SLT 2022
9th to 12th January 2023
Doha, Qatar
Languages of the world
Poster ID
1-1-18-MMP

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

Introduction

Audio-Visual Active Speaker Detection (AVASD)

- **Goal:** Determine if visible person in the video is speaking
- **TalkNet:** One of SOTA AVASD models as shown in Figure 1 (a)
- **Applications:** An indispensable front-end for user authentication
- **Challenges:** The adversarial robustness hasn't been investigated

Contributions

- Expose that AVASD are susceptible to multi-modal attacks
- Propose audio-visual interaction loss (AVIL) enlarges inter-class difference and intra-class similarity for improving robustness
- The AVIL outperforms adversarial training by 33.14% mAP (%)

Multi-Modal Adversarial Attacks

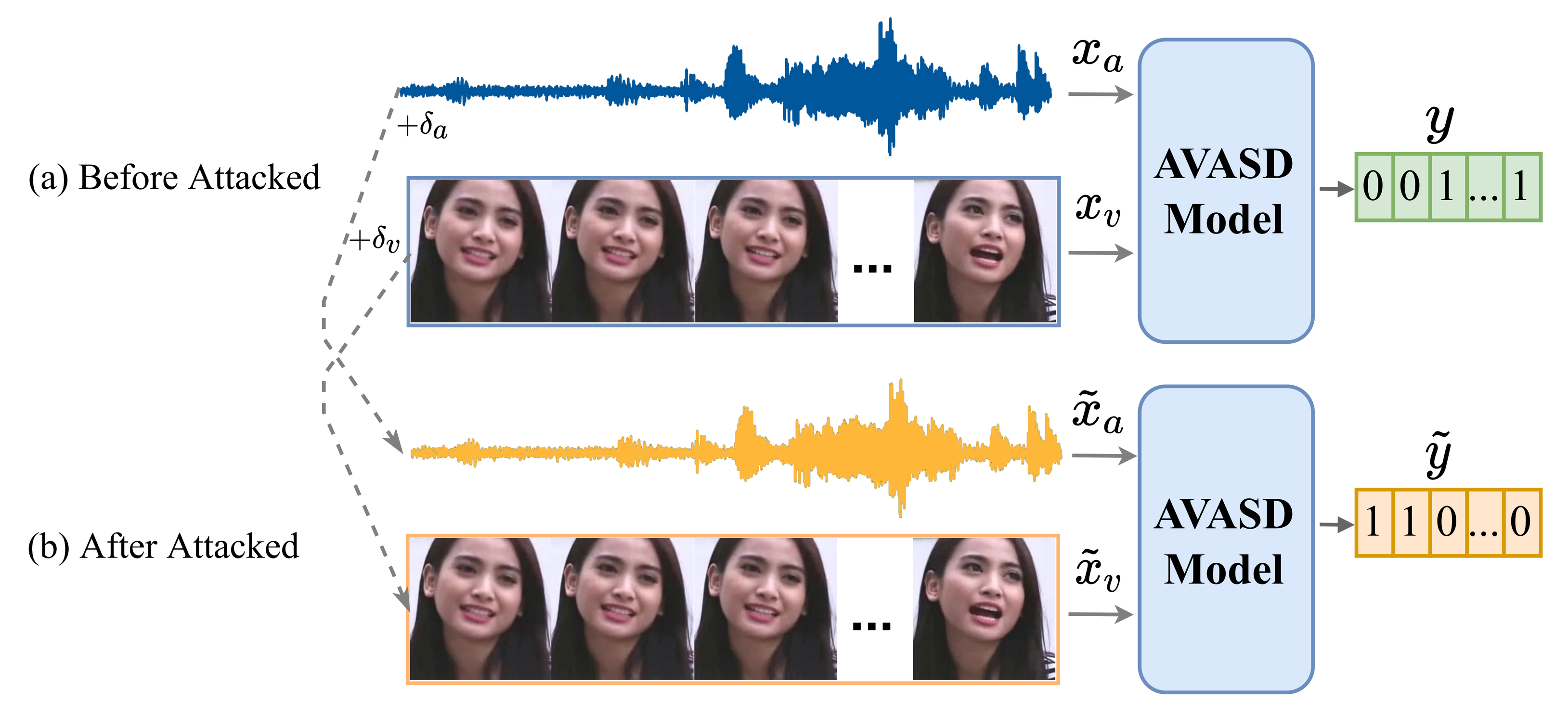

Figure 1. The multi-modal adversarial attack framework. x_a and x_v are audio and visual samples, y is ground-truth for the input. δ_a and δ_v are the adversarial perturbations for x_a and x_v. $\tilde{y}$ is the prediction for the adversarial samples $\{\tilde{x}_a, \tilde{x}_v\}$.

Attacks Objective Function

- **Goal:** Use perturbations to make model predictions wrong
- **Perturbation:** Maximize cross entropy loss $\mathcal{L}_{CE_{all}}$ difference:

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

where ϵ_a, ϵ_v are attack budget, $||\cdot||_p$ is the p-norm.

Attacks Algorithms

- Momentum-based Iterative Method (MIM)
- Projected Gradient Descent (PGD)

Attacks Defense by Audio-Visual Interaction Loss (AVIL)

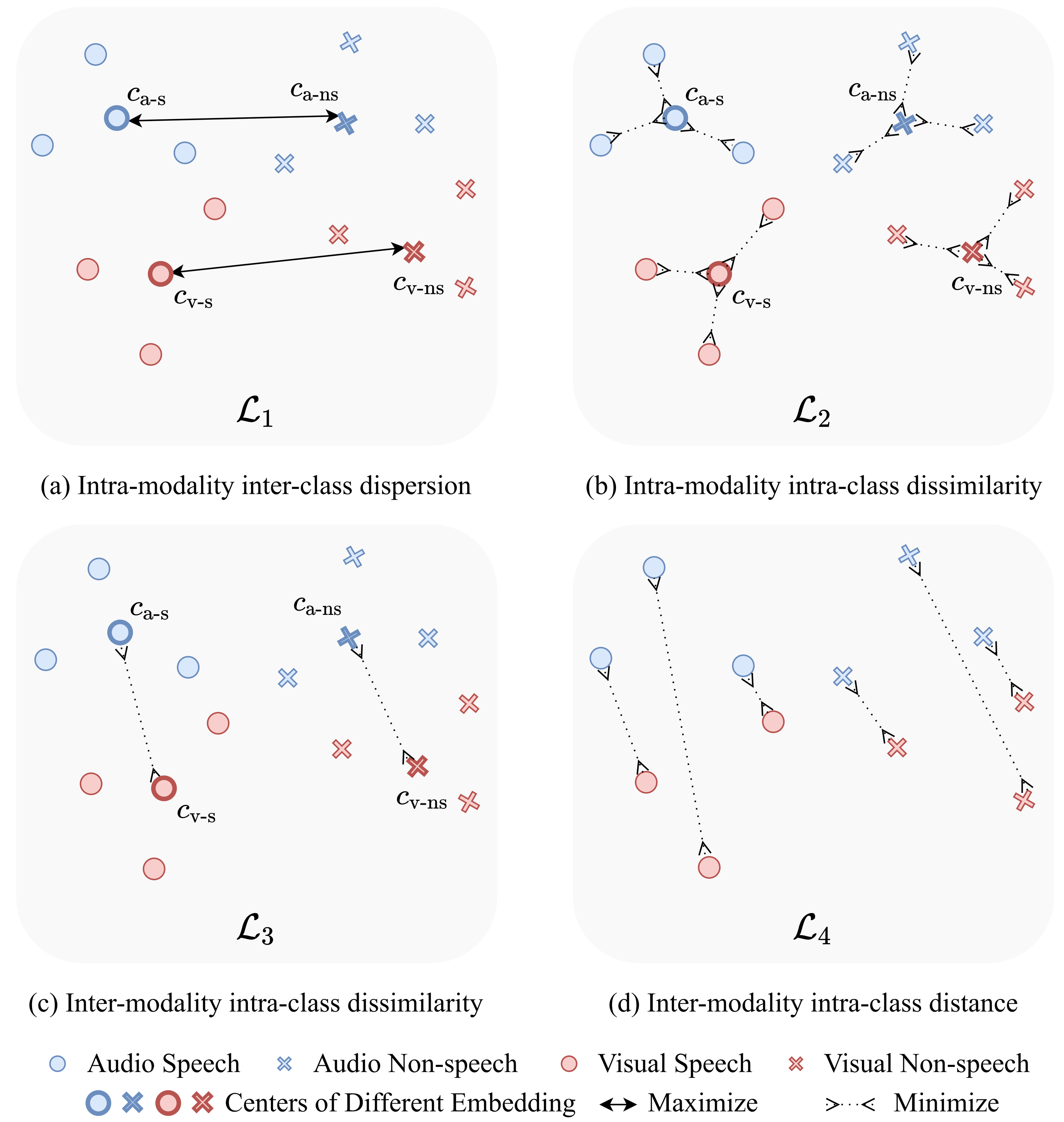

Figure 2. The Audio-Visual Interaction Loss.

Training Objective Function

- Optimize cross entropy loss $\mathcal{L}_{CE_{all}}$ and AVILs during training

Rationale of AVILs

- **Goal**: Enable the model less susceptible to adversarial attacks
- $\mathcal{L}_1$: Equip the model with better discrimination of embeddings
- $\mathcal{L}_2$-$\mathcal{L}_4$: Force the model to render compact intra-class features

Experimental Setup

- **Dataset:** AVA-ActiveSpeaker;
- **Evaluation Metric:** Mean average precision (mAP (%))

Experiment

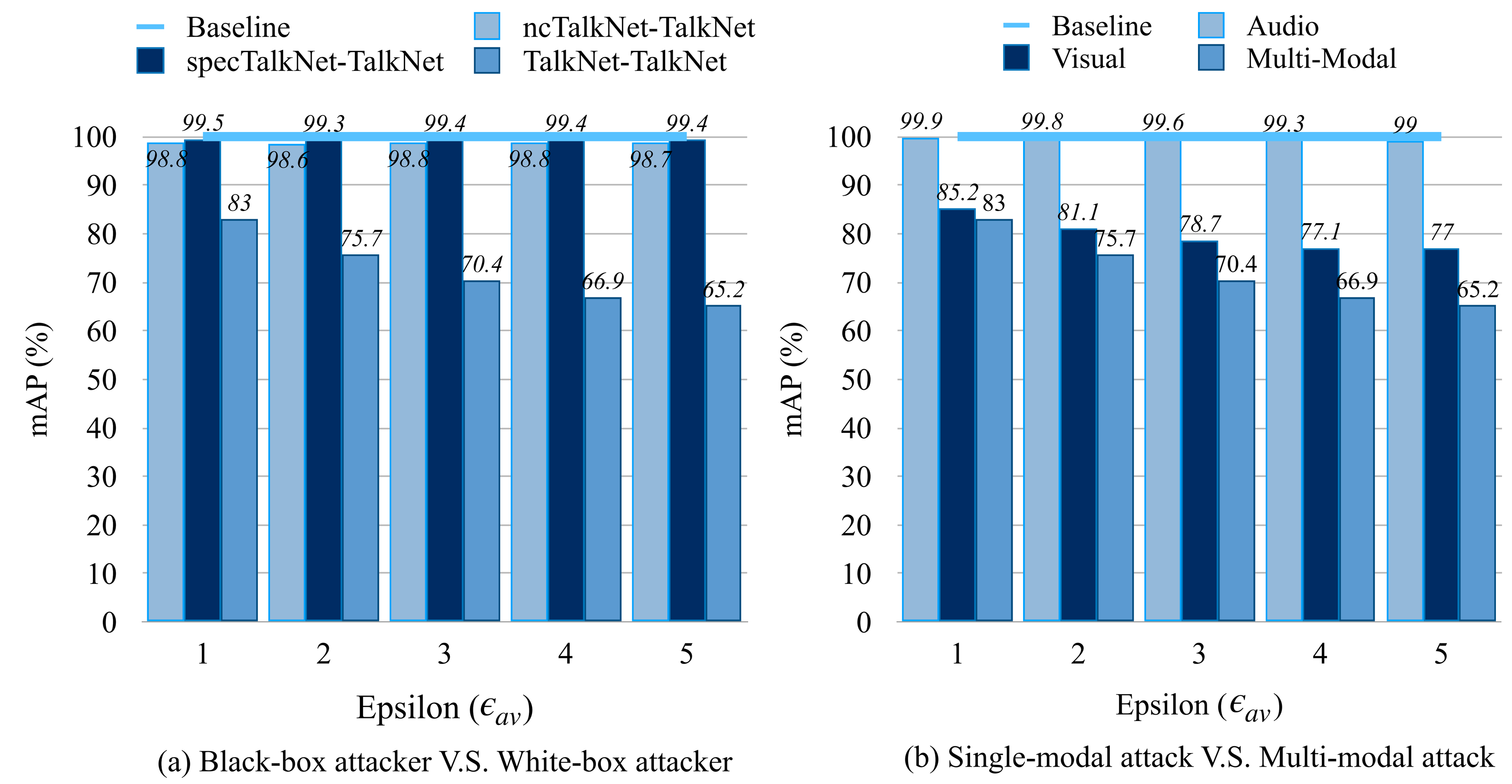

Figure 3. Adversarial attack performance of AVASD models under PGD. Black-box attackers are specTalkNet and ncTalkNet. White-box attacker is TalkNet. $\epsilon_a = \epsilon_{av} \times 10^{-4}$ and $\epsilon_v = \epsilon_{av} \times 10^{-1}$.

	Model	Adversarial training	Clean mAP (%)	MIM mAP (%)	PGD mAP (%)
(A)	$\mathcal{L}_{CE_{all}}$	✗	92.58	49.30	47.79
(B1)	$\mathcal{L}_{CE_{all}}$	MIM	91.34	52.18	54.23
(B2)	$\mathcal{L}_{CE_{all}}$	PGD	91.68	58.3	56.06
(D1)	$\mathcal{L}_{CE_{all}} + \mathcal{L}_1 + \mathcal{L}_2$	✗	92.46	67.89	64.11
(D2)	$\mathcal{L}_{CE_{all}} + \mathcal{L}_1 + \mathcal{L}_3$	✗	92.20	47.92	49.27
(D3)	$\mathcal{L}_{CE_{all}} + \mathcal{L}_1 + \mathcal{L}_4$	✗	91.81	93.34	93.15
(D4)	$\mathcal{L}_{CE_{all}} + \mathcal{L}_2 + \mathcal{L}_3$	✗	92.27	63.36	61.54
(D5)	$\mathcal{L}_{CE_{all}} + \mathcal{L}_2 + \mathcal{L}_4$	✗	91.93	66.28	67.75
(D6)	$\mathcal{L}_{CE_{all}} + \mathcal{L}_3 + \mathcal{L}_4$	✗	91.70	92.48	91.01
(E1)	$\mathcal{L}_{CE_{all}} + \mathcal{L}_1 + \mathcal{L}_4$	MIM	91.70	99.98	99.97
(E2)	$\mathcal{L}_{CE_{all}} + \mathcal{L}_1 + \mathcal{L}_4$	PGD	91.88	97.47	98.67

Table 1. AVASD mAP(%) of different models under MIM and PGD. The test data from doing the intersection of the data with the correct prediction for model (A)-(E2).

Attacker Perspective

- **Figure 3 (a):** TalkNet is vulnerable to white-box attacks
- **Figure 3 (b):** TalkNet is vulnerable to multi-modal and visual attacks

Defense Perspective

- **Table 1:** Combining AVIL with adversarial training can leverage their complementary to reach the best adversarial robustness.

Paper Demos Xuanjun Chen Haibin Wu