Push-Pull: Characterizing the Adversarial Robustness for Audio-Visual Active Speaker Detection

Xuanjun Chen [1*] Haibin Wu [12*] Helen Meng [2] Hung-Yi Lee [1] Jyh-Shing Roger Jang [1]

[1]National Taiwan University [2]The Chinese University of Hong Kong

SLT 2022
9th to 12th January 2023
Doha, Qatar
Languages of the world

Introduction

Audio-Visual Active Speaker Detection (AVASD)

- **Goal.** Predict whether the face region corresponds to the current speaker.
- **Applications.** AVASD is well-developed and now is an indispensable front-end for several multi-modal applications, such as user authentication, etc.
- **Challenges.** The adversarial robustness of AVASD models hasn't been investigated, not to mention the effective defense against such attacks.

Contributions

- We first expose that audio-visual active speaker detection models are highly susceptible to adversarial attacks.
- We propose the audiovisual interaction loss to enlarge the inter-class difference and intraclass similarity, resulting in more robust AVASD models.

Methodology

AVASD Model

- **TalkNet [6].** TalkNet is one of the state-of-the-art models for Audio-visual active speaker detection (AVASD). The architecture is shown in Figure 1 (a).

Multi-Modal Adversarial Attacks

- **Framework.** The objective function for multi-modal attacks is as follows:

$$\arg\max_{\delta_a,\delta_v} \mathcal{L}(\tilde{x}_a, \tilde{x}_v, y), s.t.\ ||\delta_a||_p \leq \epsilon_a, \quad ||\delta_v||_p \leq \epsilon_v,$$

where $\tilde{x}_a = x_a + \delta_a$, $\tilde{x}_v = x_v + \delta_v$, $\mathcal{L}(\cdot)$ is the objective function to make the outputs of the audio-visual model as different as possible to y, $||\cdot||_p$ is the p-norm, and ϵ_a and ϵ_v are audio and visual perturbation budgets. In Figure 1 (b), the multi-modal attack is jointly optimized on audio-visual modality.
- **Three attack algorithms.** BIM [3], MIM [1], PGD [4].

Audio-Visual Interaction Loss (AVIL)

- **Implementation.** Suppose we have K frames for one batch and let K_s and K_n be the speech and non-speech frame numbers, respectively. Let $\mathbb{S}$ and $\mathbb{N}$ denote the index sets for speech and non-speech. We can get the center of audio speech embeddings by equation $c_{a\text{-}s} = \frac{1}{K_s}\sum_{i\in\mathbb{S}} e_{a,i}$. Similarly, we can get the other three centers $c_{a\text{-}ns}$, $c_{v\text{-}s}$, $c_{v\text{-}ns}$, which denote the centers for audio non-speech embeddings, visual speech embeddings, visual non-speech embeddings, respectively. In Figure 2, the centers are denoted with bold borders and $\mathcal{L}_1$-$\mathcal{L}_4$ are different interaction losses.
- **Training objective function.** Summing $\mathcal{L}_{CE_a}$, $\mathcal{L}_{CE_v}$, $\mathcal{L}_{CE_{av}}$ and AVILs.
- **Rationable of AVIL.** Minimizing $\mathcal{L}_1$ will equip the model with better discrimination capacity between speech and non-speech embeddings, resulting in higher inter-class differences from the models' perspective. Maximizing $\mathcal{L}_2$, $\mathcal{L}_3$ and minimizing $\mathcal{L}_4$ will force the model to render more compact intra-class features. Incorporating $\mathcal{L}_1$-$\mathcal{L}_4$ in the training process, we can simultaneously urge the model to learn both discriminative inter-class features, and compact intra-class features, leading the model less susceptible to adversarial attacks.

Experimental Setup

- **AVA-ActiveSpeaker Dataset [5].** Every face region in each frame is annotated with a bounding box that is connected over time. Randomly selected 450 genuine samples in the validation set (225 speaking and 225 non-speaking) with the correct predictions to conduct adversarial attacks.
- **Evaluation Metric.** Mean average precision (mAP(%)).

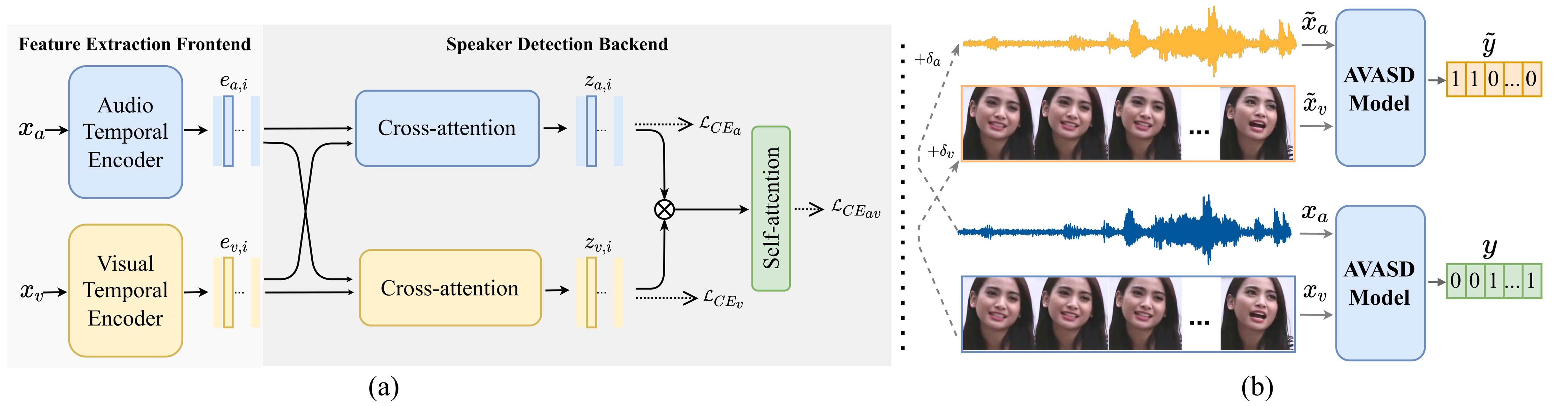

Figure 1. (a) The TalkNet framework. x_a and x_v are the audio and visual inputs, respectively. $\otimes$ denotes the concatenation procedure. $\mathcal{L}_{CE_a}$, $\mathcal{L}_{CE_v}$ and $\mathcal{L}_{CE_{av}}$ are the cross entropy losses for audio-only prediction head, visual-only prediction head, and audio-visual prediction head, respectively. (b) The audio-visual attack framework for AVASD. x_a and x_v are the audio and visual samples respectively, y is the ground-truth for the multi-sensory input $\{x_a, x_v\}$. δ_a and δ_v are the adversarial perturbations for x_a and x_v, respectively. $\tilde{y}$ is the prediction for the adversarial samples $\{\tilde{x}_a, \tilde{x}_v\}$. The adversarial attack aims at maximizing the difference between y and $\tilde{y}$.

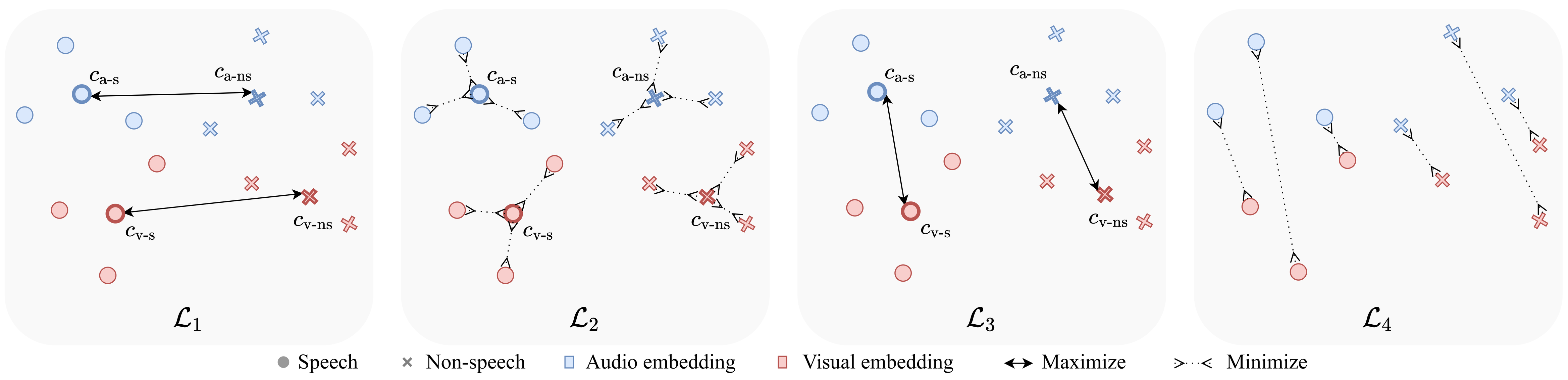

Figure 2. The Audio-Visual Interaction Loss. The circle and cross fork denote the speech and non-speech embeddings, respectively. The colors blue and red present the audio and visual embeddings, respectively. The centers are those with bold borders.

Experiment

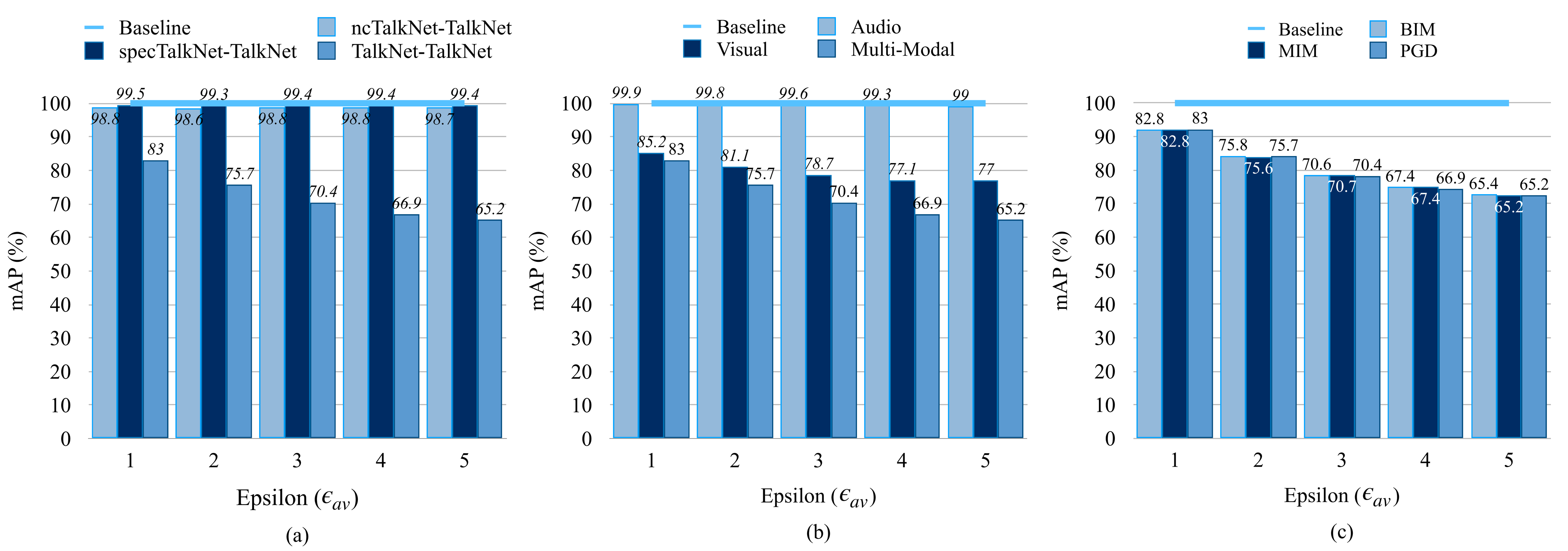

Figure 3. Adversarial attack performance of AVASD models. (a) White-box and black-box attackers under multi-modal attack with PGD method. specTalkNet and ncTalkNet are black-box attackers and TalkNet is the white-box attacker. (b) Single-modal and multi-modal attack under white-box attacker with PGD method. (c) Different attack algorithms under white-box attacker with multi-modal attack. The attack budgets of audio and visual modals are $\epsilon_a = \epsilon_{av} \times 10^{-4}$ and $\epsilon_v = \epsilon_{av} \times 10^{-1}$, respectively.

	Model	Adversarial training [2]	Clean mAP (%)	Different attack methods with $\mathcal{L}_{CE_{all}}$		
				BIM	MIM	PGD
				mAP (%)	mAP (%)	mAP (%)
(A)	$\mathcal{L}_{CE_{all}}$	✗	**92.58**	49.53	49.30	47.79
(B1)	$\mathcal{L}_{CE_{all}}$	BIM	92.15	62.7	59.26	60.01
(B2)	$\mathcal{L}_{CE_{all}}$	MIM	91.34	54.66	52.18	54.23
(B3)	$\mathcal{L}_{CE_{all}}$	PGD	91.68	58.29	58.3	56.06
(D1)	$\mathcal{L}_{CE_{all}} + \mathcal{L}_1 + \mathcal{L}_2$	✗	92.46	66.91	67.89	64.11
(D2)	$\mathcal{L}_{CE_{all}} + \mathcal{L}_1 + \mathcal{L}_3$	✗	92.20	48.16	47.92	49.27
(D3)	$\mathcal{L}_{CE_{all}} + \mathcal{L}_1 + \mathcal{L}_4$	✗	91.81	93.86	93.34	93.15
(D4)	$\mathcal{L}_{CE_{all}} + \mathcal{L}_2 + \mathcal{L}_3$	✗	92.27	57.02	63.36	61.54
(D5)	$\mathcal{L}_{CE_{all}} + \mathcal{L}_2 + \mathcal{L}_4$	✗	91.93	68.12	66.28	67.75
(D6)	$\mathcal{L}_{CE_{all}} + \mathcal{L}_3 + \mathcal{L}_4$	✗	91.70	91.79	92.48	91.01
(E1)	$\mathcal{L}_{CE_{all}} + \mathcal{L}_1 + \mathcal{L}_4$	BIM	90.63	97.85	97.6	97.47
(E2)	$\mathcal{L}_{CE_{all}} + \mathcal{L}_1 + \mathcal{L}_4$	MIM	91.70	**99.99**	**99.98**	**99.97**
(E3)	$\mathcal{L}_{CE_{all}} + \mathcal{L}_1 + \mathcal{L}_4$	PGD	91.88	97.68	97.47	98.67

Table 1. AVASD mAP(%) of different models under three attack algorithms. To conduct fair comparison, we get the data with correct prediction for model (A)-(E3), and do intersection of such data to get the testing data.

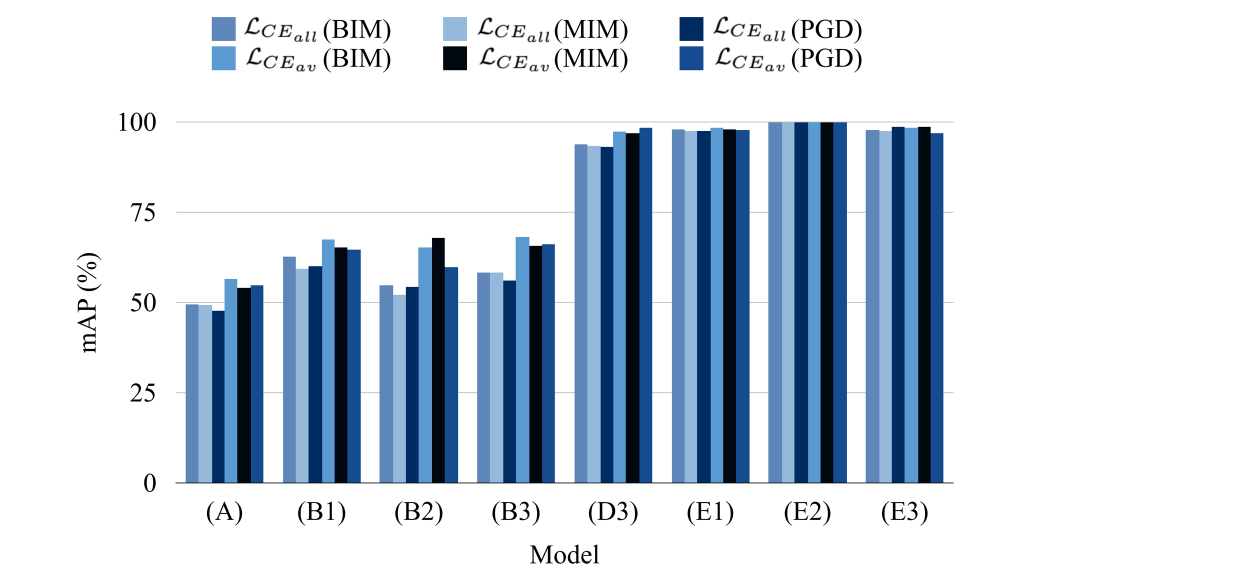

Figure 4. Training-aware ($\mathcal{L}_{CE_{all}}$) attack and inference-aware ($\mathcal{L}_{CE_{av}}$) attack scenarios.

Attacker Perspective

- **Figure 3 (a).** TalkNet is vulnerable to white-box attacks but robust to black-box attacks.
- **Figure 3 (b).** TalkNet is vulnerable to multi-modal and visual attacks but robust to audio attacks.
- **Figure 3 (c).** TalkNet has a similar degraded performance when suffering from white-box multi-modal attackers with different attack algorithms.

Defense Perspective.

- **Table 1.** Combining AVIL with adversarial training can leverage their complementary to reach the best adversarial robustness.
- **Figure 4.** The inference-aware attack scenario has the same trend as the training-aware attack scenario.

References

[1] Yinpeng Dong et al. Boosting adversarial attacks with momentum. arxiv preprint. *arXiv preprint arXiv: 1710.06081*, 2017.

[2] Ian J Goodfellow et al. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[3] Alexey Kurakin et al. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.

[4] Aleksander Madry et al. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[5] Joseph Roth et al. Ava active speaker: An audio-visual dataset for active speaker detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4492–4496. IEEE, 2020.

[6] Ruijie Tao et al. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3927–3935, 2021.

Full paper: arxiv.org/pdf/2210.00753.pdf / **Demo:** xjchen.tech/Push-Pull/index.html

R09922165@csie.ntu.edu.tw / xjchen.tech