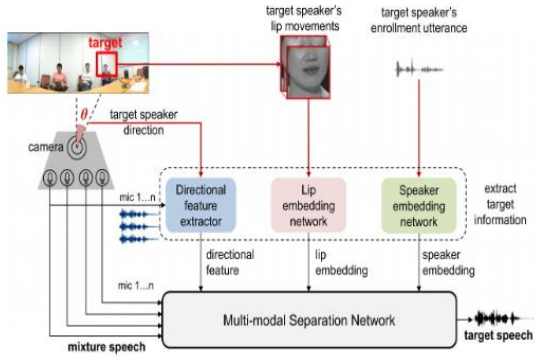
Four speakers are talking in a boisterous environment with background music. Everyone’ s mouth is moving, but only the girl on the right is speaking.

Feature work(maybe)

- Improve the anti-attack ability of the multi-modal speaker recognition system

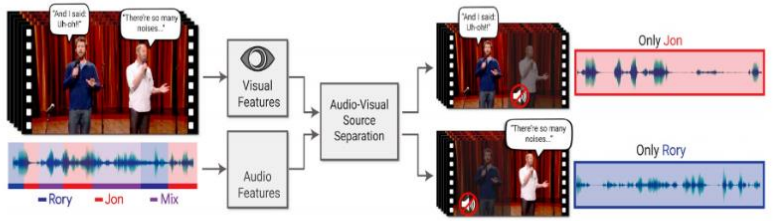
- `Improve the performance of speech separation tasks`

Visual-audio speaker separation



Gu et al., *Multi-modal Multi-channel Target Speech Separation.* IEEE Journal of selected topics in signal processing, 2020.

Visual-audio speaker separation



Ephrat A et al., *Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation*. ACM Transactions on Graphics, 2018.

- `Improve the performance of speech denoising tasks`