JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

A Light Weight Model for Active Speaker Detection

Junhua Liao[1], Haihan Duan[2,3], Kanghui Feng[1], Wanbing Zhao[1], Yanbing Yang[1,3], Liangyin Chen[1,3]

1. College of Computer Science, Sichuan University, Chengdu, China

2. The Chinese University of Hong Kong, Shenzhen, China

3.The Institute for Industrial Internet Research, Sichuan University, Chengdu, China

CVPR 2023   THU-PM-222

四川大學 SICHUAN UNIVERSITY

香港中文大學（深圳）The Chinese University of Hong Kong, Shenzhen

Highlights

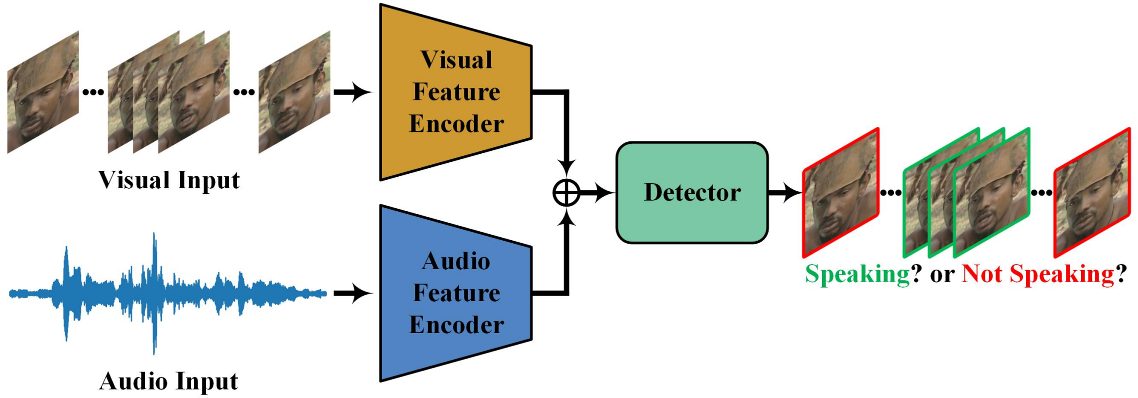


Figure 1. Overview of the proposed framework.

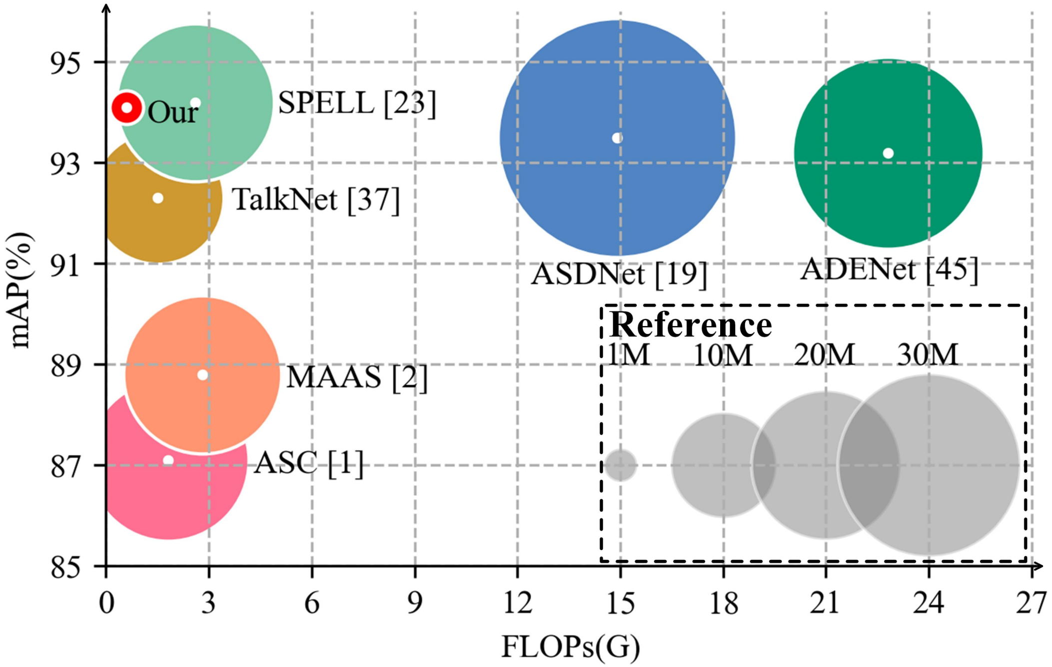


Figure 2. mAP vs. FLOPs, size ∝ parameters.

Contributions

- A lightweight design is developed from the three aspects of information input, feature extraction, and cross-modal modeling; subsequently, a lightweight and effective end-to-end active speaker detection framework is proposed. In addition, a novel loss function is designed for training.

- Experiments on AVA-ActiveSpeaker, a benchmark dataset for active speaker detection released by Google, reveal that the proposed method is comparable to the state-of-the-art method, while still reducing model parameters by 95.6% and FLOPs by 76.9%.

- Ablation studies, cross-dataset testing, and qualitative analysis demonstrate the state-of-the-art performance and good robustness of the proposed method.

Encoders

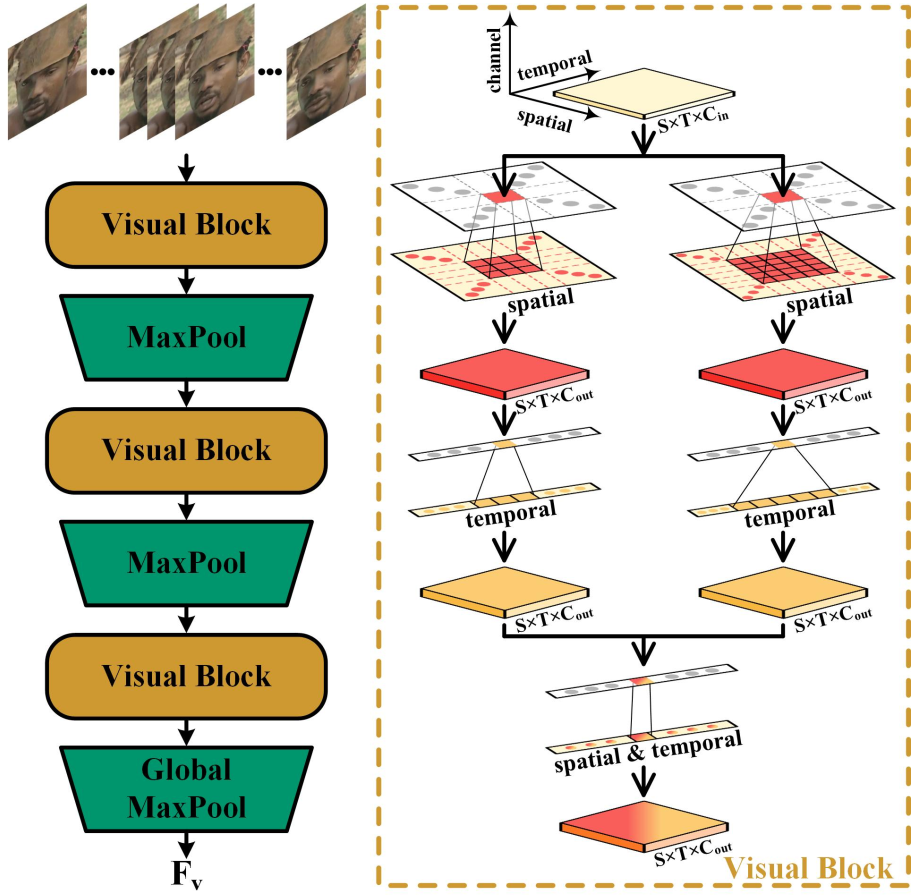


Figure 3. The architecture of visual feature encoder.

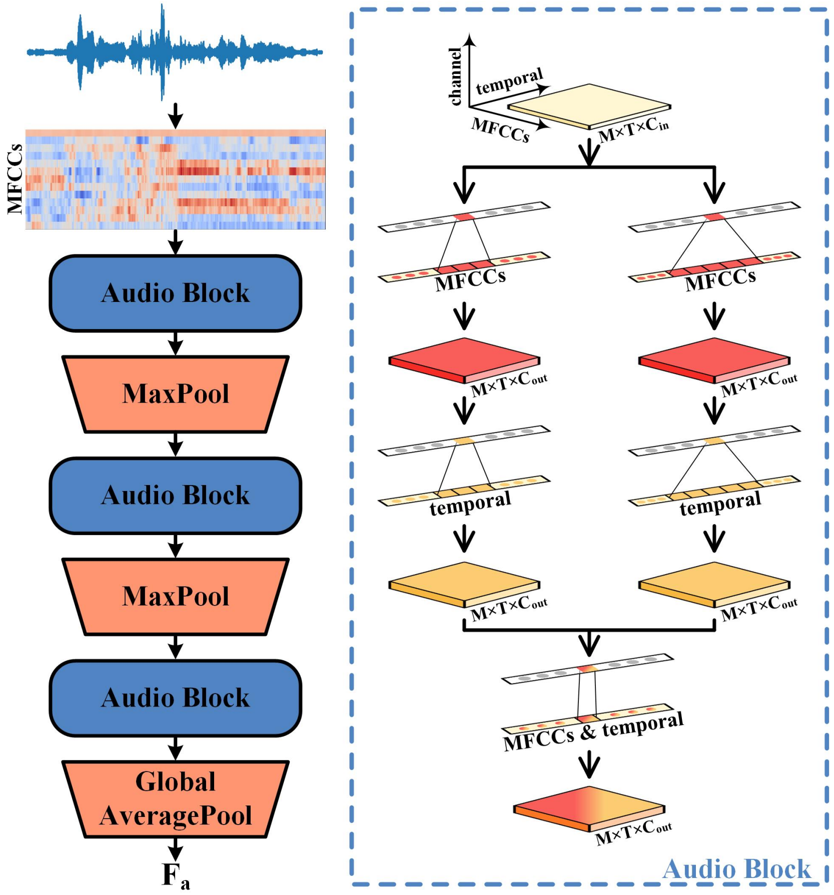


Figure 4. The architecture of the audio feature encoder.

Detector & Loss

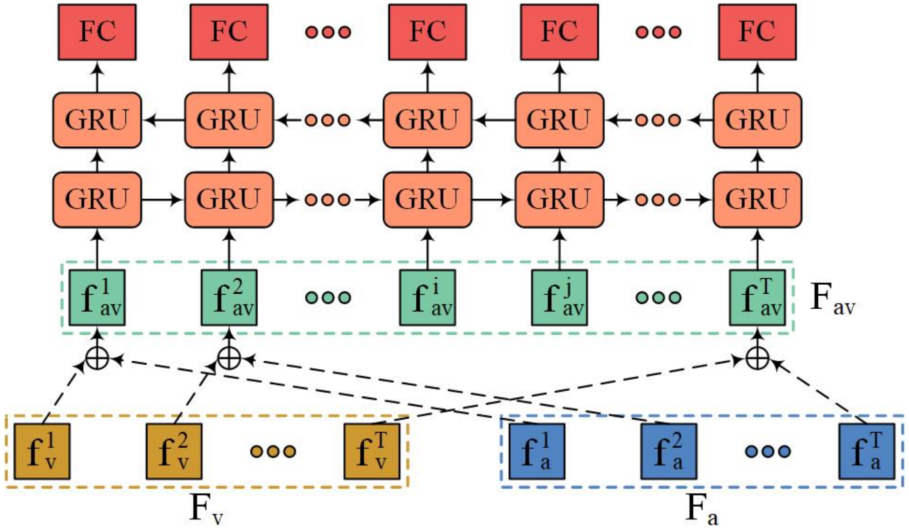


Figure 5. The architecture of the detector.

$$p_s = \frac{\exp(r_{speaking} / R)}{\exp(r_{speaking} / R) + \exp(r_{no\_speaking} / R)} \quad \cdots(1)$$

$$R = R_0 - \alpha E \quad \cdots\cdots\cdots\cdots\cdots(2)$$

$$l = -\frac{1}{T}\sum_{i=1}^{T}(g^i \log(p_s^i) + (1-g^i)\log(1-p_s^i)) \quad \cdots(3)$$

$$L_{asd} = l_{av} + \lambda l_v \quad \cdots\cdots\cdots\cdots\cdots(4)$$

Experiments

Method	Single candidate?	Pre-training?	E2E?	Params(M)	FLOPs(G)	mAP(%)
ASC (CVPR'20) [1]	✗	✓	✗	23.5	1.8	87.1
MAAS (ICCV'21) [2]	✗	✓	✗	22.5	2.8	88.8
Sync-TalkNet (MLSP'22) [44]	✓	✗	✓	15.7	1.5(0.5×3)	89.8
UniCon (MM'21) [47]	✗	✓	✗	>22.4	>1.8	92.2
TalkNet (MM'21) [37]	✓	✗	✓	15.7	1.5(0.5×3)	92.3
ASD-Transformer (ICASSP'22) [9]	✓	✗	✓	>13.9	>1.5(0.5×3)	93.0
ADENet (TMM'22) [45]	✓	✗	✓	33.2	22.8(7.6×3)	93.2
ASDNet (ICCV'21) [19]	✗	✓	✗	51.3	14.9	93.5
EASEE-50 (ECCV'22) [3]	✗	✓	✓	>74.7	>65.5	94.1
SPELL (ECCV'22) [23]	✗	✓	✗	22.5	2.6	**94.2**
Our Method	✓	✗	✓	**1.0**	**0.6**(0.2×3)	94.1

Table 1. Performance comparison for methods on the validation set of the AVA-ActiveSpeaker dataset.

Method	Speaker					
	Bell	Boll	Lieb	Long	Sick	Avg
TalkNet [37]	43.6	66.6	68.7	43.8	58.1	56.2
LoCoNet [43]	54.0	49.1	80.2	**80.4**	76.8	68.1
Our Method	**82.7**	**75.7**	**87.0**	74.5	**85.4**	**81.1**

Table 2. Comparison of F1-Score (%) on the Columbia dataset.

Ablation Studies

Kernel size	Params(M)	FLOPs(G)	mAP(%)
3	0.50	0.21	93.0
5	0.77	0.42	93.4
7	1.12	0.72	93.4
3 and 5	1.02	0.63	94.1

Table 3. Impact of convolutional kernel size.

Encoder	Params(M)	FLOPs(G)	mAP(%)
TalkNet [37]	13.68	1.53	92.8
3D convolution	2.06	1.56	92.9
Our Method	1.02	0.63	94.1

Table 4. Impact of visual feature encoder.

Encoder	Params(M)	FLOPs(G)	mAP(%)
ResNet-18 [13]	11.98	0.69	93.4
2D convolution	1.12	0.63	93.6
Our Method	1.02	0.63	94.1

Table 5. Impact of audio feature encoder.

Detector	Params(M)	FLOPs(G)	mAP(%)
None	0.82	0.63	88.0
Transformer [41]	1.02	0.63	91.5
Forward GRU	0.92	0.63	92.6
Bidirectional GRU	1.02	0.63	94.1

Table 6. Impact of the detector.

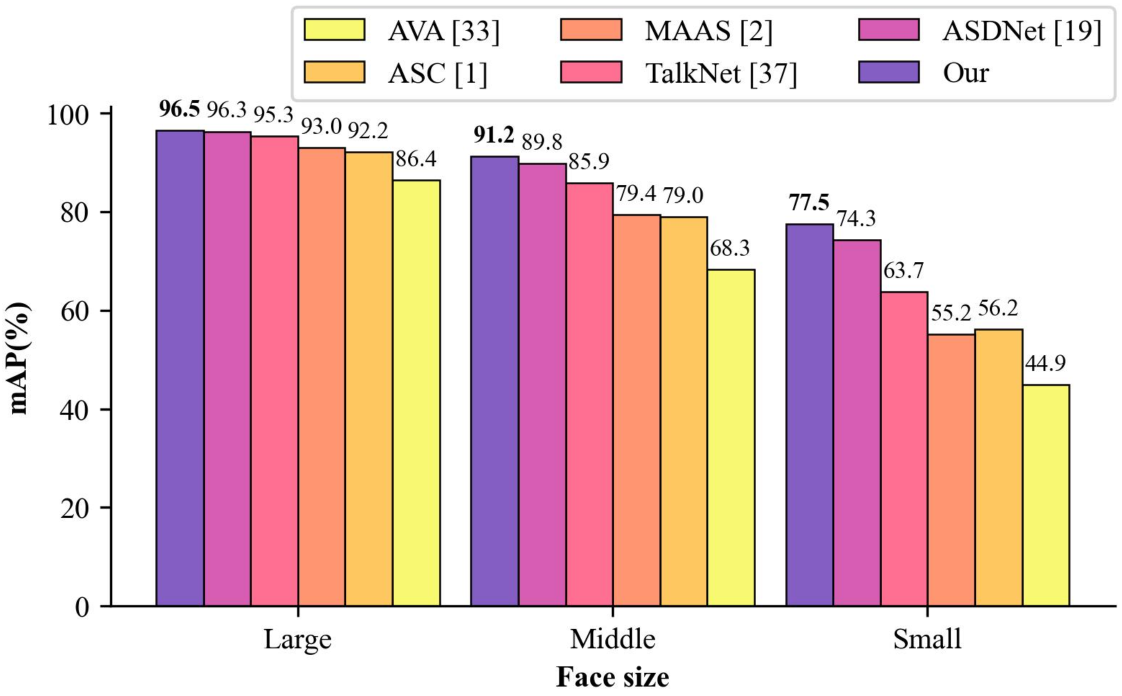
Method	Params(M)	FLOPs(G)	mAP(%)
Our (without L_{asd})	1.02	0.63	93.1
Our (with L_{asd})	1.02	0.63	94.1

Table 7. Impact of the loss function.

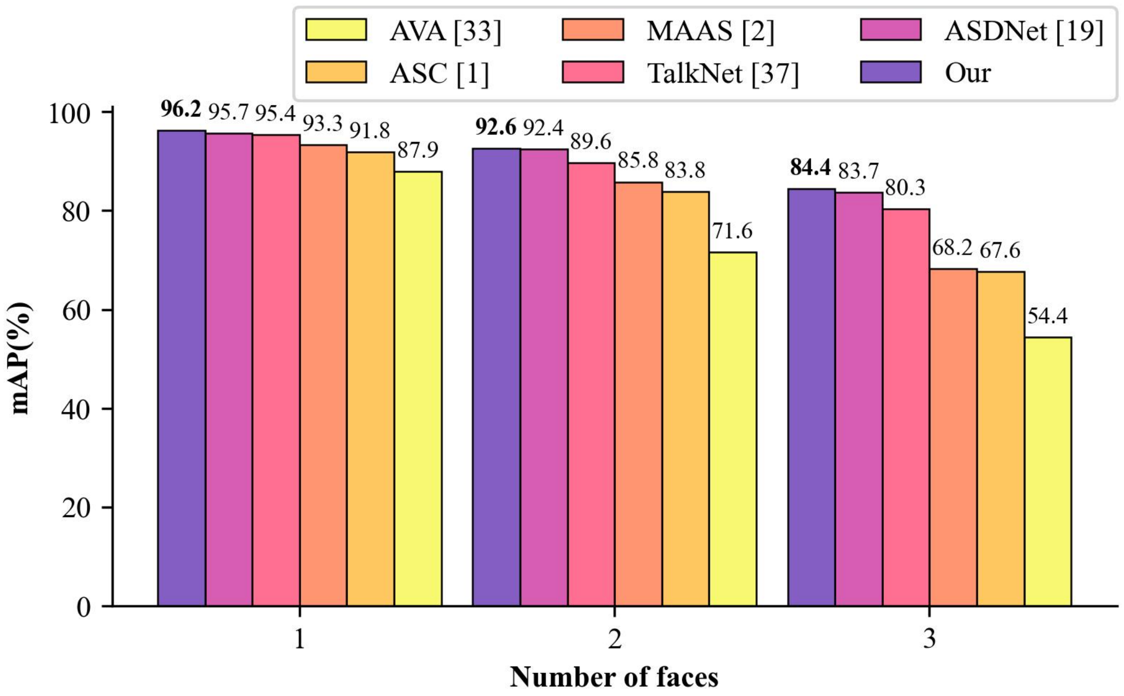
Video frames	Inference time(ms)	FPS
1 (about 0.04 seconds)	4.49	223
500 (about 20 seconds)	50.28	9944
1000 (about 40 seconds)	96.04	10412

Table 8. Impact of the number of frames on the detection speed.

Qualitative Analysis

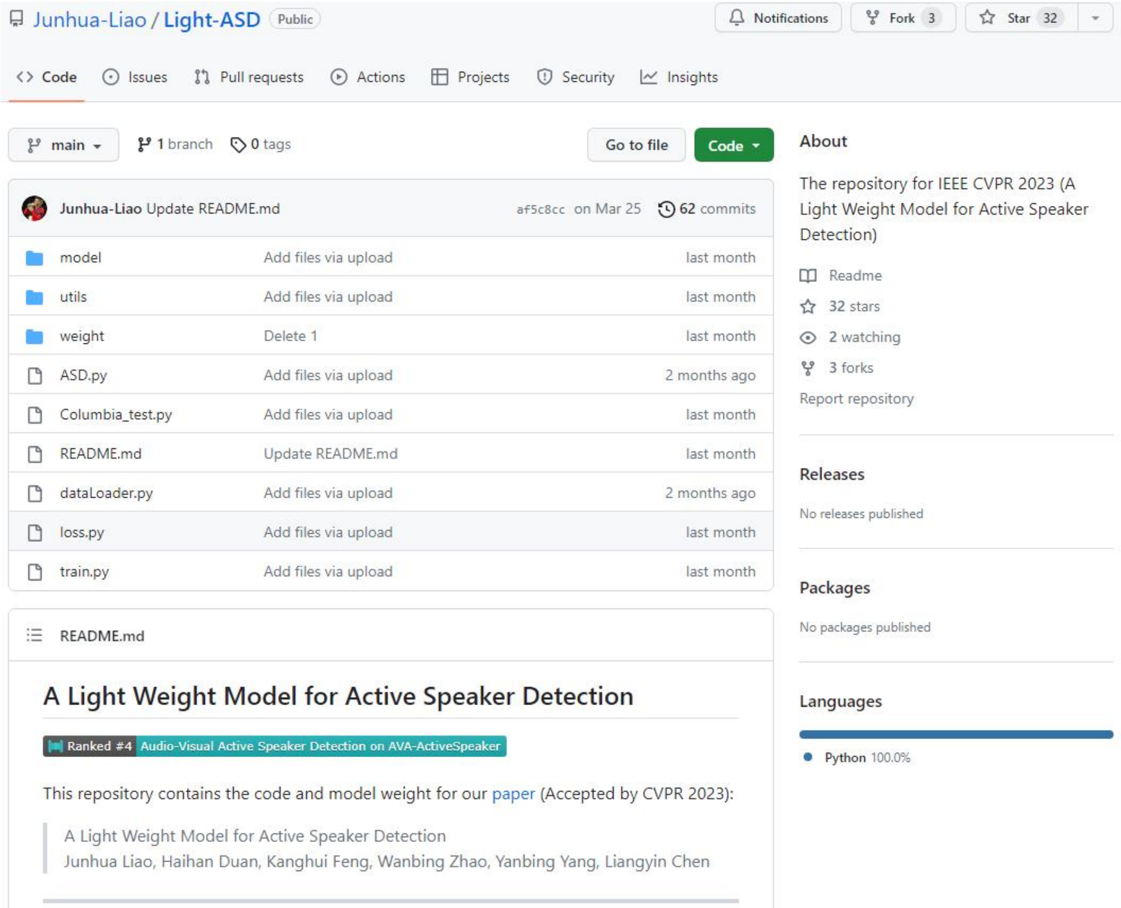


Performance comparison by face size.



Performance comparison by the number of faces on each frame.

Project Page



Project page: ***https://github.com/Junhua-Liao/Light-ASD***

Thank you!