Push the Limit of Adversarial Example Attack on Speaker Recognition in Physical Domain

Presenter: Qianniu Chen

Qianniu Chen[1], Meng Chen[1], Li Lu[1], Jiadi Yu[2], Yingying Chen[3], Zhibo Wang[1], Zhongjie Ba[1], Feng Lin[1], Kui Ren[1]

[1] *Zhejiang University, Zhejiang, China*
[2] *Shanghai Jiao Tong University, Shanghai, China*
[3] *Rutgers University, New Brunswick, NJ, USA*

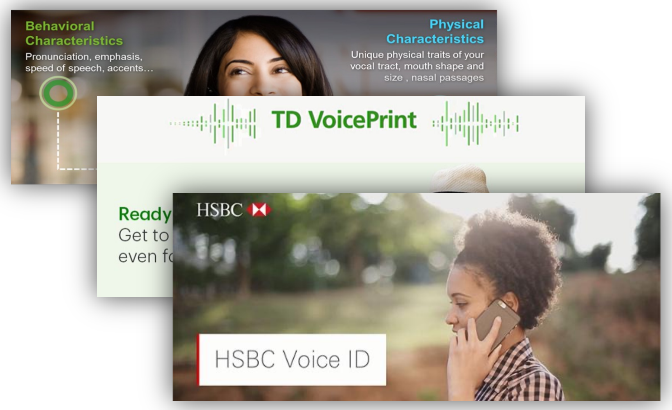
Background

- Speaker Recognition (SR) achieves wide applications in our daily life

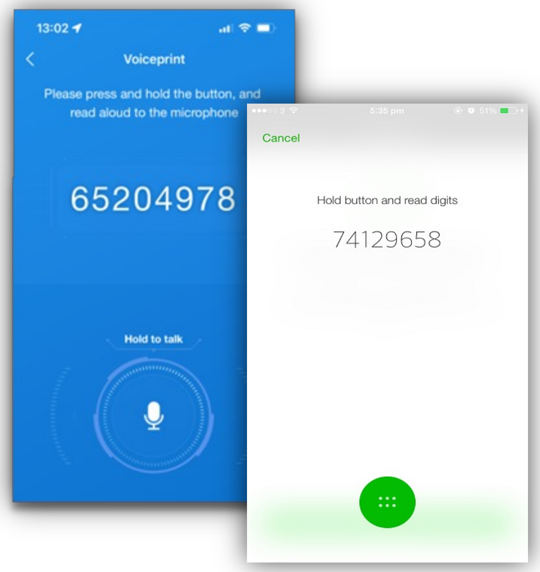
Background

- Speaker Recognition (SR) achieves wide applications in our daily life

Voice Assistant



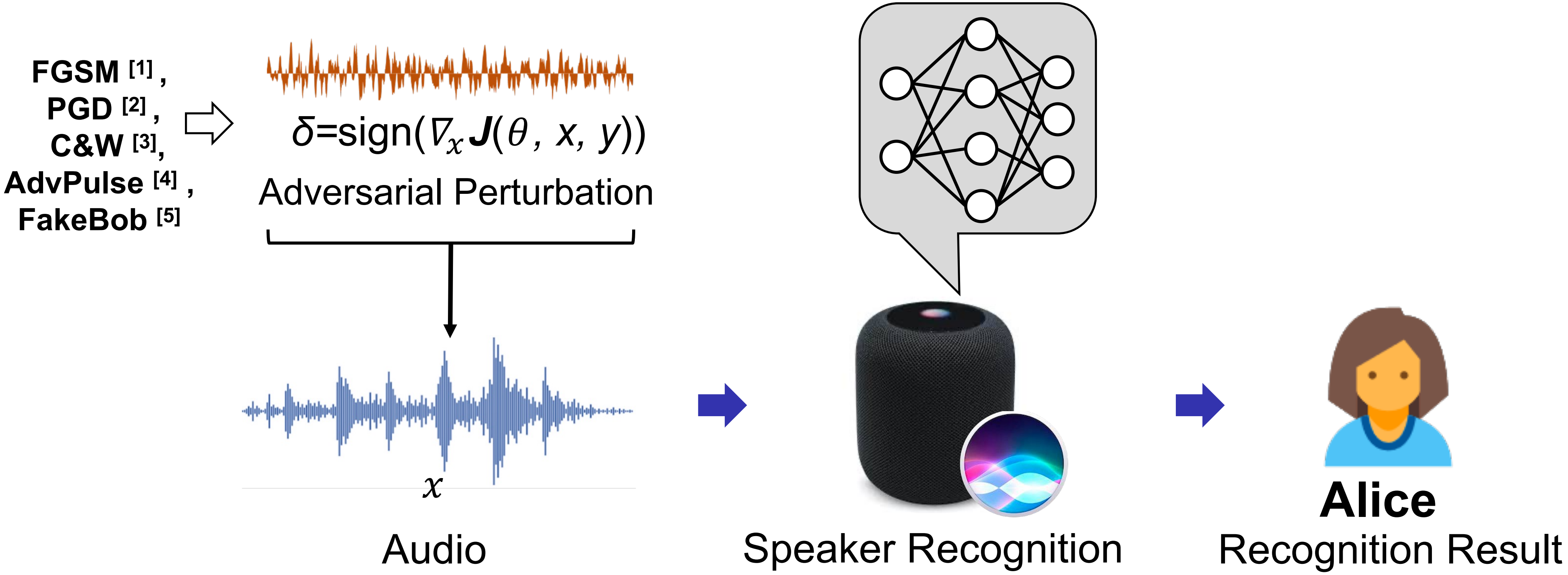
Mobile Bank Voice Password



APP Voice Lock

Background

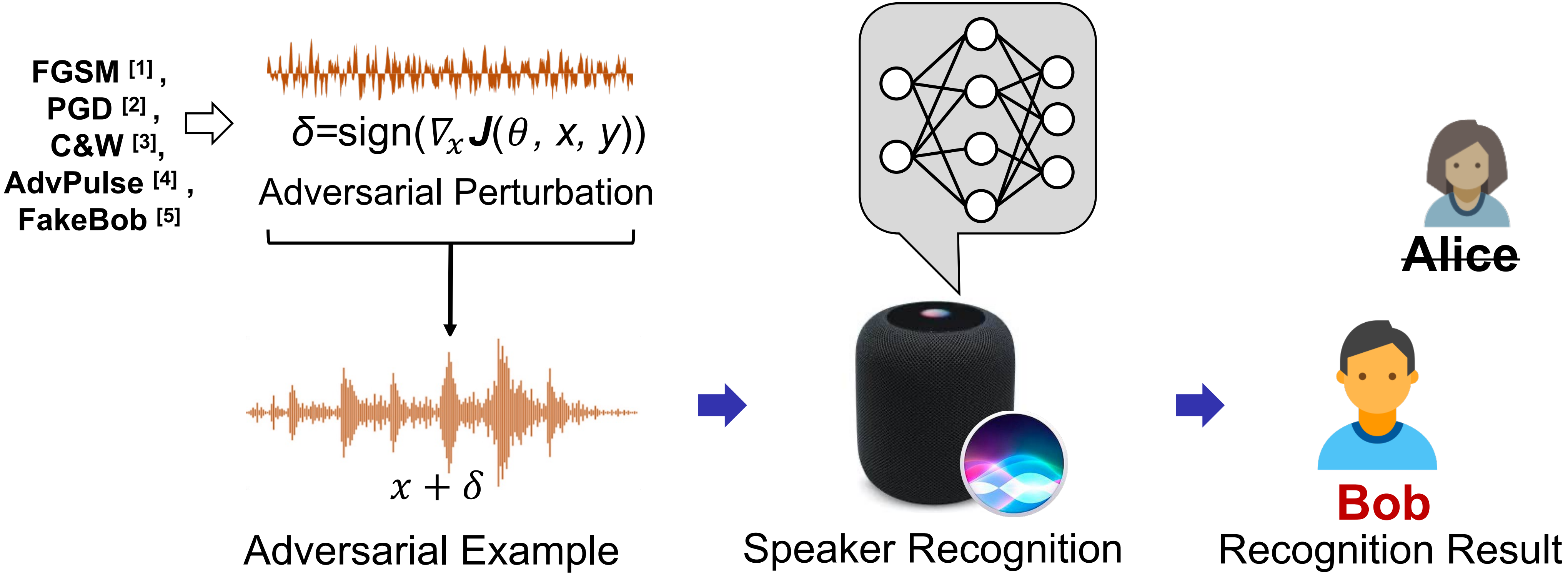
- Speaker Recognition System **& Audio Adversarial Example**



[1] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples," in Proceedings of ICLR. 2015.
[2] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. in Proceedings of ICLR. 2018.
[3] N. Carlini and D. A. Wagner. Towards evaluating the robustness of neural networks. in Proceedings of IEEE S&P. 2017.
[4] Z. Li, Y. Wu, J. Liu, et al. Advpulse: Universal, synchronization-free, and targeted audio adversarial attacks via subsecond perturbations. in Proceedings of ACM CCS. 2020.
[5] G. Chen, S. Chen, L. Fan, et al. Who is real bob? adversarial attacks on speaker recognition systems. in Proceedings of IEEE S&P. 2021.

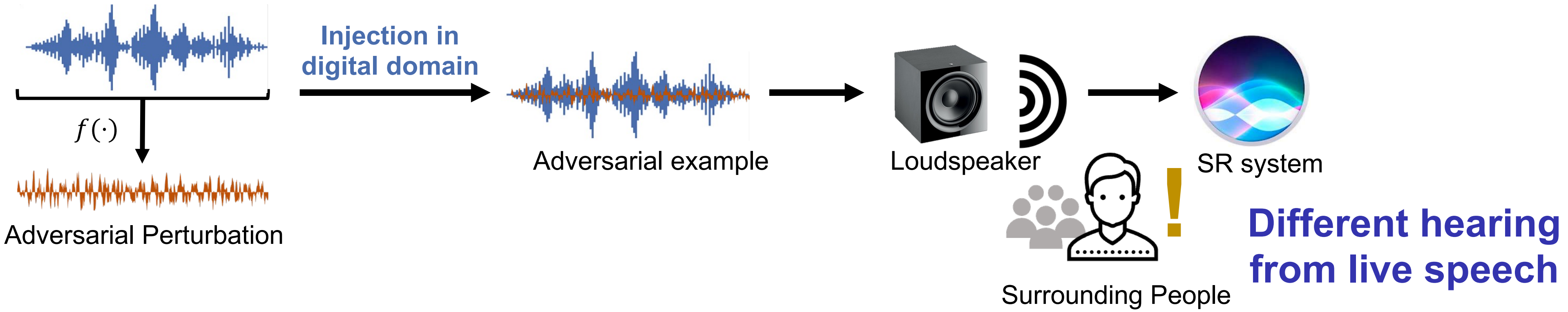
Background

- Speaker Recognition System **& Audio Adversarial Example**



[1] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples," in Proceedings of ICLR. 2015.
[2] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. in Proceedings of ICLR. 2018.
[3] N. Carlini and D. A. Wagner. Towards evaluating the robustness of neural networks. in Proceedings of IEEE S&P. 2017.
[4] Z. Li, Y. Wu, J. Liu, et al. Advpulse: Universal, synchronization-free, and targeted audio adversarial attacks via subsecond perturbations. in Proceedings of ACM CCS. 2020.
[5] G. Chen, S. Chen, L. Fan, et al. Who is real bob? adversarial attacks on speaker recognition systems. in Proceedings of IEEE S&P. 2021.

Threat Model



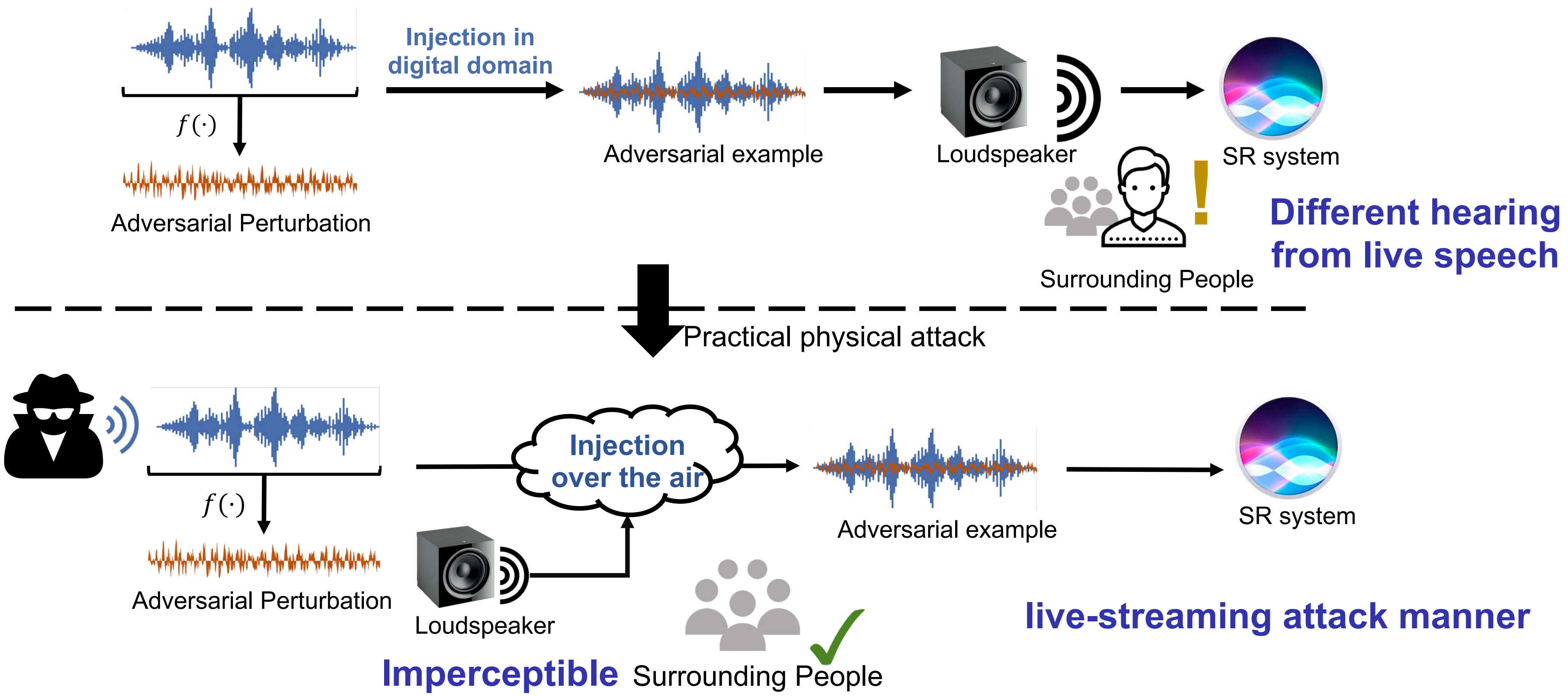
Replay **VOICE AUDIO** with adversarial perturbation

Limited to an ideal attack scenario without others around

Threat Model

Replay **VOICE AUDIO** with adversarial perturbation

Limited to an ideal attack scenario without others around

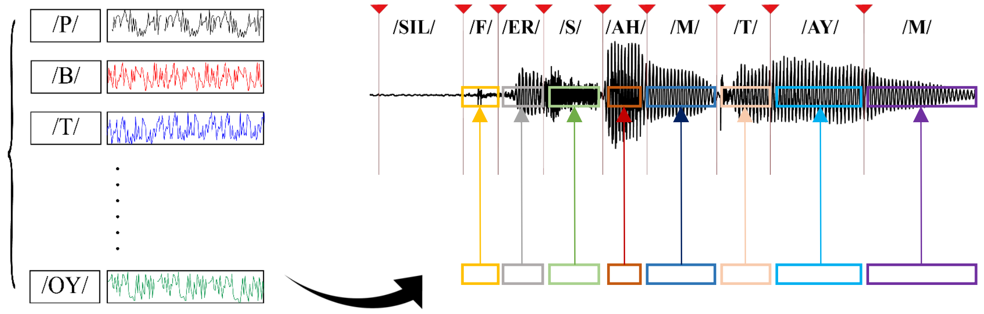


Attack Overview

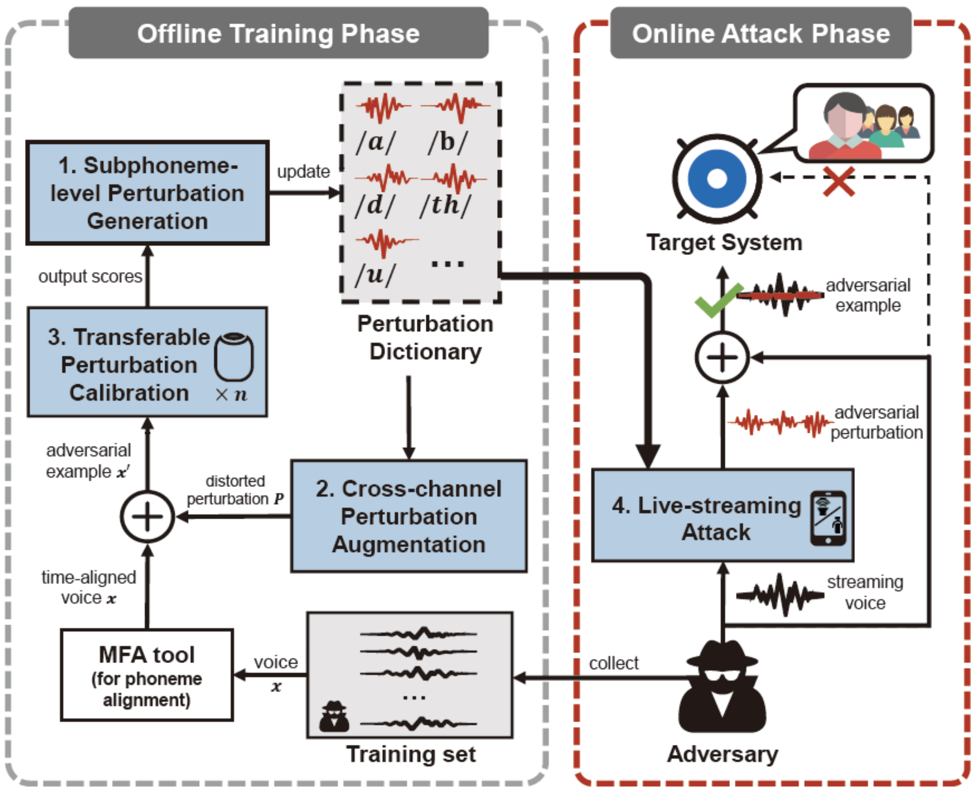
PhyTalker: a live-streaming, channel-robust, transferable audio adversarial example attack

Universal Adversarial Perturbation on Phonemes

- **Combinability**: fast generate perturbation for any speech according to its phoneme sequence.
- **Stability:** cause stable attack effectiveness on the stable acoustic characteristics of phonemes.



Function of perturbation on phonemes

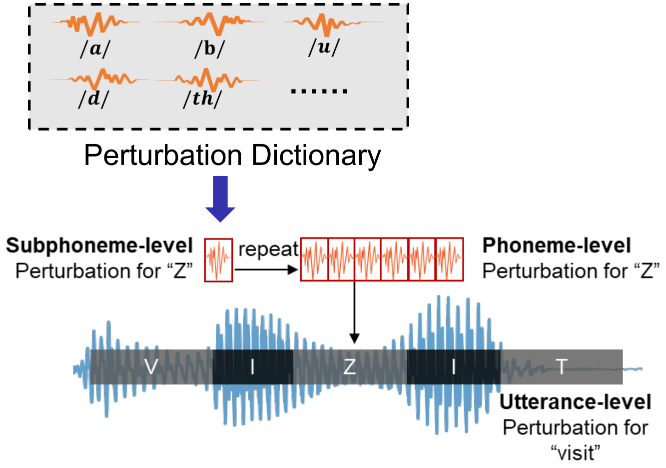


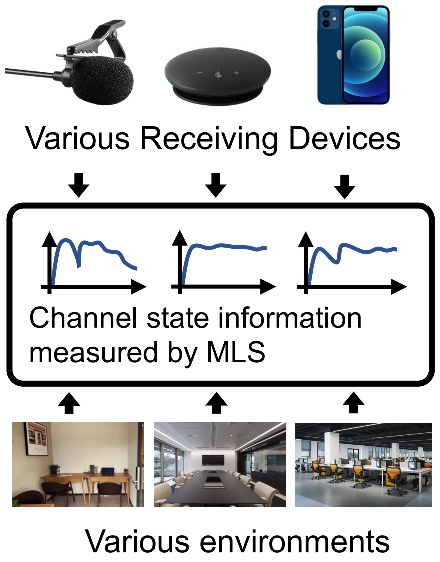
Attack overview of PhyTalker

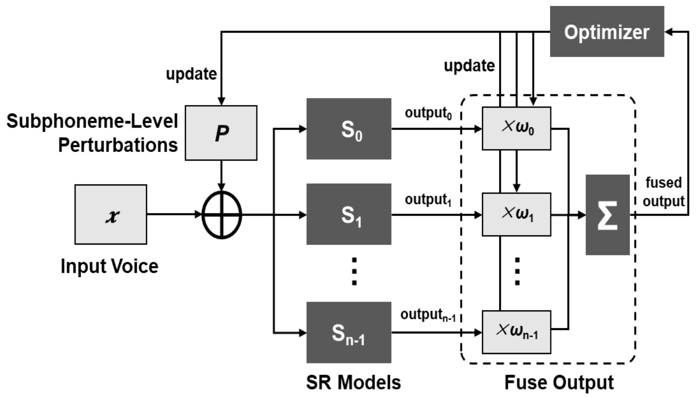
Design

Offline Training Phase

- **Subphoneme-level Perturbation**: use fixed short perturbation (<25ms) to form phoneme-level perturbation with various duration(>50ms), repetitively
- **Channel Augmentation**: explore real channel state information for data augmentation with MLS[6]
- **Transferable Calibration:** employ the ensemble learning method to improve transferability
- **Expectation Optimization：**train on a large training set instead of the specific audio







Algorithm 1 Global Optimization Procedure.

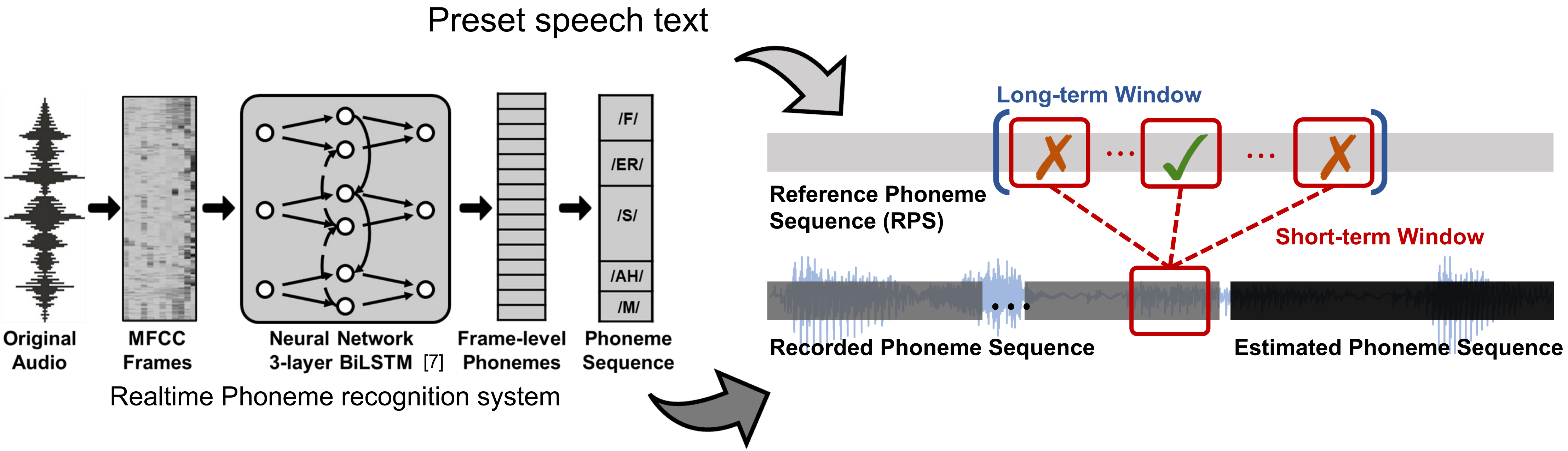
```
Input:
    Dataset χ = {(x_0, t0), (x_1, t_1), ··· , (x_{n−1}, t_{n−1})}, SR classifiers S_i with threshold θ_i (i = 0, 1, ··· , k − 1), UIR set
    C target user y_t, amplitude upper bound ε.
Output:
 Initialize P with normal distribution
 repeat
     for each (X, T) sampled from χ do
         L_t ← 0
         for each c in C do
             X_adv ← G(X, P ∗ c)
             s ← Σ_0^{k−1} w_i · L_{S_i,θ_i}(X_adv, y_t)
             L ← max{θ, max_{i≠y_t} s_i} − s_{y_t}
             L_t ← L_t + L
         end for
         Minimize L_t to update P
     end for
 until Early Stopping
 return P;
```

[6] Douglas D. Rife and John Vanderkooy. Transfer-function measurement with maximum-length sequences. Journal of the Audio Engineering Society 37, 6 (June). 1989.

Design

Online Attack Phase

- **Real-time phoneme extraction:** extract current phoneme sequence from the live speech with a fast neural phoneme recognition system
- **Phoneme Alignment：**locate the current phoneme in the RPS with long-short term

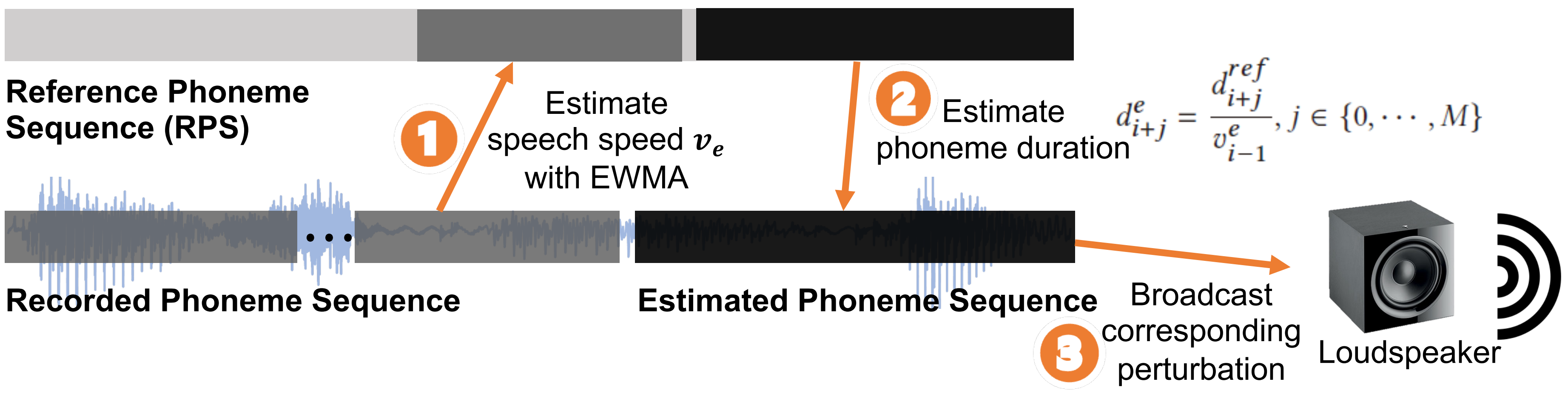


[7]Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks 18, 5-6. 2005.

Design

Online Attack Phase

- **Real-time phoneme extraction:** extract current phoneme sequence from the live speech with a fast neural phoneme recognition system
- **Phoneme Alignment :** locate the current phoneme in the RPS with long-short term
- **Phoneme Estimation :** Estimate speech voice and patch phoneme durations by referring RPS



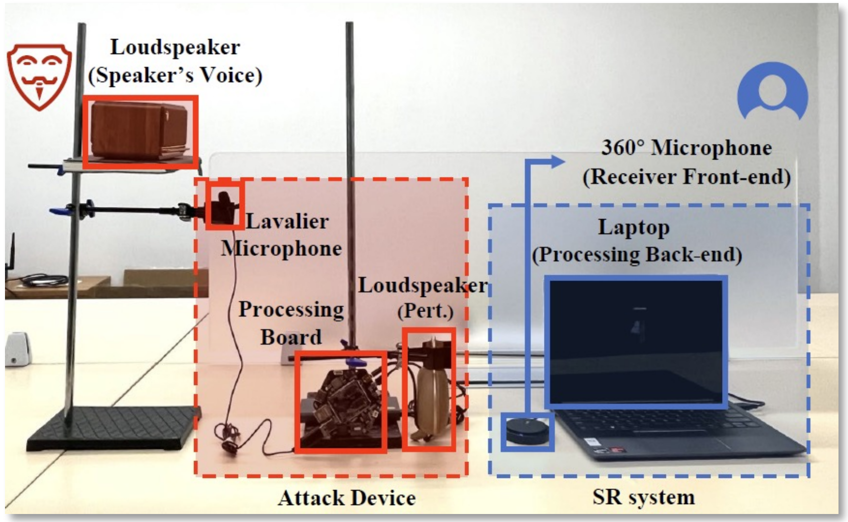
Evaluation

Target Systems Setting

- Architectures: x-vector[8] / d-vector[9] / DeepSpeaker[10]
- Training Set: Voxceleb [11] corpus
- Test Set: LibriSpeech[12] corpus
- Enrollers: 5 speakers(3 males and 2 females)
- Backend: Lenovo Xiaoxin Pro 13

Attack Setting

- Adversaries: 10 speakers(5 males and 5 females)
- Attack Device: ReSpeakerCore v2
- Subphoneme-level perturbation duration: 12.5ms
- Livestreaming synchronization: 0.5s/alignment
- Channel augmentation: 8 CIRs per (receiver, environment)
- Ensemble learning: 4 ensemble models



[8] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez- Dominguez. Deep neural networks for small footprint text-dependent speaker verification. in Proceedings of IEEE ICASSP. 2014.
[9] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur. Xvectors: Robust dnn embeddings for speaker recognition. in Proceedings of IEEE ICASSP. 2018.
[10] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu. Deep speaker: an end-to-end neural speaker embedding system. CoRR, vol. abs/1705.02304. 2017.
[11] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. VoxCeleb: A Large-Scale Speaker Identification Dataset. In Processings of ISCA Interspeech. 2017.
[12] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In Processings of IEEE ICASSP. 2015.

Evaluation

- **Overall performance**
- Human Imperceptibility

Table 3: Overall ASRs, SNR, MCD and RTF of *PhyTalker* and SOTA works under physical attack scenarios.

Attack	ASR(%)			SNR	MCD	RTF
	d-vec.	x-vec.	D.S.	(dB)	(dB)	
PhyTalker	85.5	80.5	90.5	16.8	2.45	0.5
FakeBob[5]	63.3	77.4	69.8	11.6	4.15	95.3
AdvPulse[4]	N/A	89.9	N/A	4.7	N/A	<1.0

ASR: Attack Success Rate for effectiveness (the higher the better)

MCD: Mel Cepstral Distortion for audibility (the lower the better)

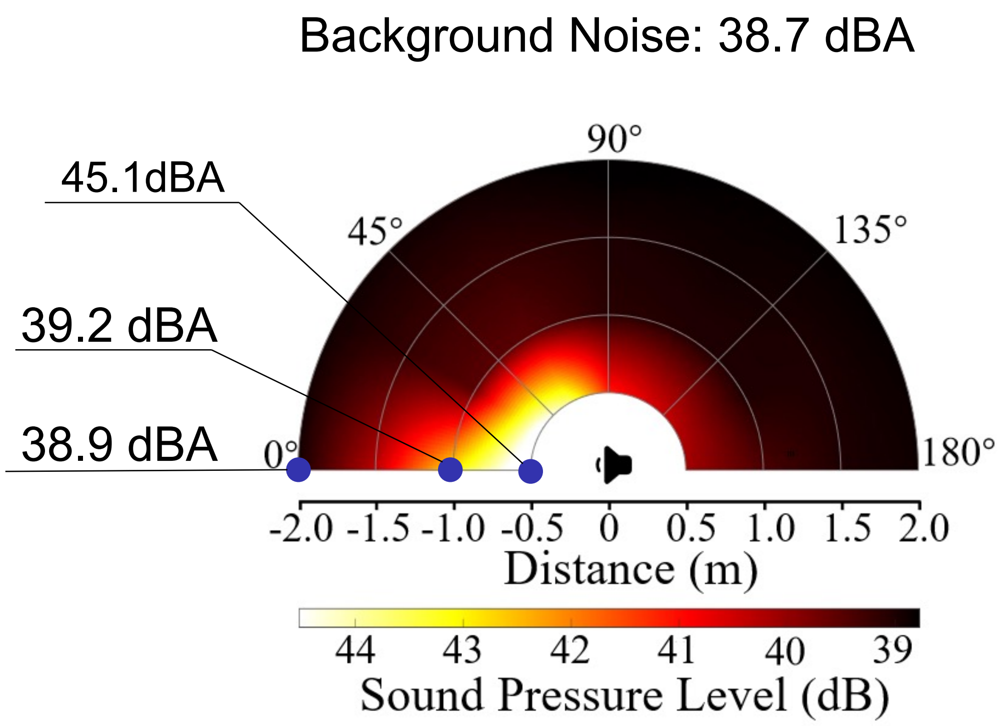
SNR: Signal-to-Noise Ratio for audibility (the higher the better)

RTF: Real Time Factor for efficiency (the lower the better)

[4] Z. Li, Y. Wu, J. Liu, et al. Advpulse: Universal, synchronization-free, and targeted audio adversarial attacks via subsecond perturbations. in Proceedings of ACM CCS. 2020.
[5] G. Chen, S. Chen, L. Fan, et al. Who is real bob? adversarial attacks on speaker recognition systems. in Proceedings of IEEE S&P. 2021.

Evaluation

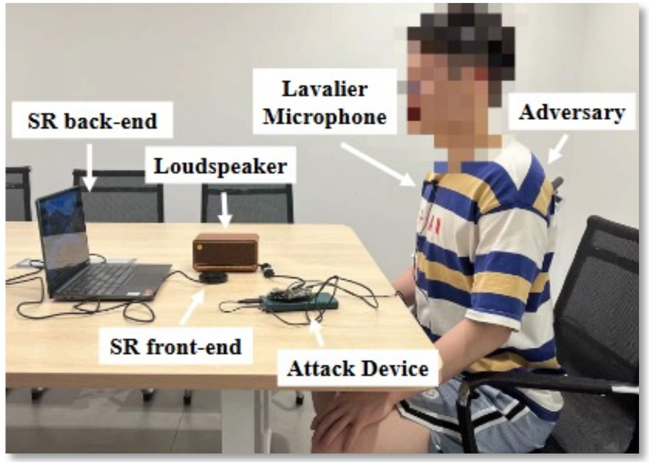
- Overall performance
- **Human Imperceptibility**

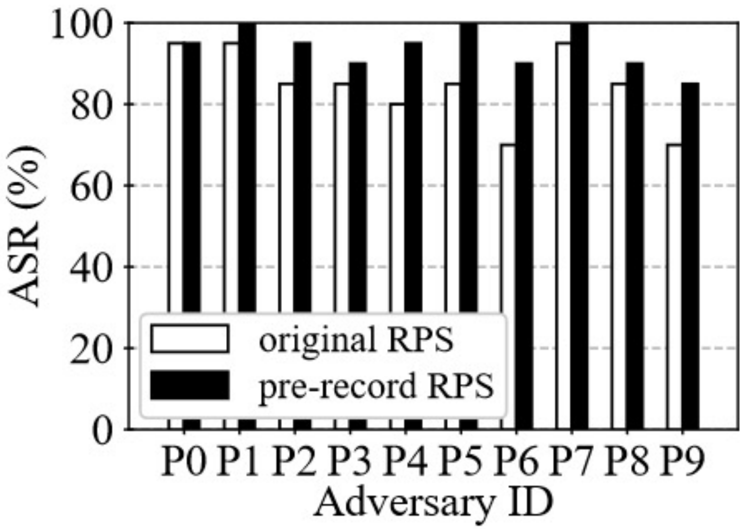


Evaluation: in-the-wild evaluation

Evaluation Setup

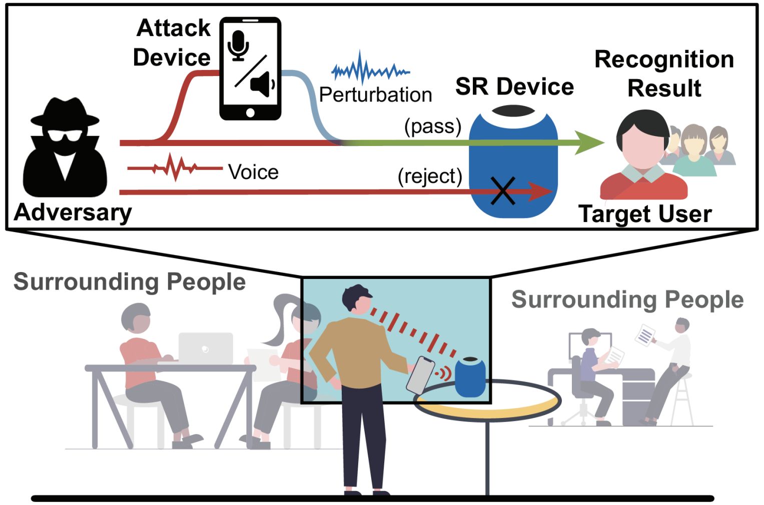
- 10 volunteers as adversaries (4 females and 6 males)
- 20-minute voice record/volunteer for training
- 10 utterances per/volunteer for evaluation





Conclusion

- Explore **three major challenges** underlying a practical physical attack scenario
- Propose a subphoneme-level, channel-robust and transferable adversarial example attack to solve the challenges
- Enables an adversary to conduct **a live-streaming attack manner** in **physical domain**

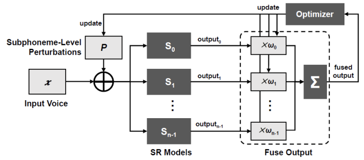


Major challenges

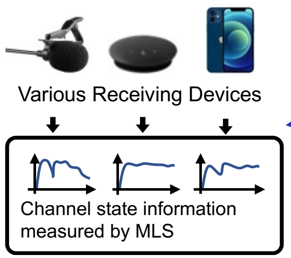
1. Livestreaming manner
2. Channel robustness
3. Black-box optimization

Conclusion

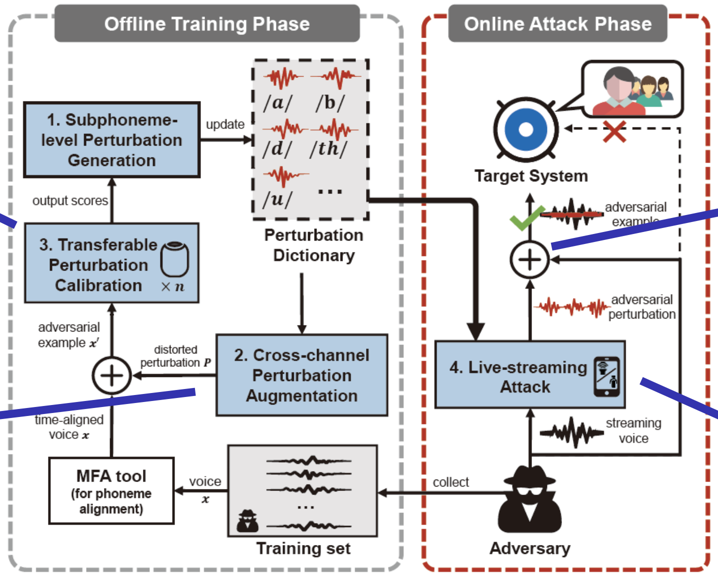
- Explore **three major challenges** underlying a practical physical attack scenario
- Propose a **subphoneme-level**, **channel-robust** and **transferable** adversarial example attack to solve the challenges
- Enables an adversary to conduct **a live-streaming attack manner** in **physical domain**

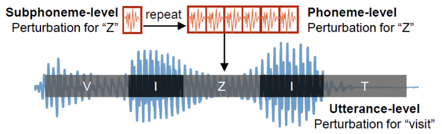


Ensemble learning

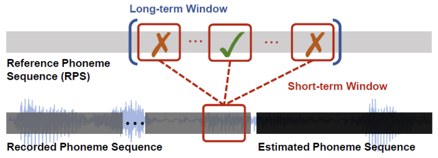


Channel augmentation





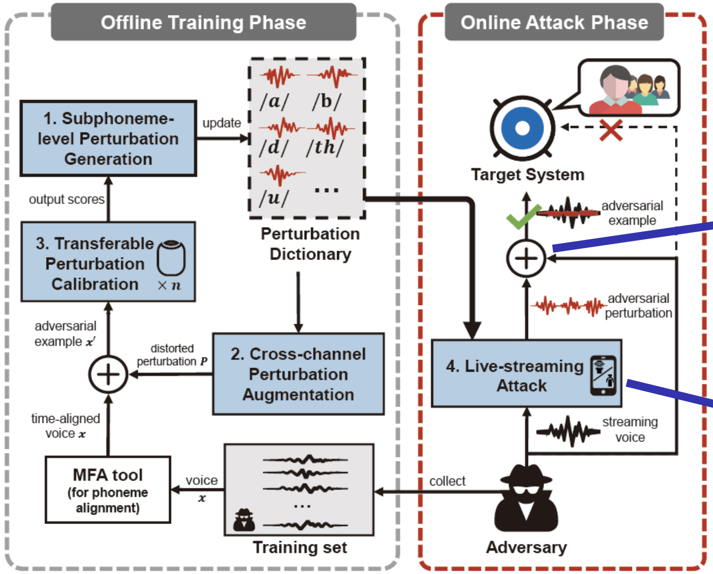
Subphoneme-level perturbation

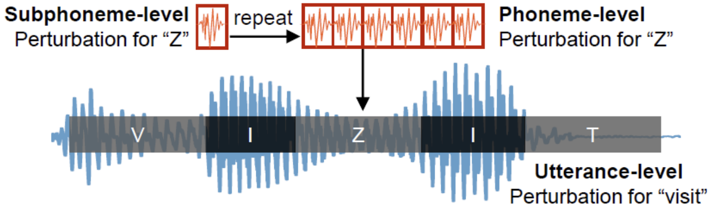


Real-time synchronization

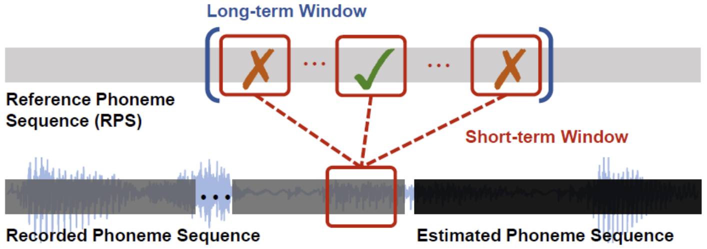
Conclusion

- Explore **three major challenges** underlying a practical physical attack scenario
- Propose a **subphoneme-level**, **channel-robust** and **transferable** adversarial example attack to solve the challenges
- Enables an adversary to conduct **a live-streaming attack manner** in **physical domain**





subphoneme-level perturbation injection



real-time synchronization mechanism

SenSys 2022

Thank you!

qianniuchen@zju.edu.cn

浙江大学 ZHEJIANG UNIVERSITY

上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

RUTGERS THE STATE UNIVERSITY OF NEW JERSEY