# Attack on Practical Speaker Verification System Using Universal Adversarial Perturbations

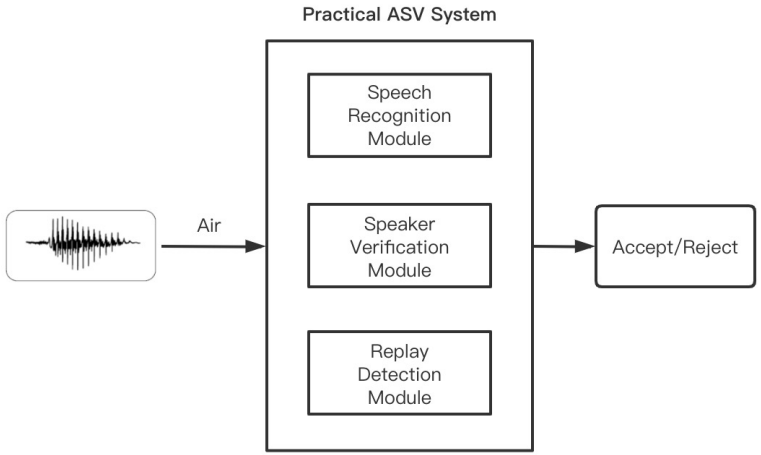
Weiyi Zhang[1], Shuning Zhao[1], Le Liu[3], Jianmin Li[1]
Xingliang Cheng[2], Thomas Fang Zheng[2], Xiaolin Hu[*1]

[1]Department of Computer Science and Technology,Tsinghua University
[2]Center for Speech and Language Technologies, BNRist, Tsinghua University
[3]Beijing d-Ear Technologies Co.,Ltd.

Contact: wy-zhang19@mails.Tsinghua.edu.cn

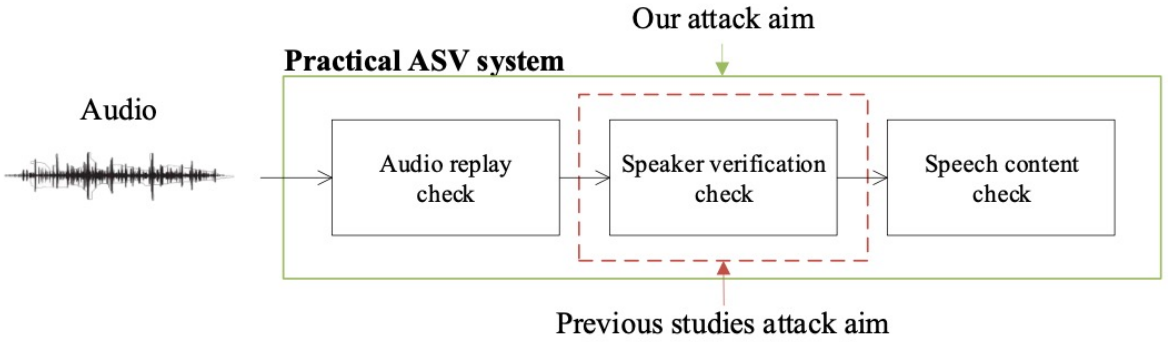
- Many authentication scenarios such as device access control, banking activities and forensics use automatic speaker verification (ASV) system for verification.
- Using **dynamic text** and **speaker verification** to ensure security.
- Performing attack on the **practical ASV system**.


From: http://www.d-ear.com/article.jsp?s=1



- Threat model : speech recognition module and replay detection module are black box, speaker verification module is white box.

- Goal of attack :
  - *a.* $R(x+\delta') = R(x)$, where $\delta' = Crop([\delta, \delta, \dots, \delta], l)$
  - *b.* $s\big(V(x+\delta'), V(y)\big) > \theta$
  - *c.* $D(x+\delta') = D(x) = True$
  - *d.* $\delta$ is independent of the text of $x$
  - *e.* $\delta$ is robust to any transformation $T(\cdot)$



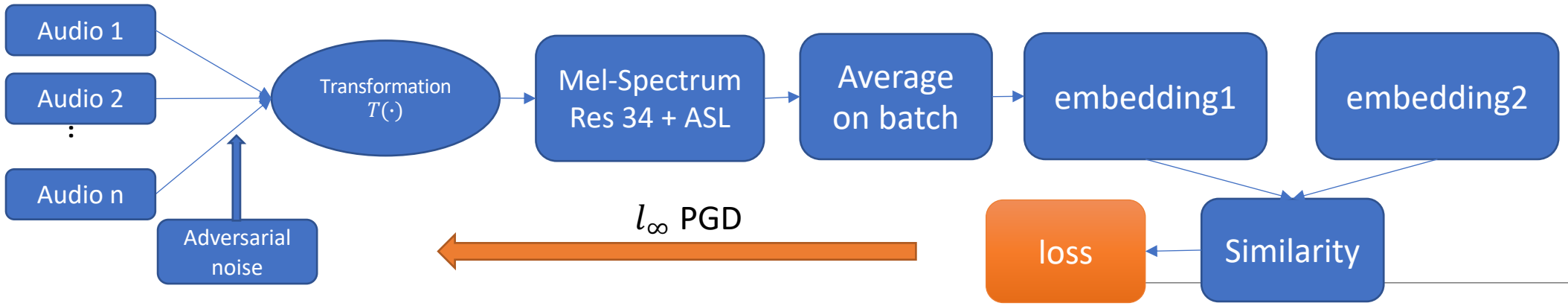
- Training set $X = \{x_1, x_2, \ldots, x_N\}$ where each $x_i$ contains different text contents. X covers great diversity about the adversary such as start offset, tune, emotion and etc.
- Loss function

$$L_1(X,\delta) = \sum_{n=1}^{N} \max(\theta - s\big(V(T(x_n) + T(\delta')), V(y)\big), -\kappa)$$

$$L_2(X,\delta) = \text{mean}(|STFT(\delta)|)$$

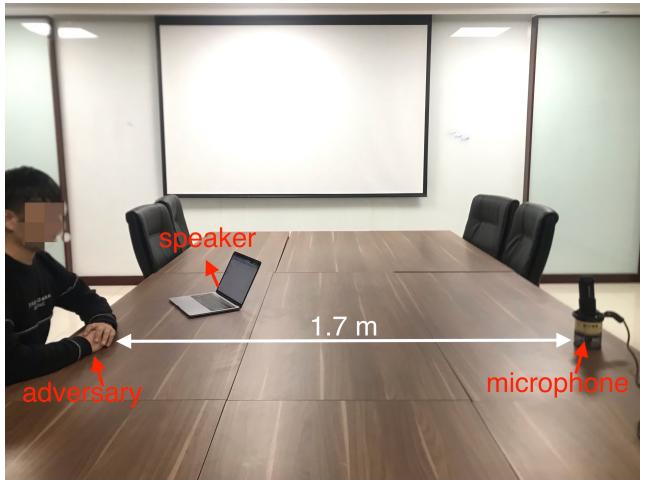
$$L(X,\delta) = L_1(X,\delta) + \gamma L_2(X,\delta)$$

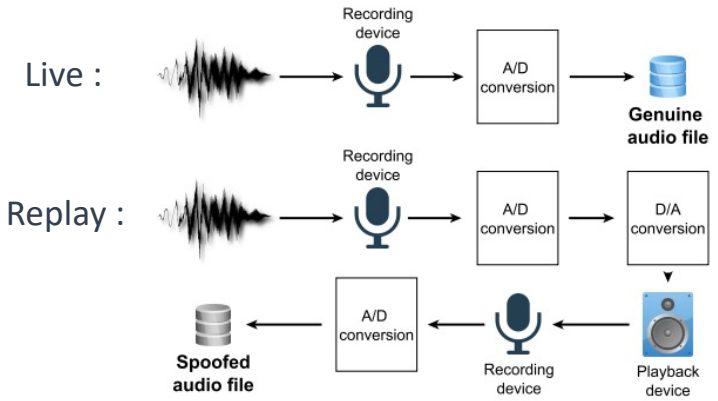
- Two-step algorithm



- Evaluation of Digital Attacks
- Evaluation of Physical Attacks
- We play the adversarial perturbation as a separate source when the adversary is speaking.
- Our adversarial examples have a high success rate to pass the replay detection.

| Attack type | Steps | ASR(%) | WER(%) | SNR(dB) |
|---|---|---|---|---|
| Clean data | N/A | 0 | 12.95 | N/A |
| intra-gender/baseline | 236 | 98.43 | 32.33 | 16.90 |
| intra-gender/ours | 846 | 98.65 | 19.43 | 23.66 |
| inter-gender/baseline | 617 | 96.63 | 37.57 | 16.55 |
| inter-gender/ours | 1872 | 96.40 | 21.53 | 22.26 |




| Attack type | ASR(%) | WER(%) | CER(%) |
|---|---|---|---|
| Clean | 0 | 11.42 | 5.78 |
| Gaussian | 0 | 17.77 | 10.06 |
| Baseline | 80.00 | 21.82 | 14.48 |
| Ours | 100.00 | 14.97 | 7.53 |

| Method | Number | Rate(%) |
|---|---|---|
| Previous | 45 | 37.7 |
| Ours | 120 | 67.7 |

- We proposed a two-step algorithm to generate universal adversarial perturbations for attacking the practical speaker verification system.
- We study the vulnerability of PSV system in physical world and help researchers to improve the security of such applications.

清華大學
Tsinghua University

# Thanks

Presenter: Weiyi Zhang

Contact: wy-zhang19@mails.Tsinghua.edu.cn