2024 Network and Distributed System Security Symposium (NDSS)

Compensating Removed Frequency Component: Thwarting Voice Spectrum Reduction Attacks

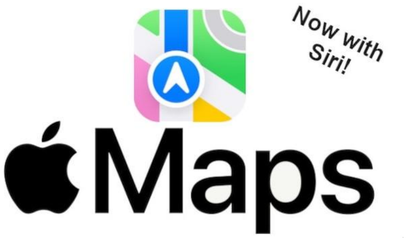
Shu Wang[1], Kun Sun[1], Qi Li[2]

[1] Center for Secure Information Systems, George Mason University
[2] Tsinghua University

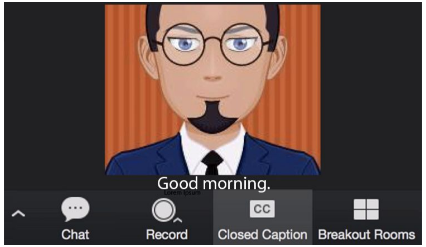
Introduction

- **Automated Speech Recognition (ASR)**
  - transcribe spoken language into text.
  - widely adopted in multiple areas.

smart home devices



navigation



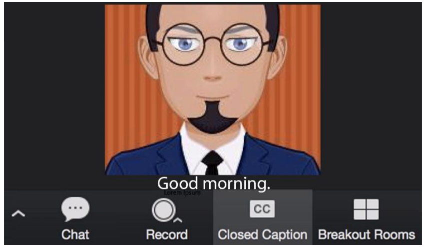
live closed captioning

Introduction

- **Automated Speech Recognition (ASR)**
  - transcribe spoken language into text.
  - widely adopted in multiple areas.

smart home devices

navigation



live closed captioning

- ASR is vulnerable to various malicious audio attacks.
  - frequency spectrum has been manipulated to achieve different attacking goals.

Spectrum-based Attacks

- **Spectrum Modification Attacks**
  - Attack: manipulating spectrum magnitude with a specific filter.
  - Defense: utilizing time-domain features.

Spectrum-based Attacks

- **Spectrum Modification Attacks**
  - Attack: manipulating spectrum magnitude with a specific filter.
  - Defense: utilizing time-domain features.
- **Spectrum Addition Attacks**[1]
  - Attack: adding high frequency components out of voice band.
  - Defense: using band-pass filters.

[1] NDSS 2019: Practical hidden voice attacks against speech and speaker recognition systems.

Spectrum-based Attacks

- **Spectrum Modification Attacks**
  - Attack: manipulating spectrum magnitude with a specific filter.
  - Defense: utilizing time-domain features.
- **Spectrum Addition Attacks**[1]
  - Attack: adding high frequency components out of voice band.
  - Defense: using band-pass filters.
- **Spectrum Reduction Attacks**[2]
  - Attack: removing spectrum magnitude under a threshold.
  - Defense: no effective methods due to the information loss.

[1] NDSS 2019: Practical hidden voice attacks against speech and speaker recognition systems.
[2] S&P 2021: Hear "No Evil", See "Kenansville": Efficient and Transferable Black-Box Attacks on Speech Recognition and Voice Identification Systems.

Spectrum Reduction Attack

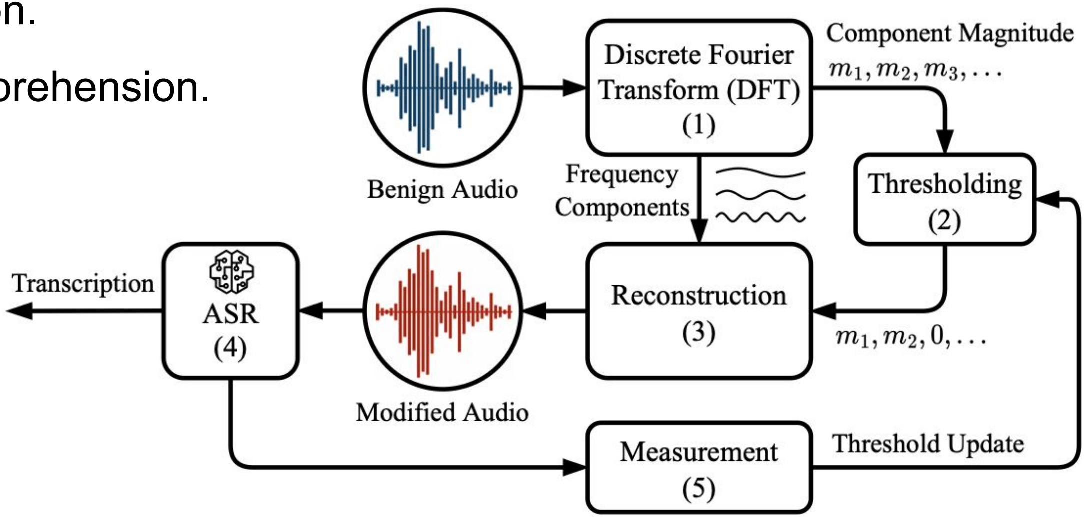
Hypothesis: some speech components are

- essential for ASR interpretation.
- non-essential for human comprehension.

Method: remove components with low magnitude.

Impact: modified audio

- can be recognized by humans.
- cannot be interpreted by ASRs.



Workflow of spectrum reduction attack.

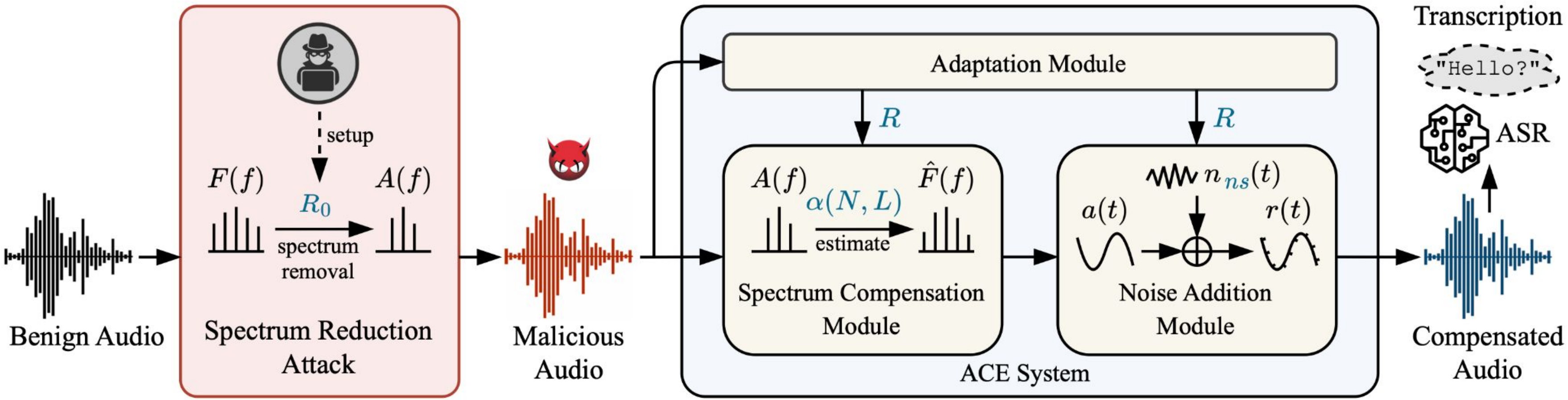
Impact of Spectrum Reduction Attack

- Content moderation systems in social media platforms
  - pre-screen and filter out harmful content (e.g., misinformation, violence).

- Malicious influencers can post and spread the videos and audios containing restricted speeches to online users without triggering any content alerts.

- The sensitive content within the audio tracks
  - cannot be noticed/detected by machine-based detection.
  - can be perceived by public audiences.

Acoustic Compensation System (ACE)

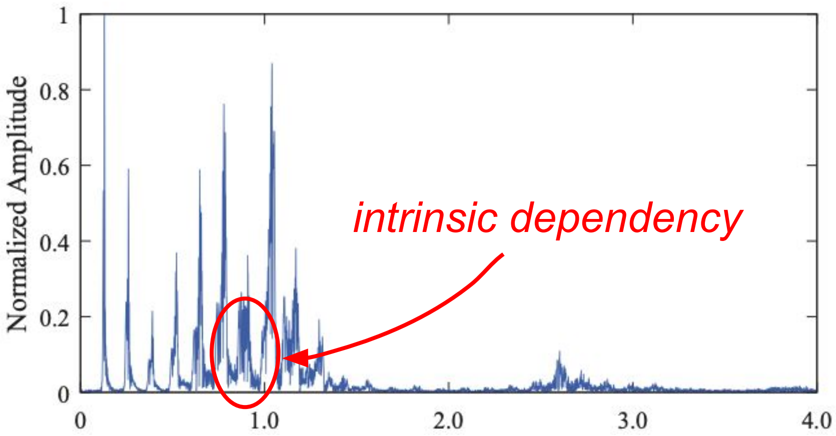


ACE consists of three modules:

- spectrum compensation module - recover missing components.
- noise addition module - improve voice recognition robustness.
- adaptation module - estimated attack parameters and adjust system parameters.

(1) Spectrum Compensation Module

Objective: recover the deleted components based on the existing ones.



Observation: frequency components with similar frequencies have high correlations.

Hypothesis:

- spectrum leakage caused by signal truncation in the DFT computation.
- aliasing caused by signal undersampling (only in low-sampling-rate devices).

(1) Spectrum Compensation Module

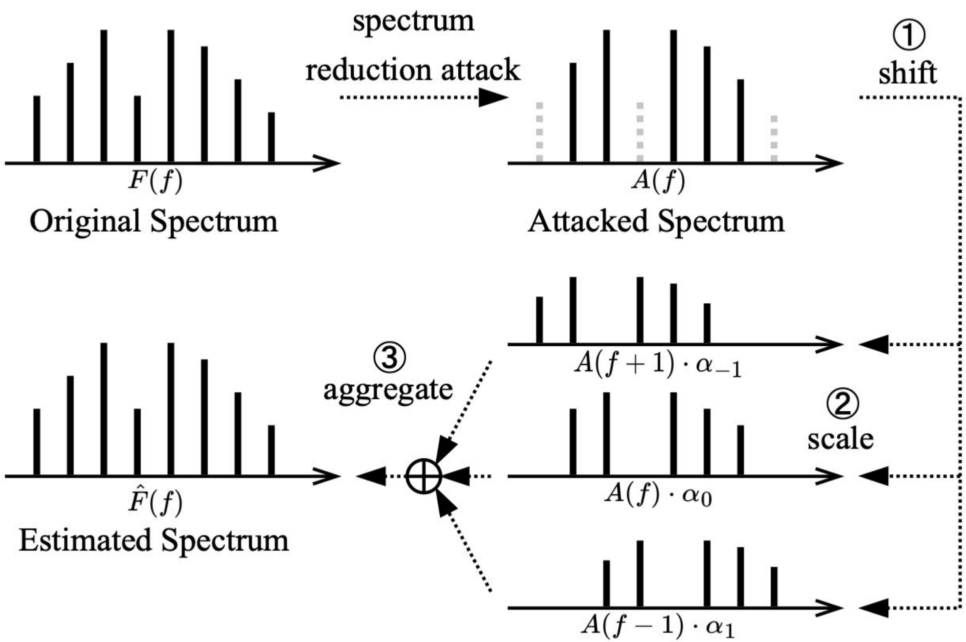
Proposed Method:

(1) shift attacked spectrum $A(f)$ by $i(-L \leq i \leq L)$ DFT units.

(2) scale shifted spectrum $A(f-i)$ with a scaling factor α_i.

(3) reconstruct $\hat{F}(f)$ by aggregating all shifted spectrums.

$$\hat{F}(f) = \sum_{-L \leq i \leq L} \alpha_i \cdot A(f-i)$$



(1) Spectrum Compensation Module

$$\hat{F}(f) = \sum_{-L \le i \le L} \alpha_i \cdot A(f-i) \qquad (0 \le f \le N-1)$$

Matrix form with a Hanker matrix:

$$\begin{bmatrix} A(-L) & A(-L+1) & .. & A(L-1) & A(L) \\ A(-L+1) & A(-L+2) & .. & A(L) & A(L+1) \\ .. & .. & .. & .. & .. \\ A(-L+N-2) & A(-L+N-1) & .. & A(L+N-3) & A(L+N-2) \\ A(-L+N-1) & A(-L+N) & .. & A(L+N-2) & A(L+N-1) \end{bmatrix} \cdot \begin{bmatrix} \alpha_{-L} \\ \alpha_{-L+1} \\ .. \\ \alpha_{L-1} \\ \alpha_L \end{bmatrix} = \begin{bmatrix} F(0) \\ F(1) \\ .. \\ F(N-2) \\ F(N-1) \end{bmatrix}$$

$$H \cdot \alpha = F$$

We can get the scaling factors with closed-form linear regression:

$$\alpha = (H^T \cdot H)^{-1} \cdot H^T \cdot F$$

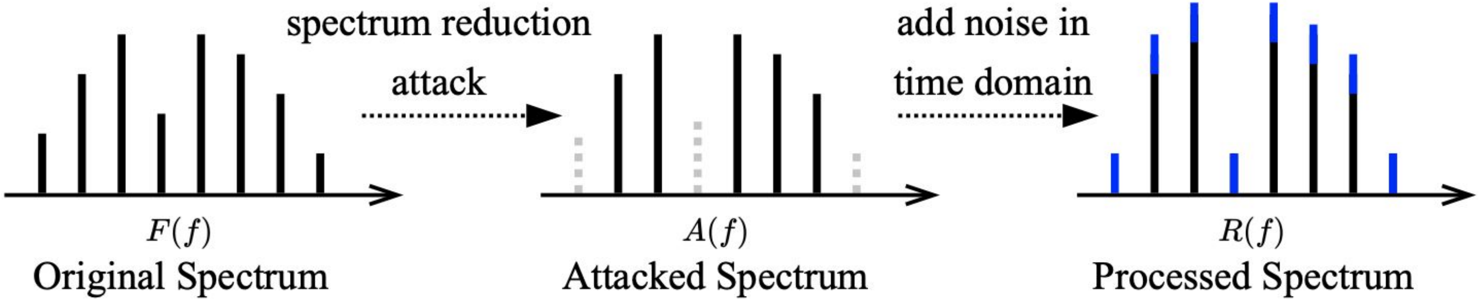
(2) Noise Addition Module

Objective: add Gaussian noise to the time-domain modified signals.

$$r(t) = a(t) + n_{ns}(t)$$

Weak noise effects:

- fill in the positions of missing weak components.
- not essentially change the distribution for strong components.



(2) Noise Addition Module

Hypothesis:

- the removed components can be seen as special adversarial noise,

$$n_{adv}(f) = -\sum_{f \in S_f} |m_f \cdot e^{j(2\pi f + \phi_f)}|$$

  whose effect is to counteract the weak components in the frequency domain.

- $n_{adv}(f)$ has a similar property with Gaussian noise of a limited intensity.
    - $n_{adv}(f)$: all magnitude are weak and under a small threshold.
    - Gaussian noise: all magnitude are equal to a specific value (threshold).

(3) Adaptation Module

Problems:

- defenders do not know the spectrum reduction ratio (R) used by attackers.
- system parameters (e.g., noise level, scaling coefficients) are related to R.

Solutions:

- estimate R in the received audio to adaptively optimize the parameters of modules.
- calculate the ratio of extremely weak components among the whole spectrum
  (i.e., magnitude is less than 0.2% of the max magnitude).

ACE Evaluation

- **Speech Datasets**:
  - TIMIT: 6,300 samples; English dialects; 16 kHz.
  - VCTK: 44,000 samples; multi-accent; 48 kHz.
- **ASR Models**:
  - DeepSpeech: support desktop, mobile, and embedded devices.
  - CMU Sphinx: designed for low-resource platforms.
- **Evaluation Metrics**:
  - WER/CER (i.e., Word/Character Error Rate)
  - WER/CER Reduction Rate

ACE Evaluation

TABLE I: The performance of ACE and its each module against the word-level/phoneme-level spectrum reduction attacks (component removal ratio is 0.85). We evaluate both the WER and CER for the attacked audio and the audio with defense.

Dataset	Attack Granularity	Evaluation Metric†	Baseline Error‡	Error w/ Attack§	Error w/ Our Defense*		
					Compensation	Noise Addition	ACE
TIMIT	phoneme-level	WER	0.217	0.597	0.336 (-68.7%)	0.322 (-72.4%)	0.314 (-74.5%)
		CER	0.107	0.386	0.203 (-65.6%)	0.190 (-70.3%)	0.187 (-71.3%)
	word-level	WER	0.217	0.794	0.593 (-34.8%)	0.570 (-38.8%)	0.568 (-39.2%)
		CER	0.107	0.562	0.396 (-36.5%)	0.372 (-41.8%)	0.370 (-42.2%)
VCTK	phoneme-level	WER	0.487	0.897	0.576 (-78.3%)	0.641 (-62.4%)	0.571 (-79.5%)
		CER	0.375	0.705	0.419 (-86.7%)	0.465 (-72.7%)	0.415 (-87.9%)
	word-level	WER	0.487	0.885	0.691 (-48.7%)	0.714 (-43.0%)	0.686 (-50.0%)
		CER	0.375	0.688	0.511 (-56.5%)	0.522 (-53.0%)	0.506 (-58.1%)

† WER: word error rate between labels and predictions; CER: character error rate between labels and predictions.
‡ Baseline Error indicates the average error rate when ASR infers original benign audio.
§ Error w/ Attack indicates the average error rate under spectrum reduction attack (including the baseline error).
* The percentage in parenthesis represents the reduction ratio to the errors caused by attacks.

Adaptive Attackers

- **Q:** Could attackers use time-varying component removal ratios to circumvent the defense if they are aware of the ACE defense system?

- ACE performance is stable due to the attackers' dilemma.
    - a smaller attack unit can increase the parameter changing frequency while decreasing the attack performance.

TABLE II: The performance of ACE system under a dynamic attack environment with different attack granularities.

attack unit (ms)	80	160	320	640	1280	2000	4000
CER w/ attack (%)	16.9	19.1	18.3	23.8	22.1	24.0	23.2
CER w/ ACE (%)	11.8	13.7	14.1	19.3	17.4	19.1	18.4
error reduction (%)	82.3	64.3	55.3	34.4	41.2	36.8	38.4

Residual Error Analysis

We find ASR recognition errors come from 6 types:

T1: Fast Speed (Elision) Errors

```
G: don't ask me to carry an oily rag like that.
T: to carry an oily rag like that.
```

T2: Rare Word Errors

```
G: iguanas and alligators are tropical reptiles.
T: quanta analogous are tropical reptiles.
```

T3: Consonant Errors

```
G: the one meat showing .. at .. doses is pork.
T: the one need showing .. and .. does is poor.
```

Residual Error Analysis

We find ASR recognition errors come from 6 types:

T4: Vowel Errors

```
G: will robin wear a .. showed pleasure.
T: well robin where a .. should pleasure.
```

T5: Shifted Phoneme Errors

```
G: the tooth fairy forgot to .. tooth fell out.
T: the two theories for that to .. to sell out.
```

T6: NLP Inference Errors

```
G: she had your dark suit in greasy wash water.
T: she had her dark suit and greasy wash water.
```

Residual Error Analysis

Benign audio

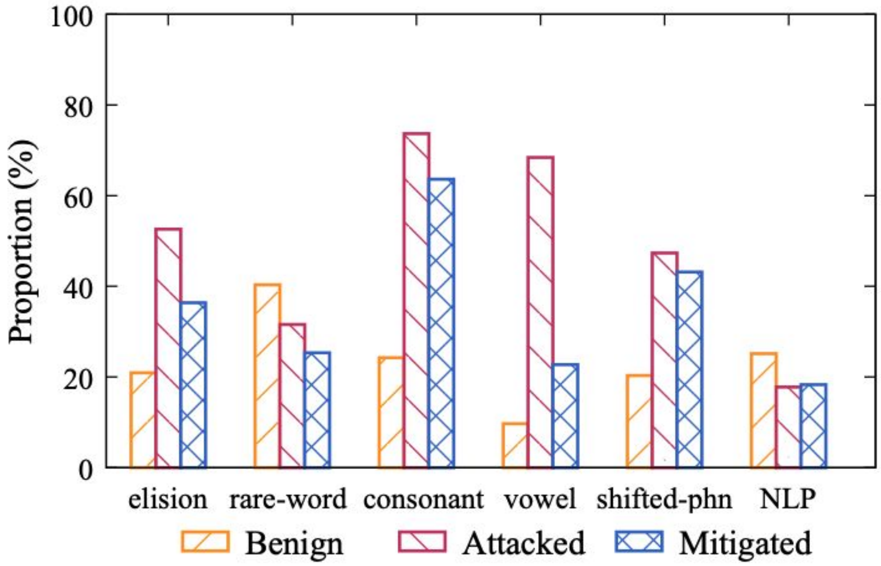
- rare word errors

Attacked Audio

- consonant & vowel errors

Mitigated Audio

- consonant errors



Reason:

- vowels are easier to recover due to higher loudness and signal strength.
- consonants are harder to recover due to light sounds and shorter durations.

Takeaways

- Mitigate spectrum reduction attacks:
    - spectrum compensation.
    - noise addition.
- ACE is stable to adaptive attacks due to attacker's dilemma.
- Residual error analysis:
    - audio attacks mainly generate phoneme errors.
    - vowels are easier to be recovered than consonants.

Thank you!

Questions and Comments?

Contact:

Email: shuvwang@gmail.com