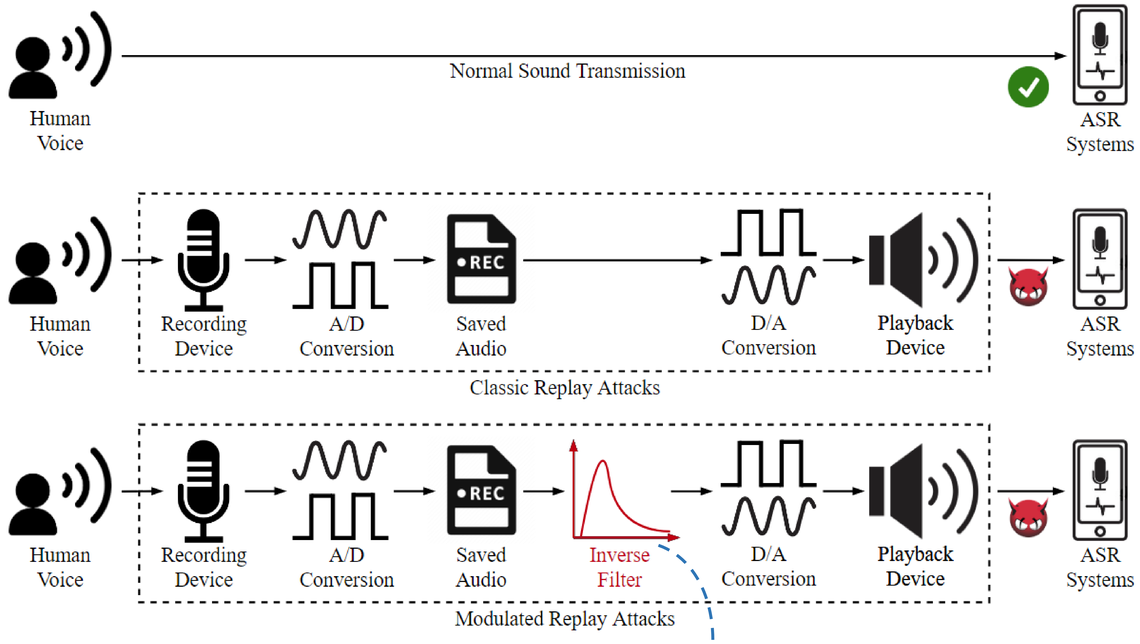
When the Differences in Frequency Domain are Compensated: Understanding and Defeating Modulated Replay Attacks on Automatic Speech Recognition

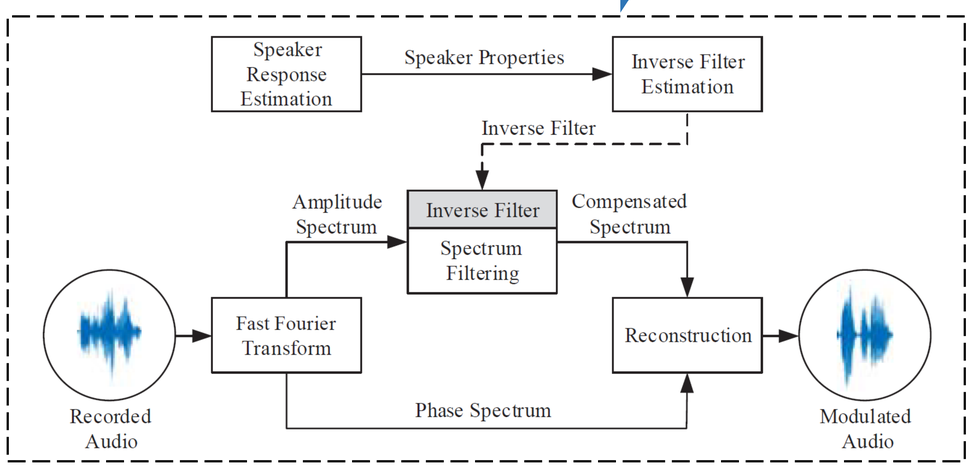
Shu Wang[1], Jiahao Cao[2], Xu He[1], Kun Sun[1], Qi Li[2]
{swang47, xhe6, ksun3}@gmu.edu, caojh15@mails.tsinghua.edu.cn, qli01@tsinghua.edu.cn
[1] Center for Secure Information Systems, George Mason University
[2] Institute for Network Sciences and Cyberspace, Tsinghua University

New Audio Replay Attack

- We propose the modulated replay attack that utilizes an **inverse filter** to compensate for the spectrum distortion brings from the loudspeaker so that the modulated replay audio can bypass all existing frequency-based defense due to the same frequency spectrum with genuine audio.



Attack overview.



The modulated processor.

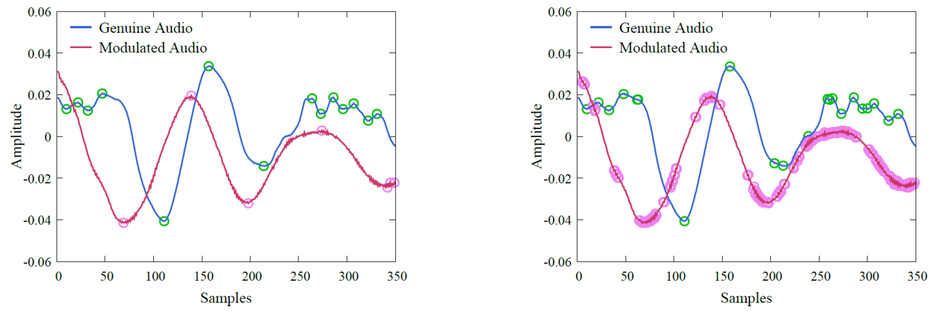
Defense Approach

- ***DualGuard*: Two-domain Defense**

  Key Insight: It is inevitable for any replay attacks to either leave *ringing artifacts* in the time domain or cause *spectrum distortion* in the frequency domain.

  - **Time-domain Defense**

  Distinguish **modulated replay audio** using the *local extrema ratio* with different granularity, which can detect the *ringing artifacts*.

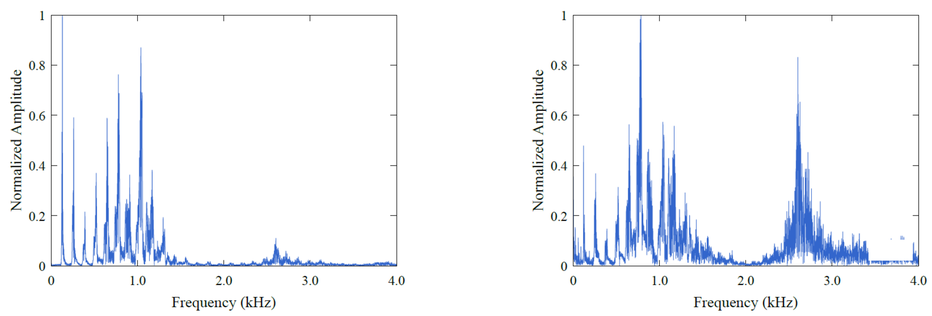


(a) Coarse granularity ($r = 10$) (b) Fine granularity ($r = 1$)

The local extrema under different granularity.

  - **Frequency-domain Defense**

  Distinguish **direct replay audio** using the area under the cdf curve of *spectral power distribution*, which can detect the *spectrum distortion*.



(a) Genuine audio (b) Direct replay audio

The frequency spectrum of genuine audio and replay audio.

- ***DualGuard*** achieves 90% accuracy for **direct replay audio** and 98% accuracy for **modulated replay audio** with different replay devices.