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

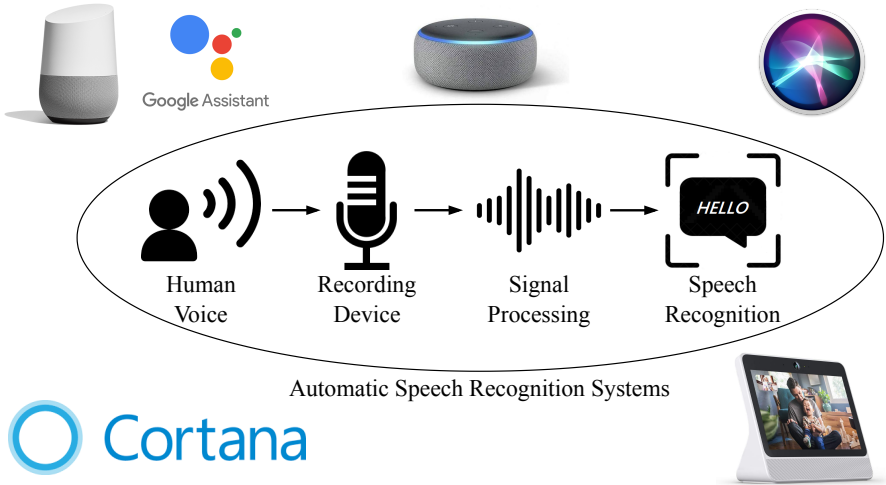
Shu Wang[1], Jiahao Cao[2], Xu He[1], Kun Sun[1], Qi Li[2]

[1] Center for Secure Information Systems, George Mason University
[2] Institute for Network Sciences and Cyberspace, Tsinghua University

清華大學
Tsinghua University

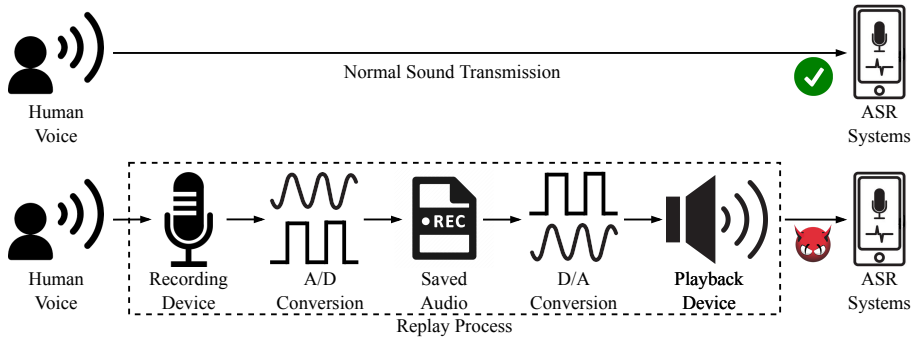
Table of Contents

Introduction



Replay Attack

- The most powerful and practical attacks on ASRs is audio replay attack.

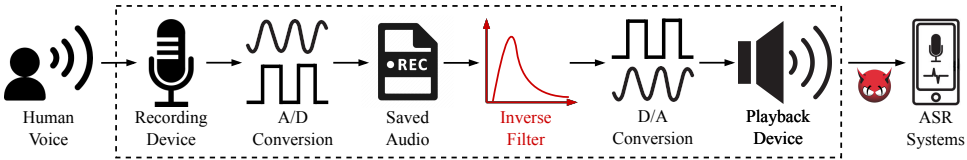


- Solution: Frequency feature detection (e.g., LPCC, MFCC, CQCC, MWPC).

Motivation

Is it possible to compensate for the effects of replay process?

Replay voice can have the same frequency features with human voice.



Modulated Replay Attack

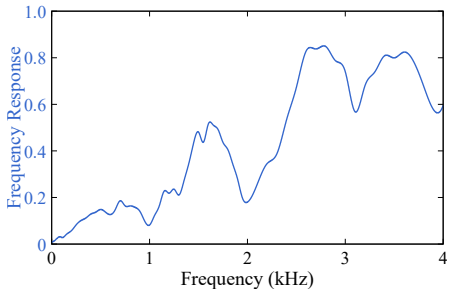
Effects of replay process can come from:

- Recording device - negligible (ambient noise, microphone non-linearity)
- A/D converter - negligible (sampling and quantization)
- D/A converter - negligible (low-pass filter)
- Playback device - significant (low-frequency response distortion)
  Amplitude response is a highpass filter with a cut-off frequency near 500 Hz.

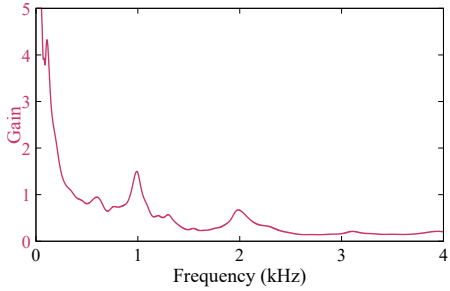
Method: design an **inverse filter** based on the loudspeaker amplitude response.

Modulated Replay Attack

1. Estimate Amplitude Response.



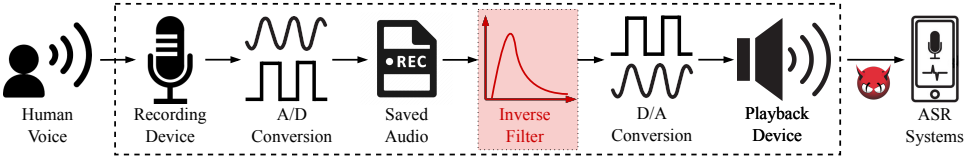
2. Construct Inverse Filter.

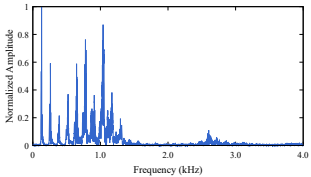


Amplitude responses of the inverse filter and the speaker can cancel each other.

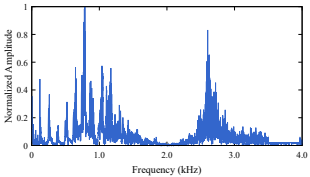
Modulated Replay Attack

3. Apply Modulation Processor.

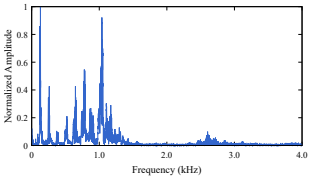




Genuine Audio



Replay Audio



Modulated Replay Audio

Modulated Replay Attack

Modulated replay attack can bypass existing frequency-based defense.

Table 1: The accuracy of different defense methods on detecting direct replay attacks and modulated replay attacks.

Detection Method	iPhone	iPad	Mi Phone	Google Nexus	BOSE	Samsung TV
CQCC	95.95% / 4.50%	95.51% / 6.31%	92.18% / 8.11%	89.93% / 2.25%	91.90% / 7.21%	95.51% / 6.76%
MFCC	90.99% / 15.51%	93.24% / 18.92%	89.64% / 24.32%	89.19% / 27.03%	91.89% / 29.73%	90.99% / 27.71%
LPCC	89.19% / 8.11%	87.84% / 9.91%	90.09% / 15.32%	86.03% / 18.92%	87.84% / 11.71%	90.54% / 11.26%
MWPC	95.05% / 46.85%	92.79% / 36.04%	90.99% / 53.15%	95.05% / 43.24%	100.0% / 50.45%	86.93% / 58.56%
Sub-band Energy	89.61% / 5.41%	89.22% / 4.50%	89.70% / 6.31%	88.61% / 10.81%	84.11% / 0.00%	85.57% / 0.90%
HF-CQCC	90.91% / 25.23%	90.91% / 22.52%	90.91% / 24.32%	90.08% / 18.02%	93.94% / 38.74%	93.94% / 11.71%
FM-AM	92.86% / 7.21%	92.86% / 17.12%	89.29% / 4.5%	92.86% / 9.91%	92.86% / 35.14%	96.43% / 12.61%
Sub-bass	99.10% / 7.66%	99.10% / 4.50%	98.20% / 5.80%	98.65% / 4.95%	96.85% / 6.76%	97.30% / 5.40%

DualGuard Defense

We propose a countermeasure DualGuard against the modulated replay attack.

Verified audio must pass two checks:

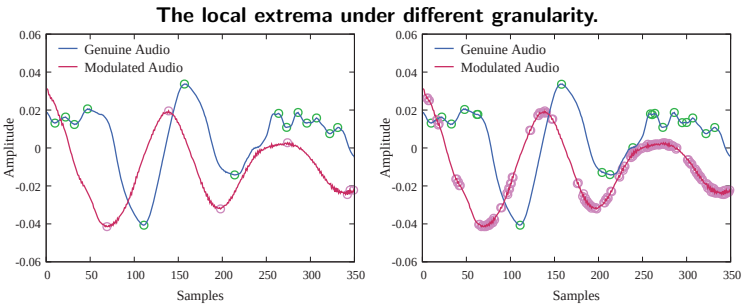
1. Time domain verification. (ringing artifacts patterns)
2. Frequency domain verification. (spectrum distortion patterns)

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.

DualGuard Defense

Time-domain Defense

Principle: Modulated replay audio will inevitably involve ringing artifacts.



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

Local extrema ratio (LER):
The ratio of the local extrema amount to the total signal length.

DualGuard Defense

Frequency-domain Defense

Principle: Spectrum distortion will lead to a different spectral power distribution.

Patterns: Cumulative density function of spectral power distribution.

$$A(n) = \sum_{i=0}^{n} D(i)$$
$$= \sum_{i=0}^{n} K^2(i) / \sum_{i=0}^{N-1} K^2(i).$$

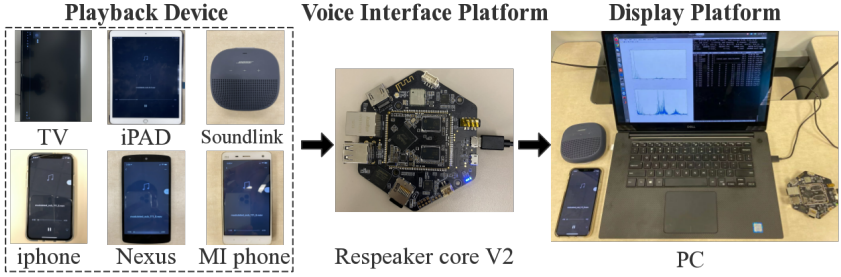
Algorithm 1 Frequency-Domain Replay Detection

Input: an audio signal $\boldsymbol{y}$, FFT point numbers N, decision threshold A_{th}
Output: whether there is a classical replay attack

/* *Calculate Normalized Signal Power Spectrum* */
$\boldsymbol{K} \leftarrow FFT(\boldsymbol{y}, N)$
$p \leftarrow \sum_{i=0}^{N-1} \boldsymbol{K}_i^2$
for $i \leftarrow 0$ to $N-1$ **do**
    $\boldsymbol{D}_i = \boldsymbol{K}_i^2 / p$
/* *Calculate the CDF and its AUC* */
$\boldsymbol{A}_0 = \boldsymbol{D}_0$
for $i \leftarrow 1$ to $N-1$ **do**
    $\boldsymbol{A}_i = \boldsymbol{A}_{i-1} + \boldsymbol{D}_i$
$AUC = \sum_{i=0}^{N-1} \boldsymbol{A}_i / N$
/* *Identify Classical Replay Attacks with AUC* */
if $AUC < A_{th}$ **then**
    output *replay attacks*
else
    output *genuine audio*

Evaluation

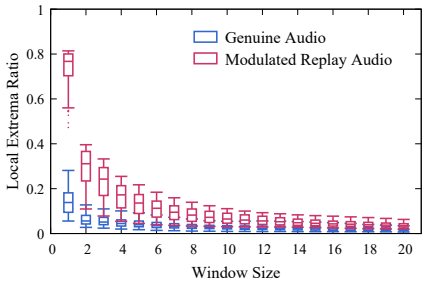
- Construct dataset containing replay audio and modulated replay audio.
- Implement DualGuard prototype in ReSpeaker core V2.
- Test 6 playback devices (i.e., iPhone X, iPad Pro, Mi Phone 4, Google Nexus 5, Bose Soundlink Micro, and Samsung UN65H6203 Smart TV).



Evaluation

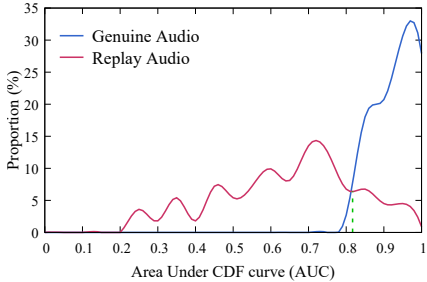
Performance of Dual-domain Defense

- **Time-domain Defense**



Local extrema patterns with different granularity.

- **Frequency-domain Defense**



The AUC distribution with the decision threshold.

Evaluation

Performance of DualGuard

- **DualGuard Performance**

Table 2: The accuracy of DualGuard on detecting direct replay attacks and modulated replay attacks.

Playback Device	Direct Replay	Modulated Replay
iPhone	91.00%	98.88%
iPad	90.54%	98.32%
Mi Phone	89.19%	97.75%
Google Nexus	90.45%	98.22%
BOSE	90.10%	97.79%
Samsung TV	89.64%	99.65%

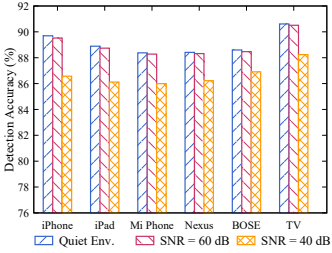
- **Overhead**

Processing time: 5.5 ms for 32 ms-length signal.
CPU usage[†]: 24.2%.
Memory usage: 12.05 MB.

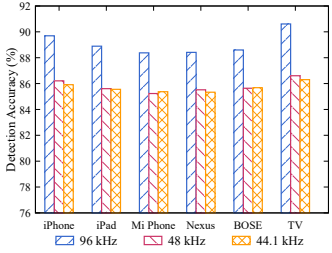
[†] Tested with C++ language in ReSpeaker Core v2 with quad-core ARM Cortex-A7 of 1.5GHz and 1GB RAM on-board.

Discussion

- Genuine audio sampling rate has no impact on DualGuard performance.
- Different recording devices have no impact on DualGuard performance.
- Noise conditions have limited impact on DualGuard performance.
- Higher ASR sampling rate can increase the detection accuracy.



accuracy vs. noise level.



accuracy vs. ASR sampling rate.

Conclusion

1. We propose a new modulated replay attack against ASR systems, utilizing a software-based inverse filter to compensate for frequency distortion.
2. We design a novel defense system DualGuard to detect all replay attacks including the modulated replay attacks by two-domain verification.
3. We implement a prototype of DualGuard on a popular voice platform and demonstrate its effectiveness and efficiency with different factors.

Thank you!

Author:
Shu Wang, Jiahao Cao, Xu He, Kun Sun, Qi Li

Questions?
My Email: swang47@gmu.edu