

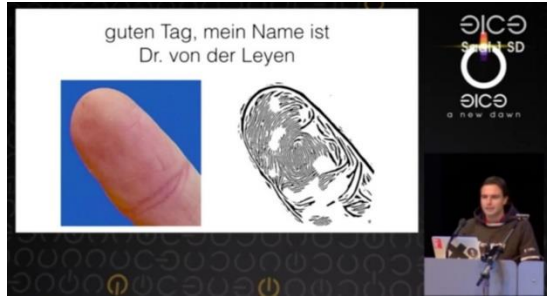
Transforming acoustic characteristics to deceive playback spoofing countermeasures of speaker verification systems

Fuming Fang, Junichi Yamagishi, Isao Echizen,
MD Sahidullah, Tomi Kinnunen



Background

Presentation attacks have been carried out with:



https://www.gizmodo.jp/2015/01/post_16271.html

printed fingerprint



https://www.theregister.co.uk/2013/09/22/iphone_5_touchid_broken_by_chaos_computer_club/

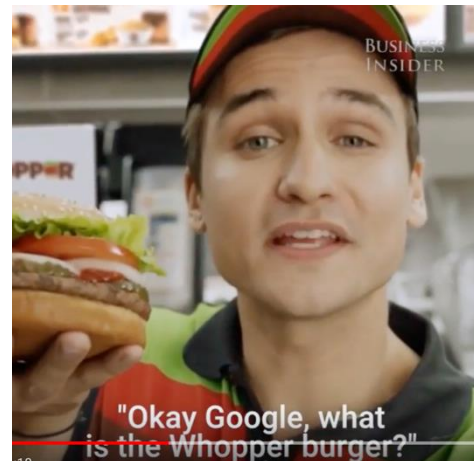


<https://www.youtube.com/watch?v=lOwuddvzl2A>

printed iris
(unlock a smart phone)



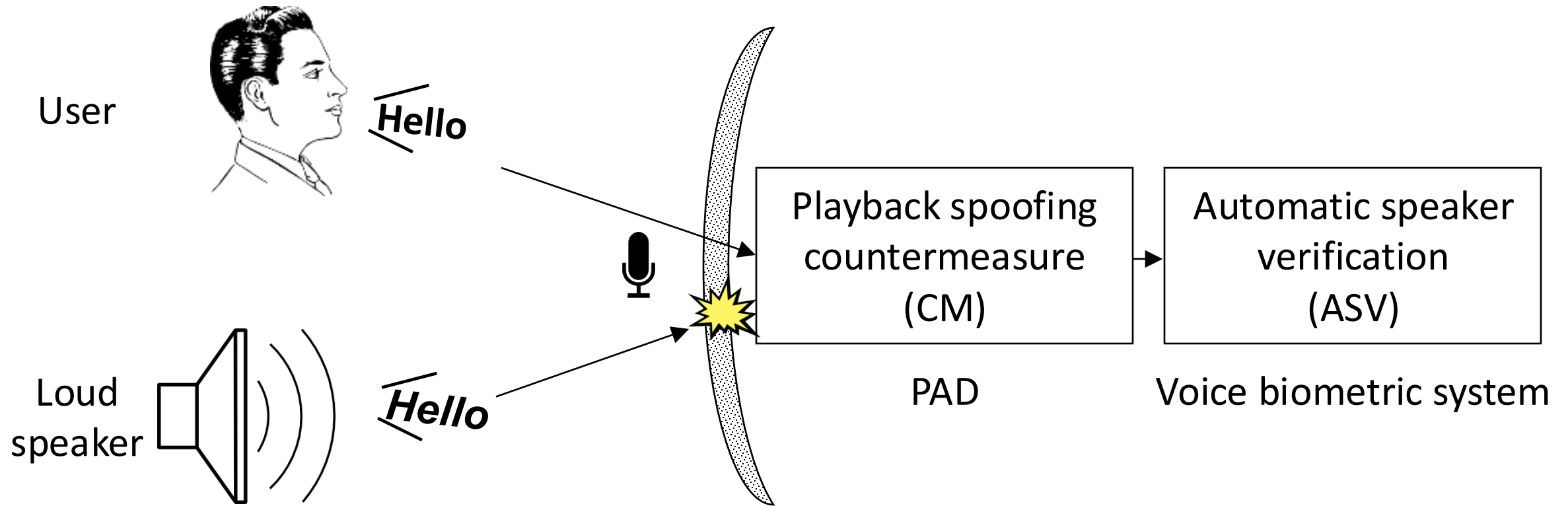
... also replayed voice



<https://www.youtube.com/watch?v=InOmTxmq1Ik>



Presentation attack detection and voice biometric system



ASVspoof challenge 2017:

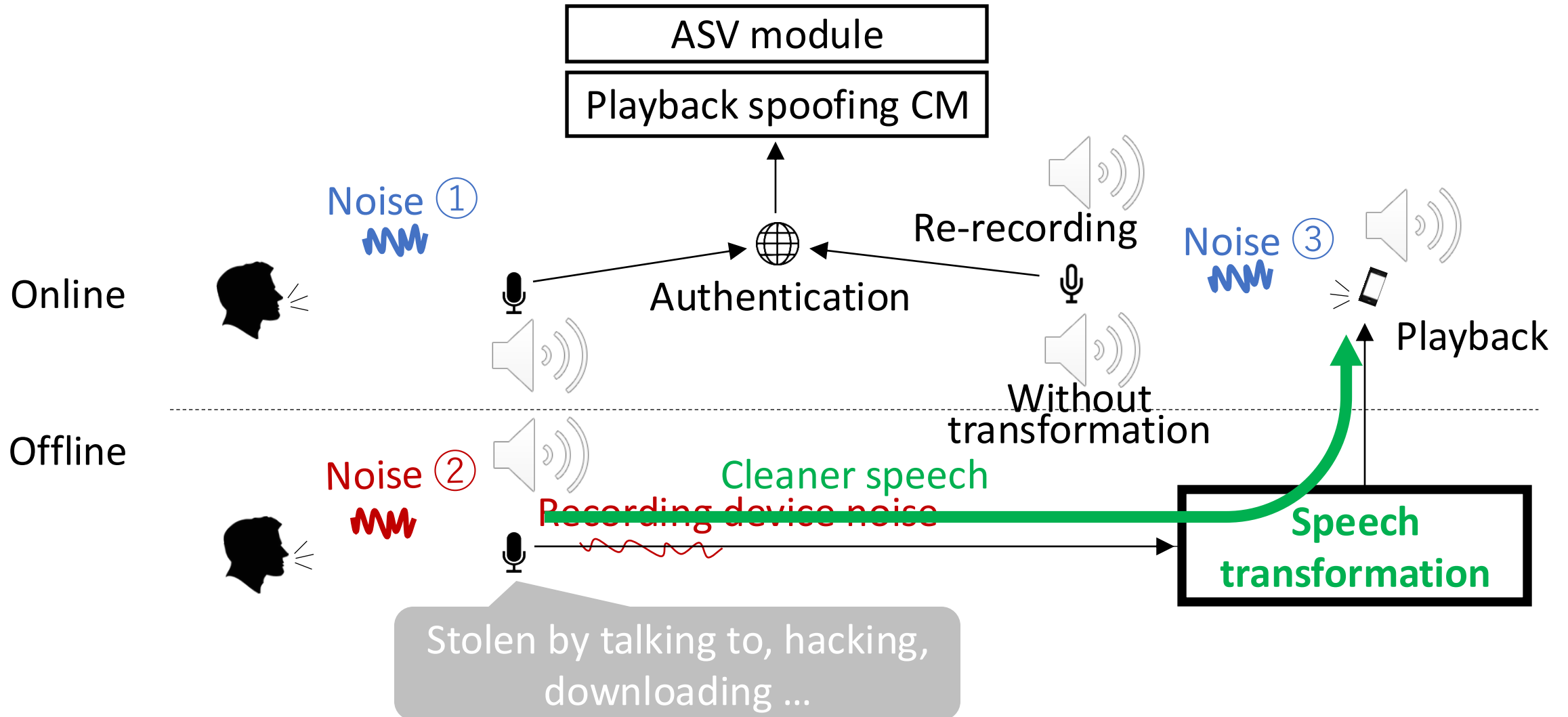
- A competition for replayed voice detection
- More than 40 teams joined
- Evaluation using the same database

Categories of the playback spoofing CM

Category	Example	Disadvantage
1 Random pass-phrase (challenge response)	Randomly promoting pass-phrase [T. Kinnunen '18, H. Zeinali '18]	Arbitrary phrase can be created if an attacker has sufficient data
2 Rule-based	Pop-noise exists? [S. Mochizuki '18]	Rules are difficult to design and implement
3 Audio fingerprinting	Incoming recording = recordings used for authentication? [J. Gonzalez-Rodriguez '18]	One billion users \times #of test trials
4 Machine learning-based	Learning the difference between human and playback speech [C. Wang '16, T. Kinnunen '17, G. Lavrentyeva '17]	Assumption: attackers have no special knowledge

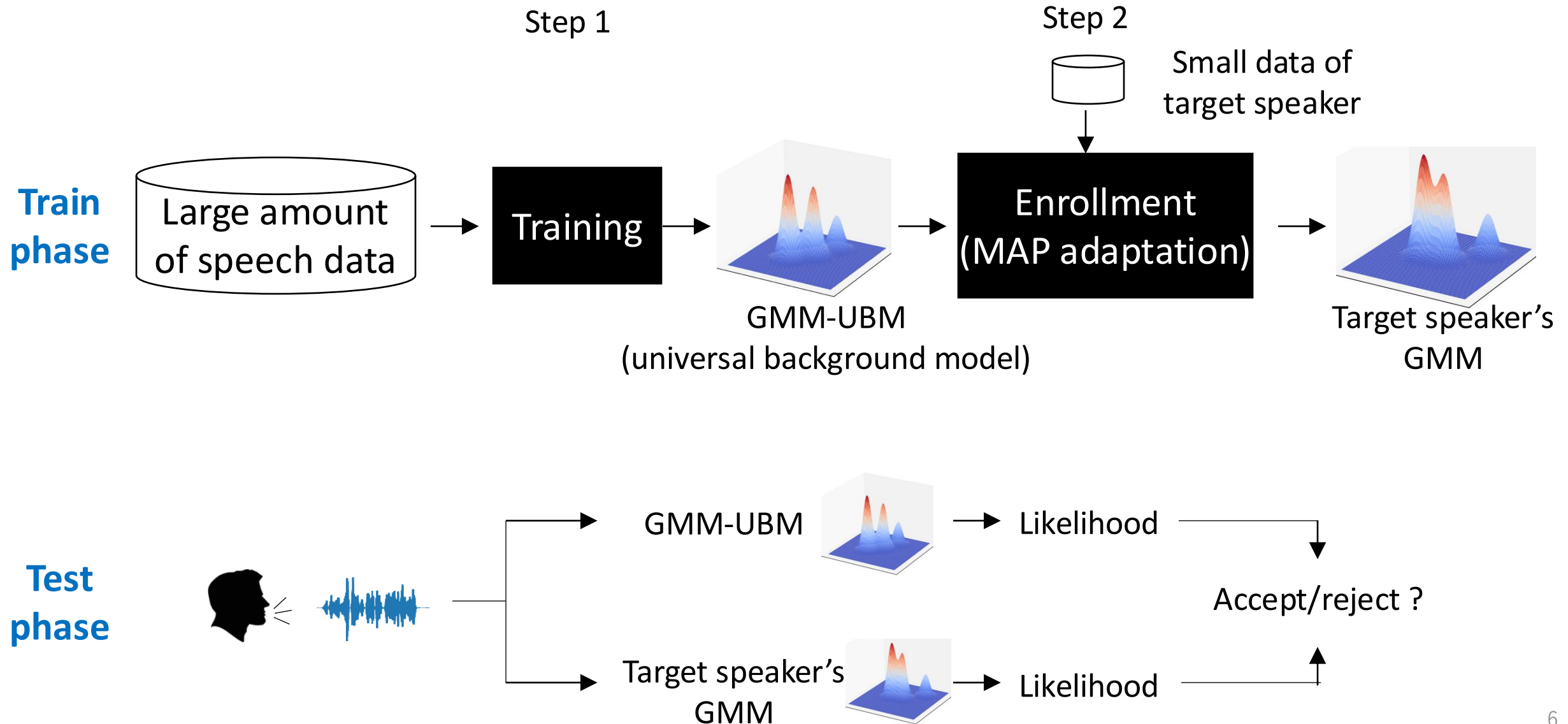
Proposed threat model

- Reduce environment noise and reverberation included in stolen speech
- Replay cleaner speech to the voice biometric system



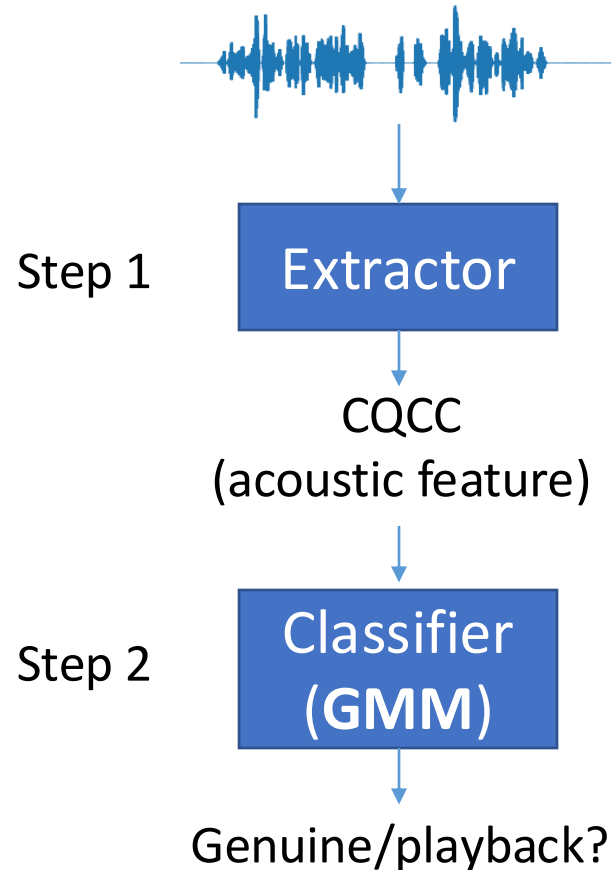
GMM-UBM-based ASV to be attacked

- A classical GMM-UBM-based method: suit for short duration utterance-based verification



Playback spoofing CMs to be attacked

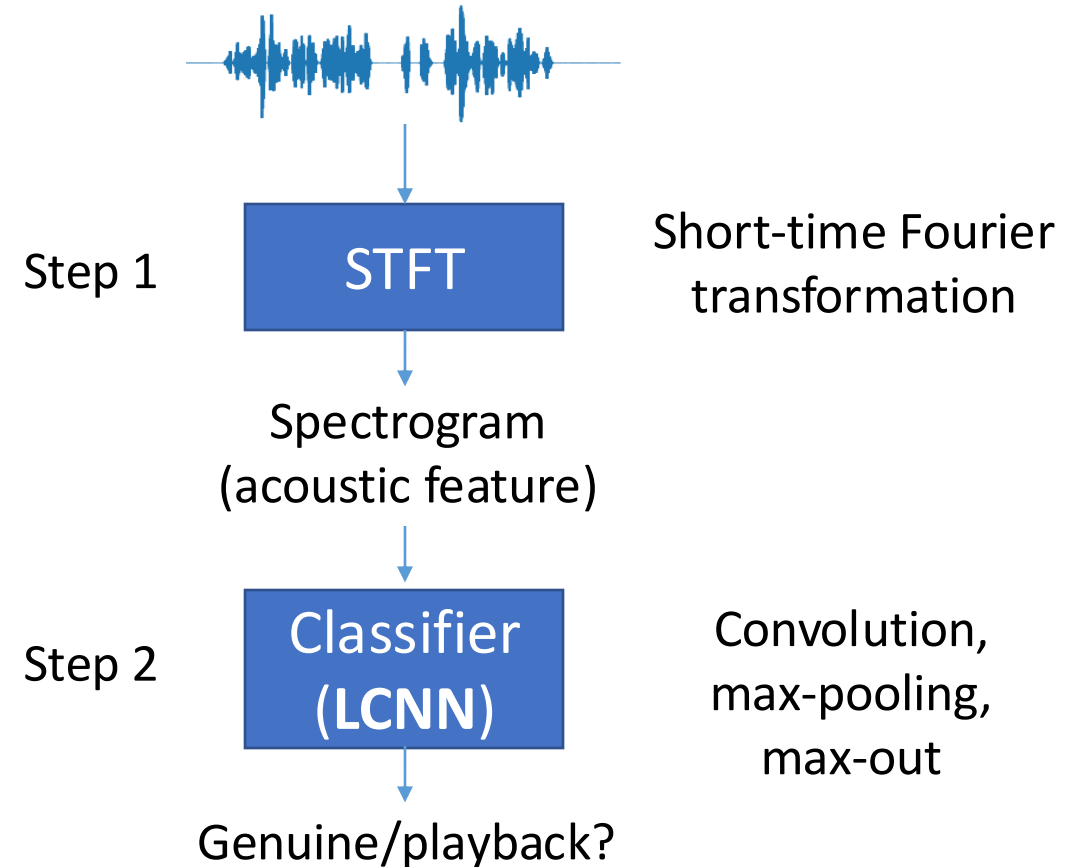
- CQCC+GMM-based method
(baseline of ASVspoof 2017)



EER = 30.60%

[T. Kinnunen et al., 2017]

- Light CNN (LCNN)-based method
(best method of ASVspoof 2017)

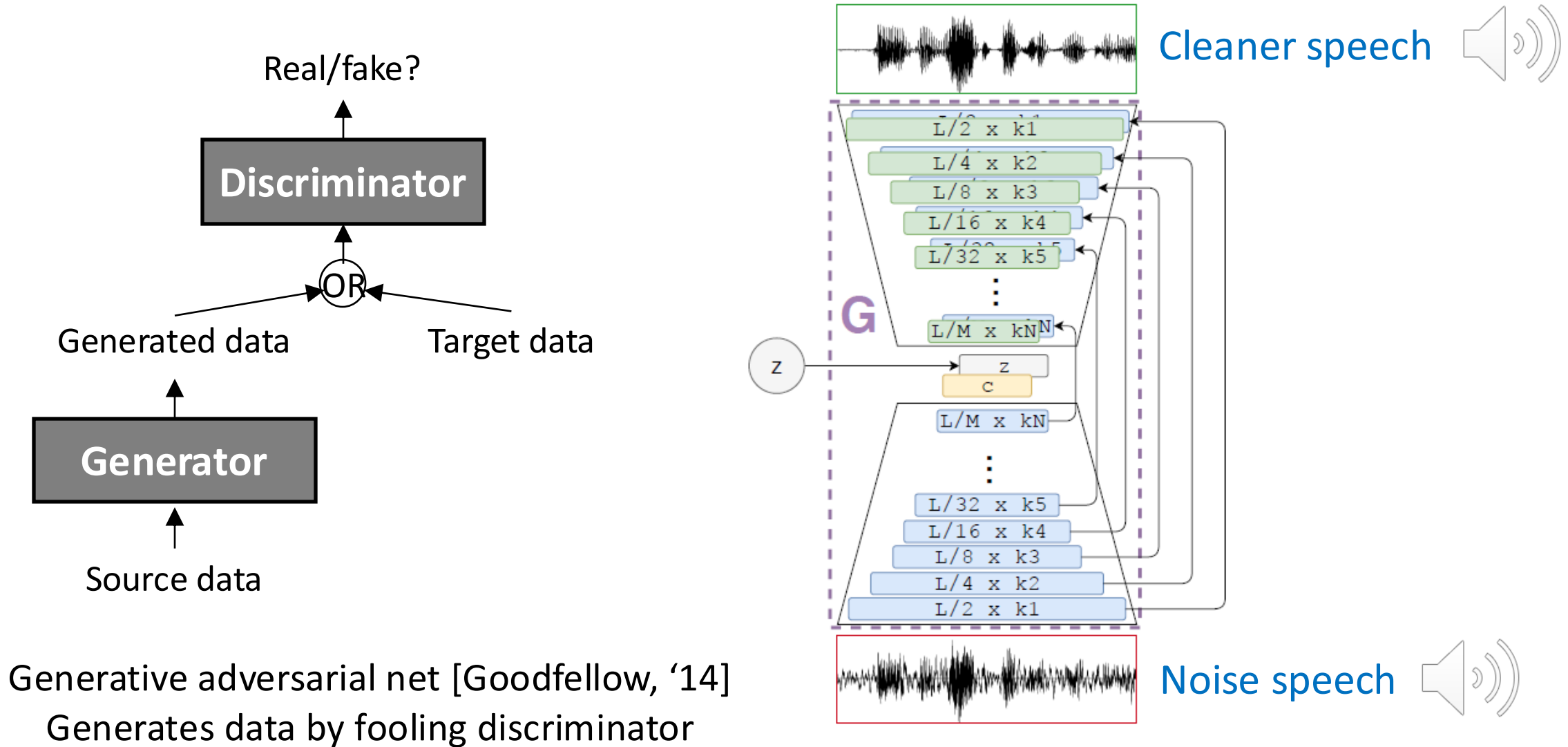


EER = 7.37%

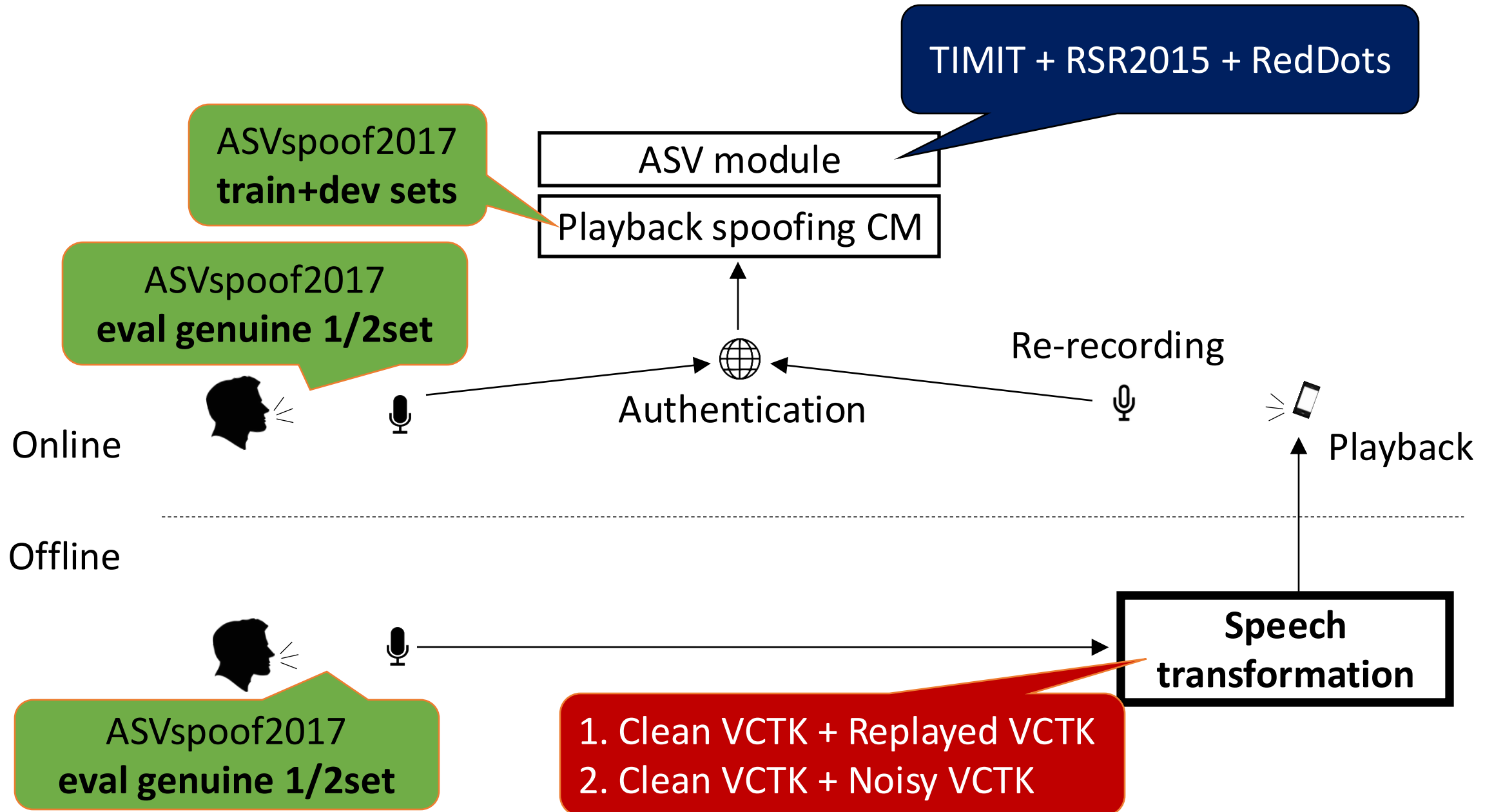
[G. Lavrentyeva et al., 2017]

Technique used by attackers: SEGAN

- Speech enhancement generative adversarial network (SEGAN) [S. Pascual et al., 2017]
- Originally proposed for end-to-end speech enhancement






Databases for training and evaluation



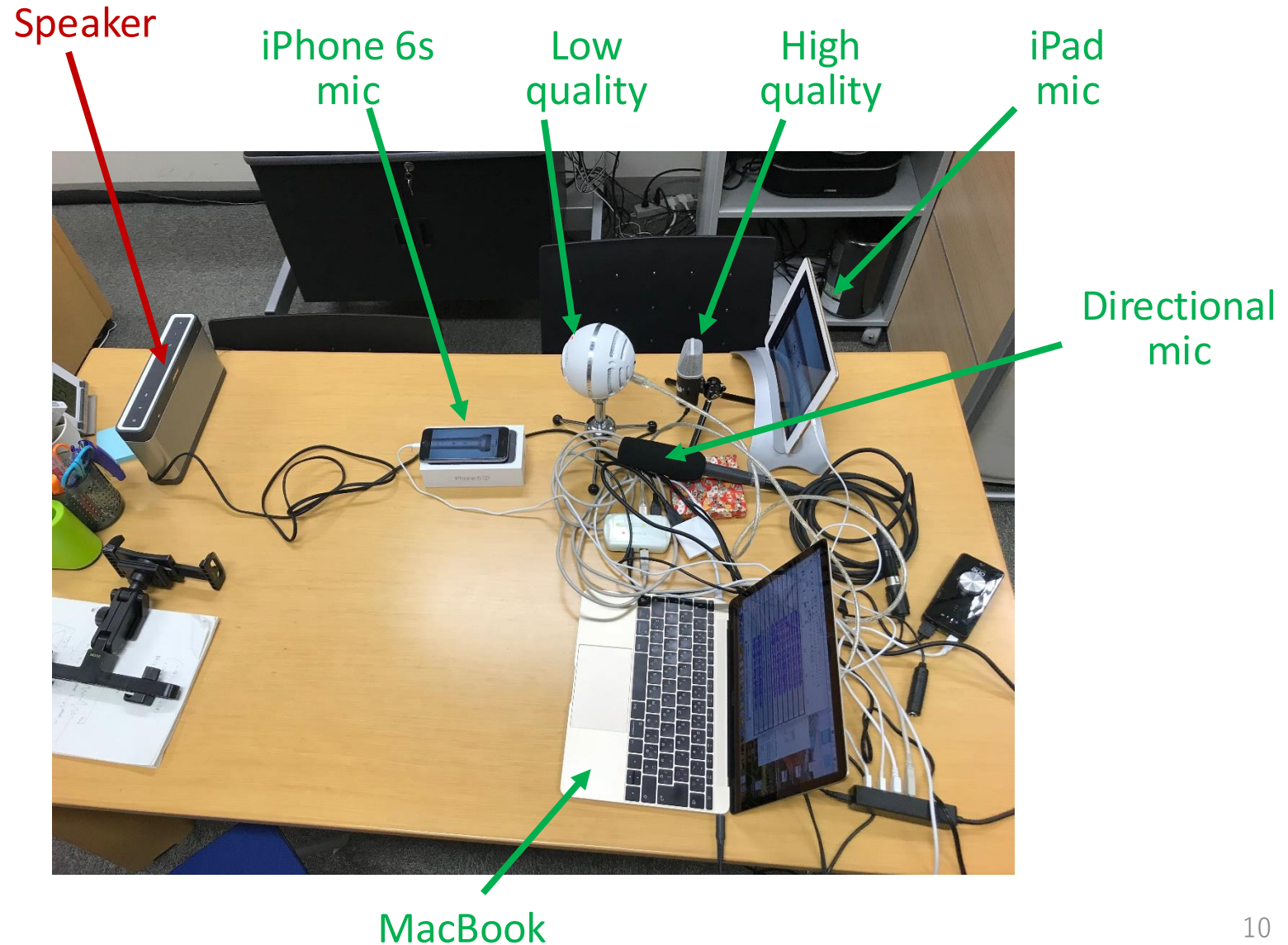
Playback attack setup

- Four types of loudspeakers and six types of microphones used for re-recording

Loudspeaker

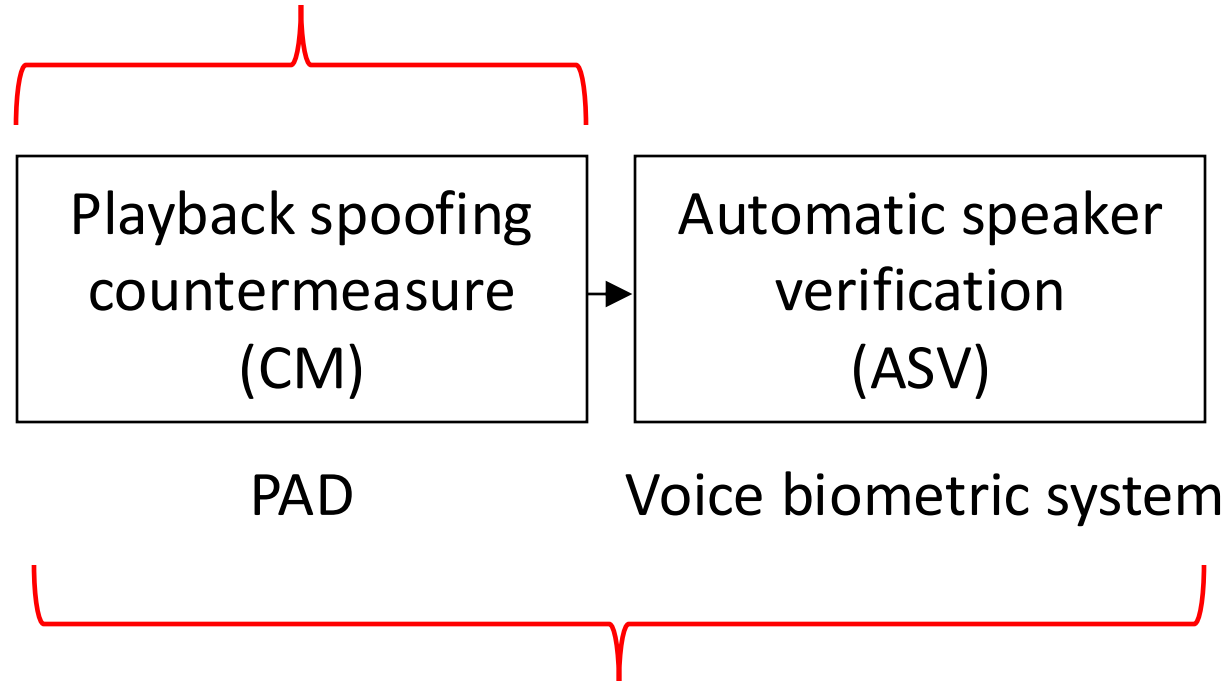
High quality	
Medium quality	
Low quality	
iPhone 6s	

* All is portable



Spoofing measures

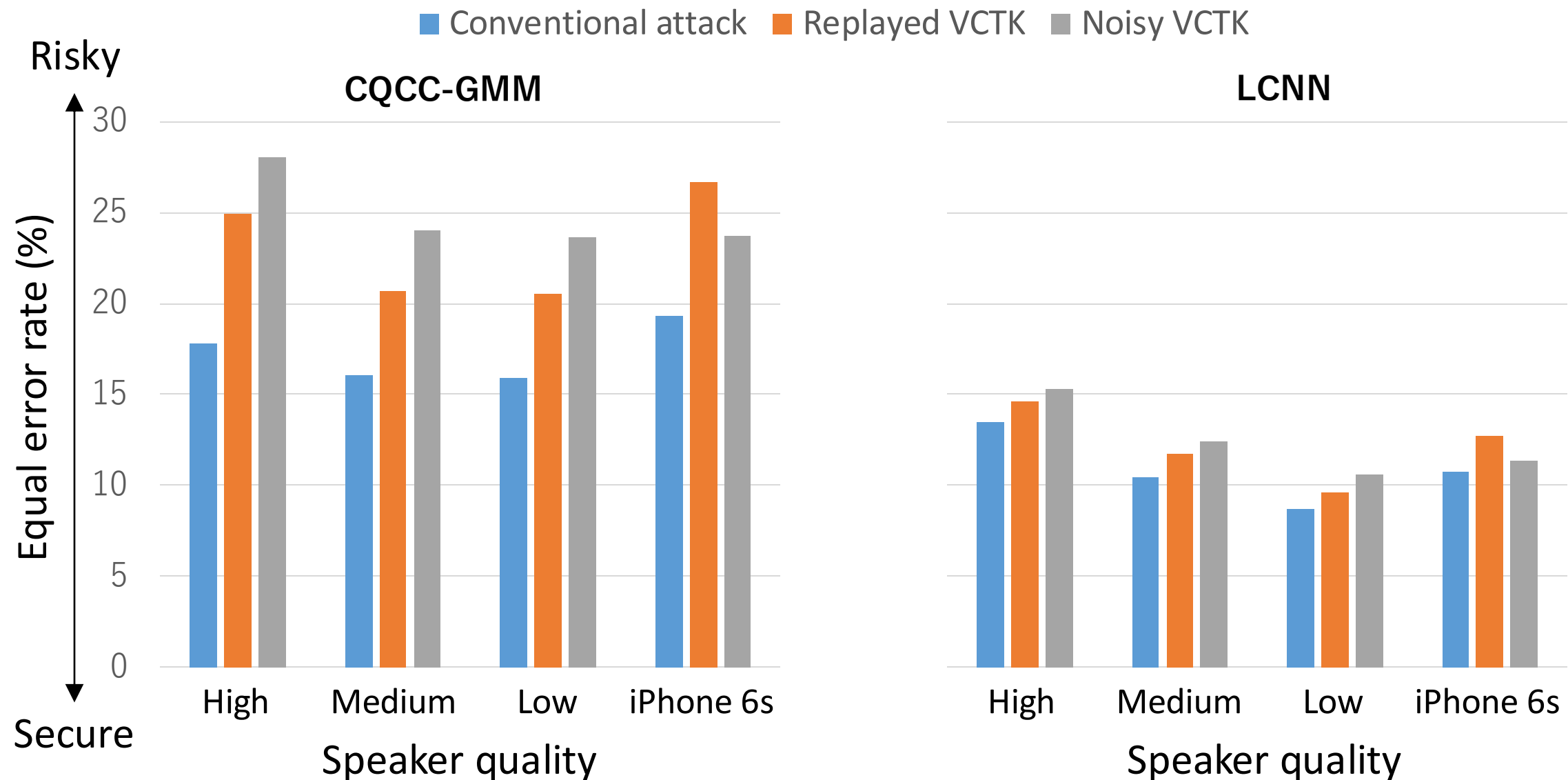
Case 1: Equal error rate (EER) measures CM



Case 2: “t-DCF” measures the whole system

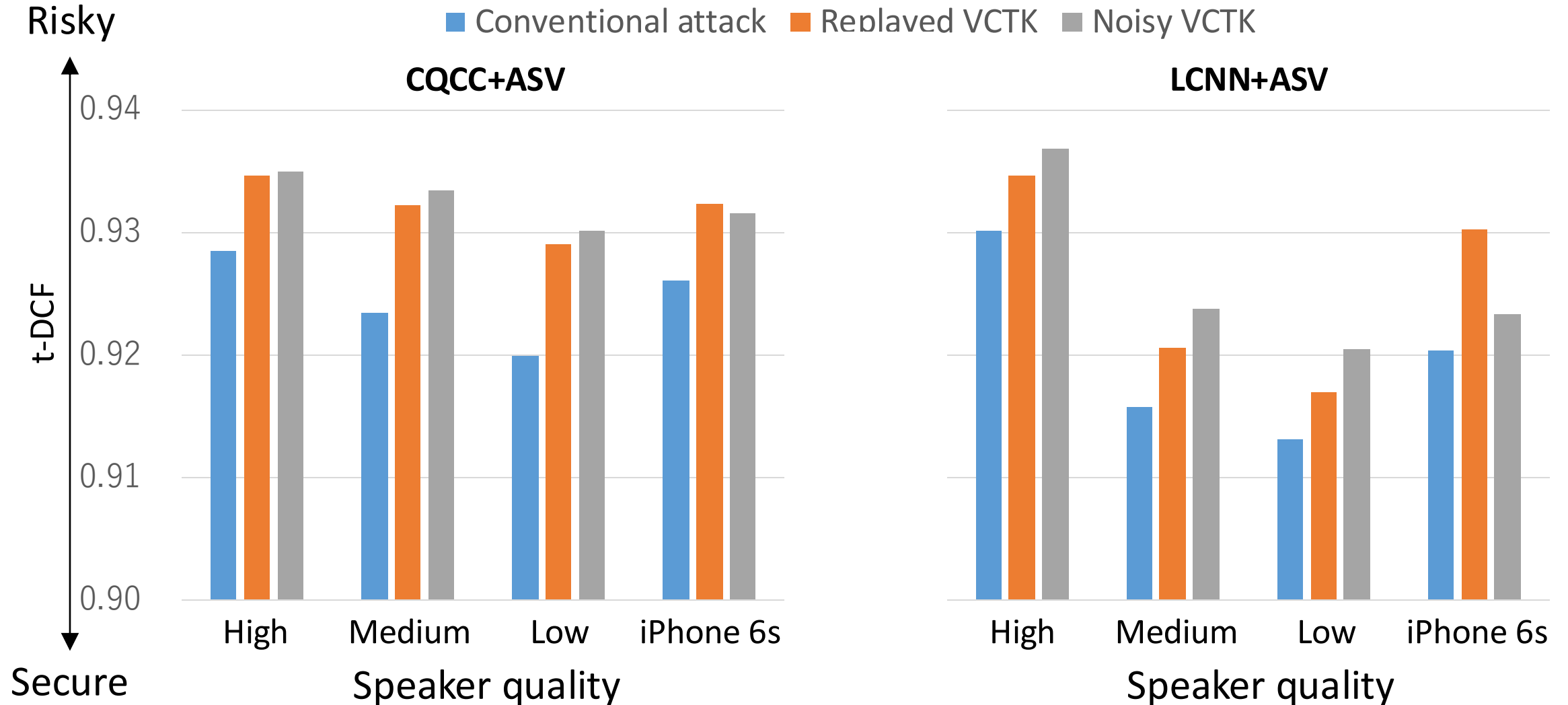
- t-DCF is a version of expanded DCF [T. Kinnunen et al., 2018]
- Considers both CM and ASV
- Higher t-DCF value = less reliable

Spoofing against CMs (EERs averaged across microphones)



The transformed voices have higher risks

Spoofing performance against CM&ASV (averaged t-DCF across microphones)



The transformed voices have higher risks

Conclusion and future work

- Proposed a playback attack method: pre-transforming speech before replay
- Increased EERs of both CQCC+GMM and LCNN-based playback spoofing CMs
- Increased t-DCF values obtained by playback spoofing CM and ASV system
- Plan to develop a robust CM against multiple transformation techniques