SIEVE: Secure In-Vehicle Automatic Speech Recognition Systems

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

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

[2] Institute for Network Sciences and Cyberspace, Tsinghua University

Outline

Introduction

System Design

Experiments

Discussion

Conclusion

Background

Self-driving cars are becoming an irreversible trend in our daily lives.

- ➢ Tesla cars with Autopilot
- ➢ Waymo's driverless cars

The latest in-vehicle voice control system provides a convenient way for drivers and passengers to interact with driverless cars.

Motivation

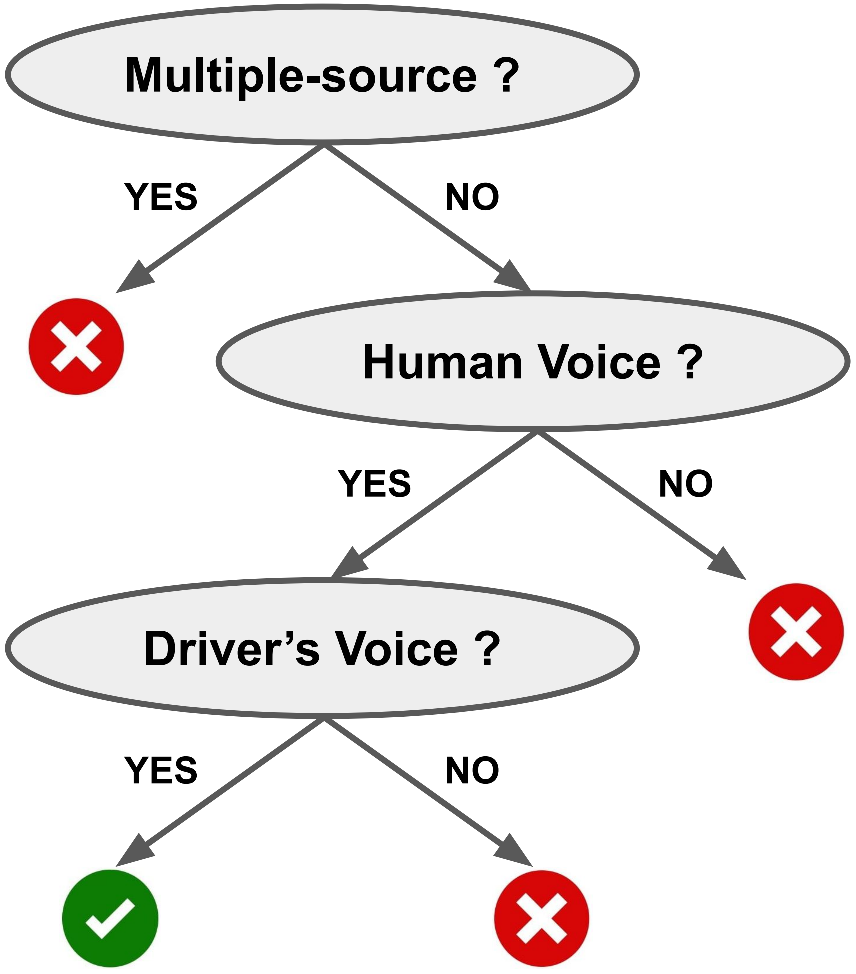
➢ Design a in-vehicle automatic speech recognition system to defeat various adversarial voice command attacks.

➢ Malicious commands may come from:

Our Work

➢ SIEVE: distinguish voice commands issued from a driver, a passenger, and non-human speakers.

➢ Three-step scheme:
- Detecting multiple speakers
- Identifying human voice
- Identifying driver’s voice



Outline

Introduction

System Design

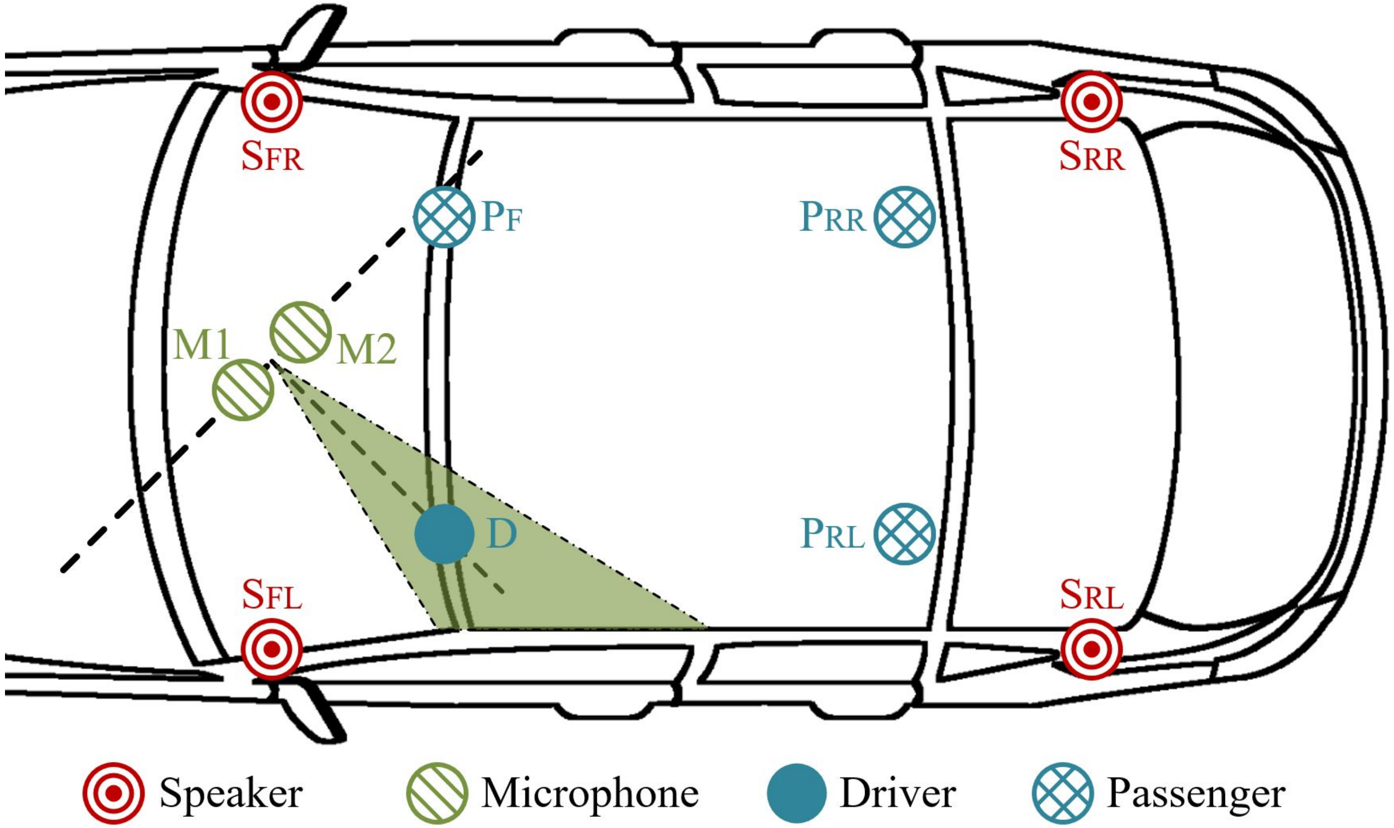
Experiments

Discussion

Conclusion

SIEVE System

Dual Microphone Scheme:



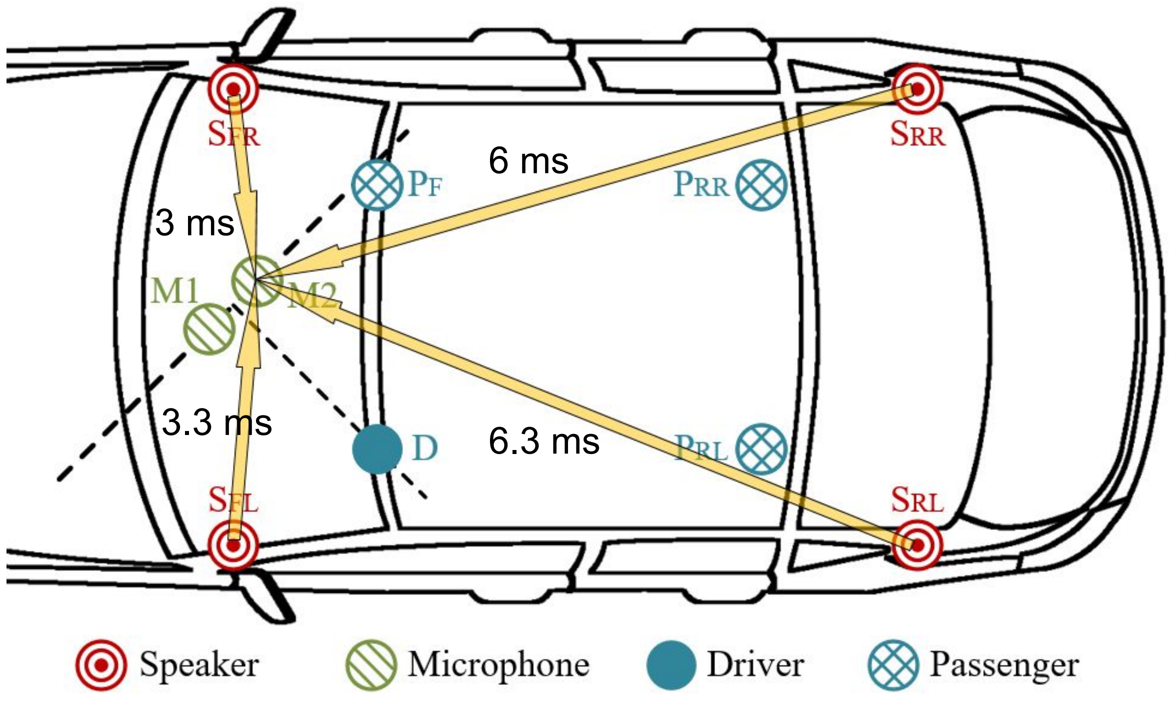
Step1: Detecting Multiple Speakers

Objective:
Filter out multiple-source voice commands.

Key features:
the overlap of the received signals will expand the signal correlations in the time domain.

Our methods:
Autocorrelation analysis

$$C(s) = \sum_{k=N}^{N+L-1} g(n) \cdot g(n+s), s \in [-S, S]$$



Step2: Identifying Human Voice

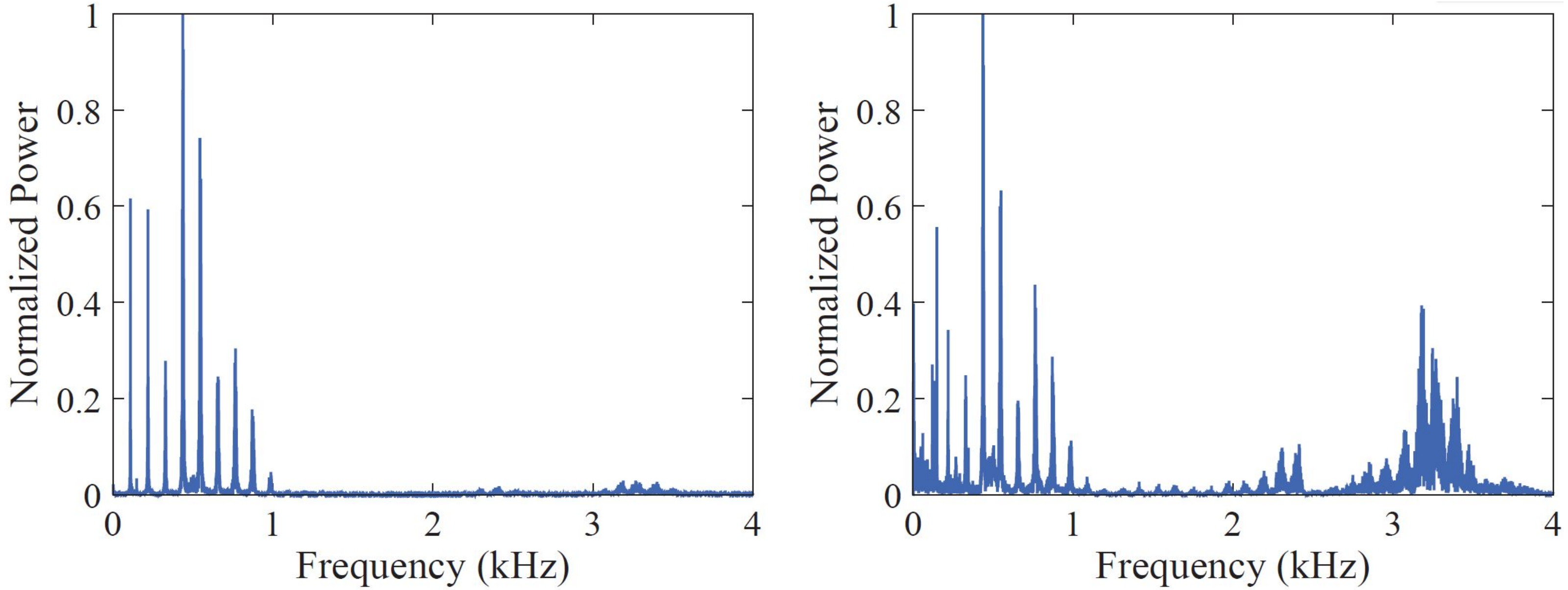
Objective:
Filter out non-human voice coming from car speakers and smartphone speakers in single-source commands.

Our method:
Voice must pass two checks:

- ➢ Frequency-domain power spectrum verification.
- ➢ Time-domain local extrema cross-check.

Frequency Domain Verification

Observation: timbre difference between human voice and replay voice.



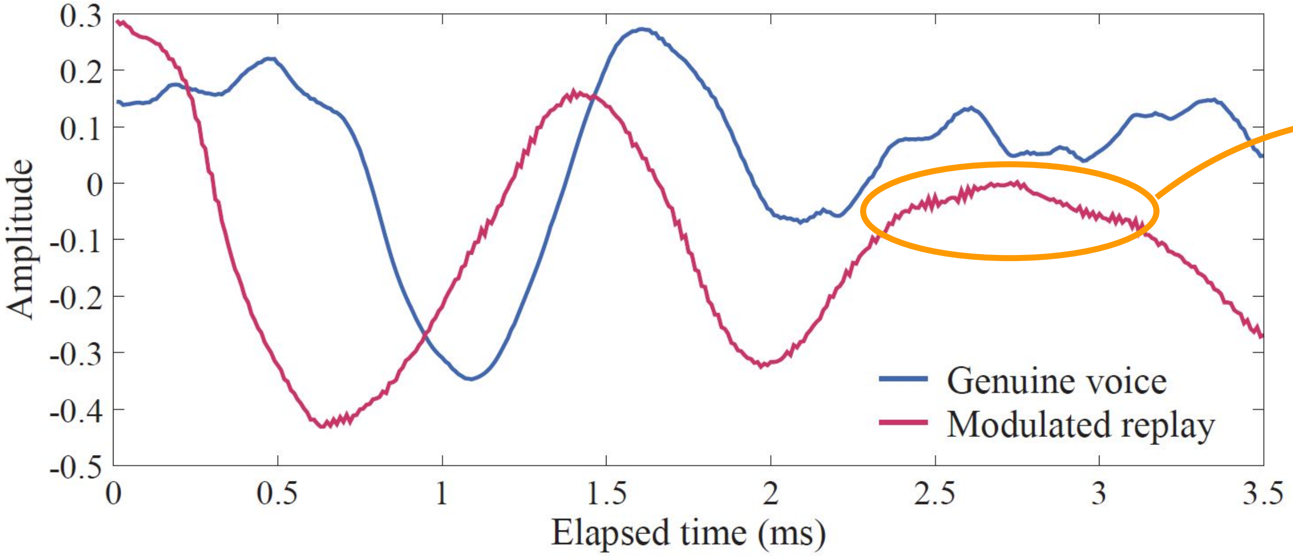
(a) The human voice (b) The replay voice

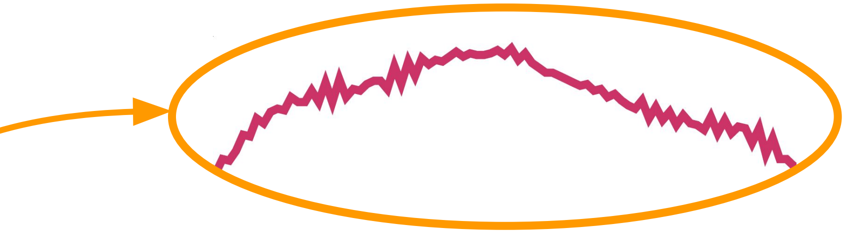
Verification: the ratio of the low frequency power to the total power.

$$R_1 = \sum_{f=85Hz}^{2kHz} A^2(f) / \sum_f A^2(f)$$

Time domain Verification

Modulated replay attacks compensate the spectrum distortion.





ringing artifacts: small oscillations in time-domain signals.

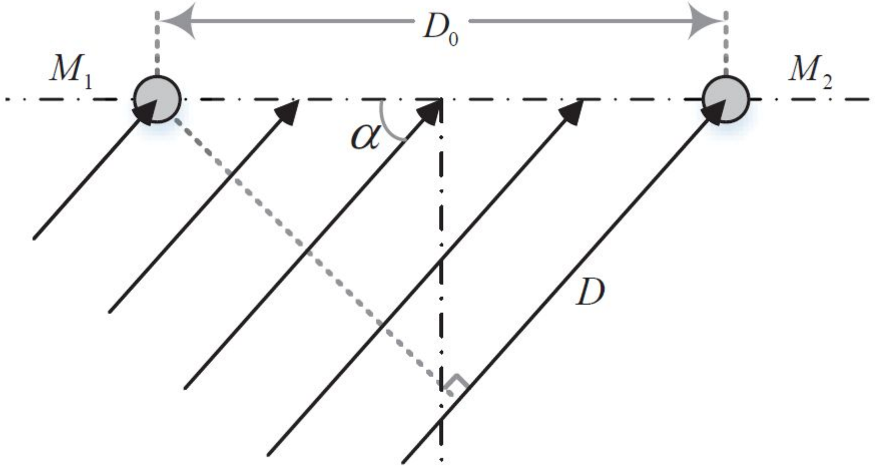
Verification: local extrema ratio. $R_2 = \frac{cnt_{w=3}}{N-2}$

Step3: Identifying Driver’s Voice

Objective:
Filter out passengers’ voice commands.

Key features:
Voice propagation direction.

Our method:
Time Difference of Arrival (TDoA).



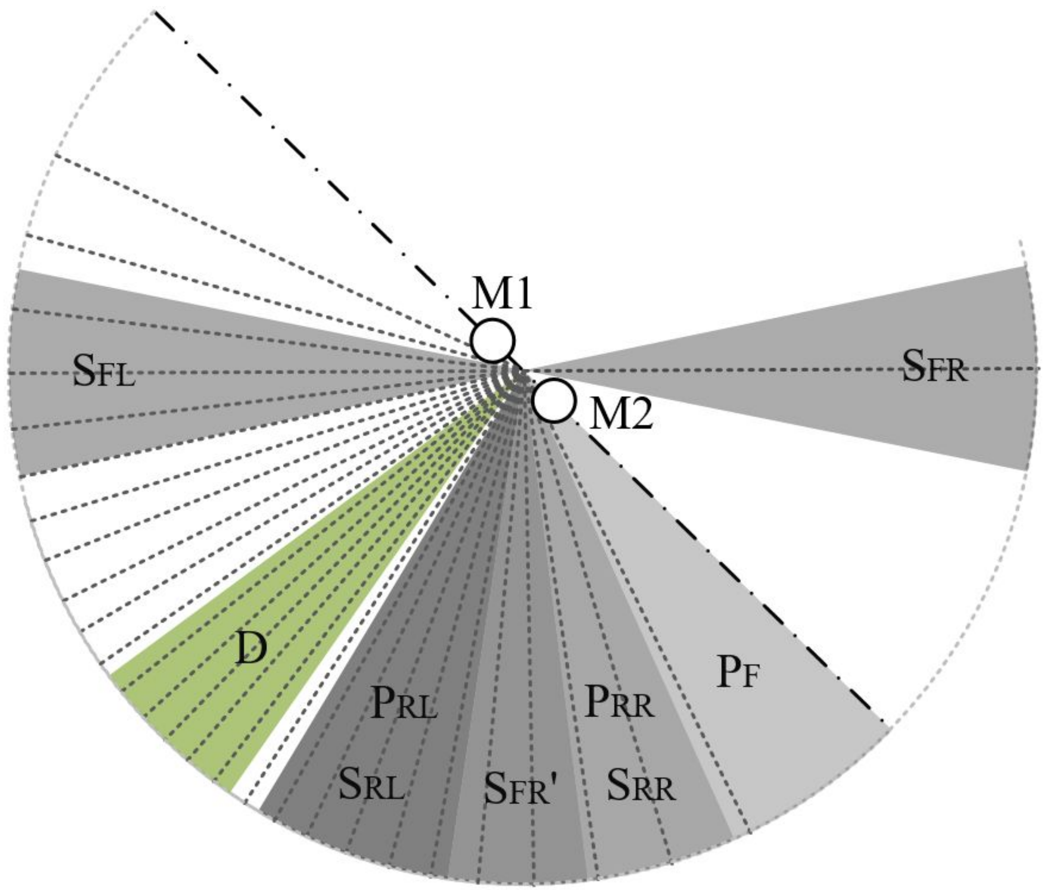
$$\alpha = arccos(\frac{D}{D_0})$$

Step3: Identifying Driver's Voice

- Propagation direction estimation

$$\alpha = arccos(\frac{\Delta N \cdot v_0}{D_0 \cdot f_s})$$

- Higher precision in the driver's direction.



Detection regions

Outline

Introduction

System Design

Experiments

Discussion

Conclusion

Experiments

Testbed:

Sedan: Toyota Camry LE

- 2 Scion TCXB 6.5-inch speakers
- 2 Kicker 43DSC69304 D-Series 6x9-inch speakers

Microphone: TASCAM DR-40

Laptop: Dell XPS15, 2.8GHz CPU

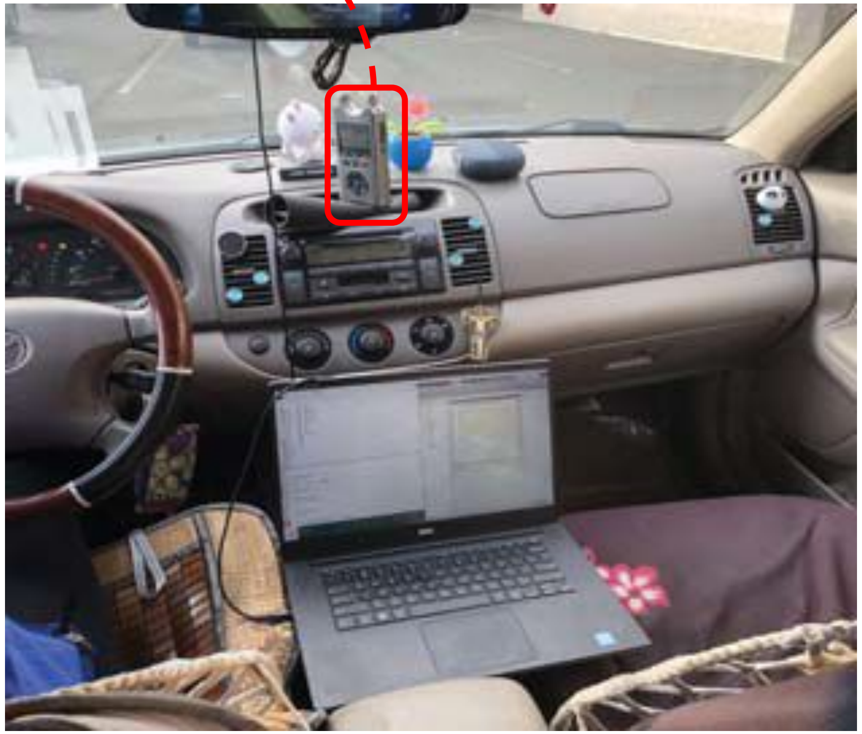
Real-World Testing:

Idling: running engine @ 0 mph

Local: 20 mph

Highway: 50 mph

TASCAM DR-40
Dual Microphone



Vehicle Testbed

Evaluation

Step 1:Detecting multiple speakers

The Detection Accuracy for Different Number of Speakers

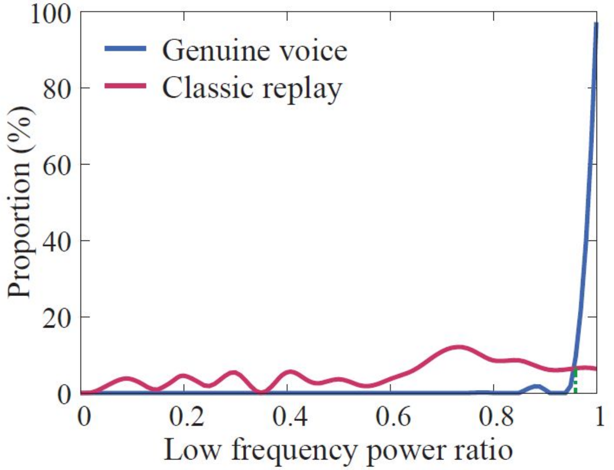
# of Speakers	Idling	Local	Highway
1	100%	83.3%	58.3%
2	66.7%	58.3%	66.7%
3	75%	66.7%	75%
4	100%	100%	100%
Total	83.3%	73.8%	71.4%

Evaluation

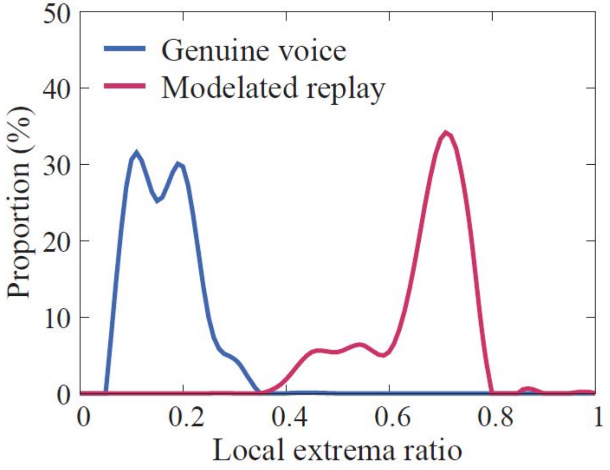
Step 2: Identifying human voice

The Detection Accuracy of Human Voice

Driving State	Accuracy
Idling	97.46%
Driving on Local Street	96.75%
Driving on Highway	94.20%



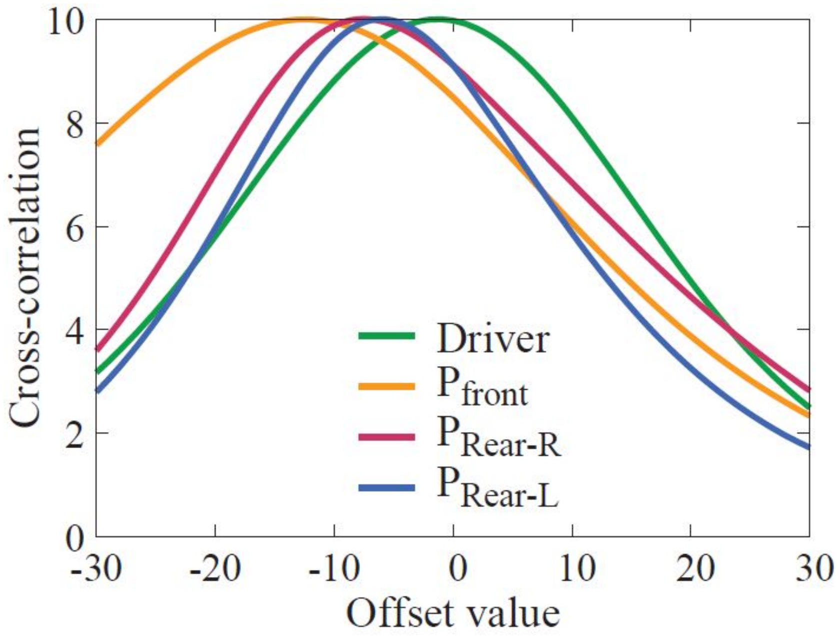
Frequency-domain check



Time-domain check

Evaluation

Step 3: Identifying driver's voice



The Peak Offsets for the Driver and Passengers

Voice Source		Idling	Local	Highway
Driver	Mean	-0.11	0.38	1.09
	Stdev	4.15	3.03	2.11
Front Passenger	Mean	-11.31	-10.99	-8.88
	Stdev	5.98	4.67	4.75
Rear Right Passenger	Mean	-8.02	-6.57	-5.31
	Stdev	4.04	3.29	5.00
Rear Left Passenger	Mean	-5.36	-5.30	-4.57
	Stdev	3.58	3.27	3.75

Evaluation

➢ Code size: 633 KB

➢ Well supported by the modern in-vehicle computing platforms.

➢ Optimized C code or assembly code may further reduce the running time.

Performance Overhead for Detection Step.

Detection Step	Running Time	Memory
Multi-speaker Detection	134 ms	111 MB
Human Voice Detection	47 ms	10 MB
Driver's Voice Identification	33 ms	23 MB
Total Overhead Costs	214 ms	144 MB

Outline

Introduction

System Design

Experiments

Discussion

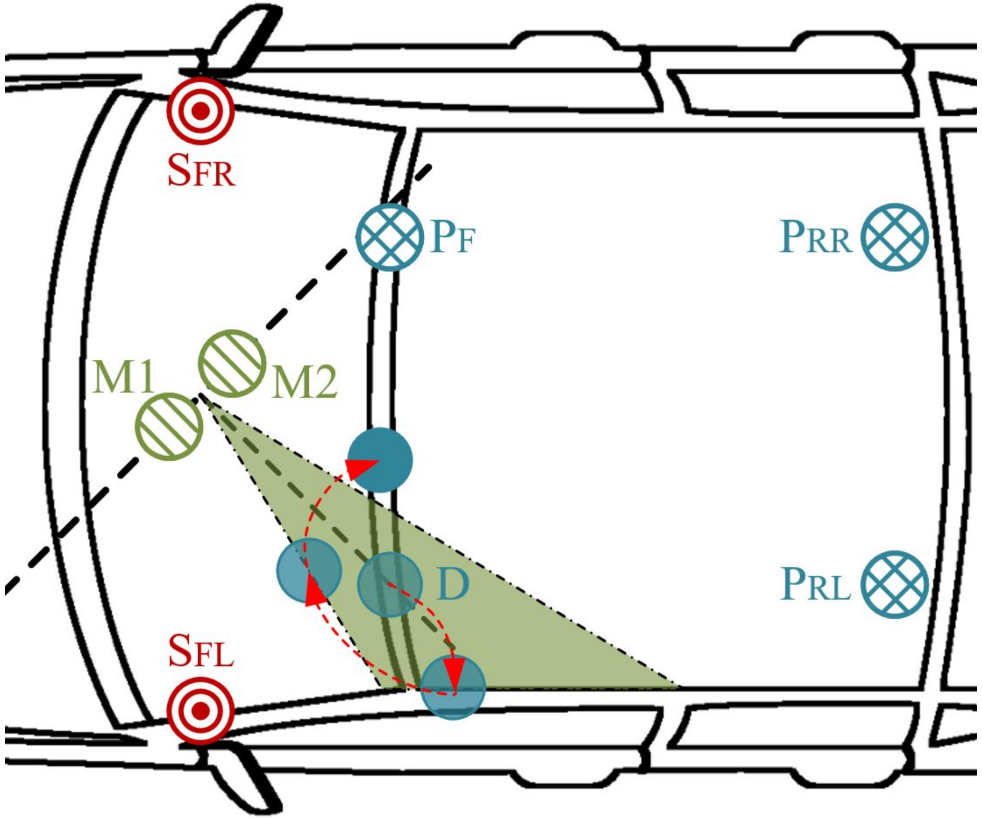
Conclusion

Discussion

➢ Spectrum-assisted Detection.

Commands must satisfy:

- the spectrum histograms of wake-up command are similar to the previous one.
- the voice movement is within an acceptable wider range.



Discussion

- SIEVE can be extended to other vehicle models or future driverless car models.
- It is also possible to deploy more microphones (or a microphone array) in the future car designs.
- Microphones with a higher sampling rate and denoising algorithms may provide a fine-grained angle measurement.
- Our techniques can be adopted in smart home systems.

Outline

Introduction

System Design

Experiments

Discussion

Conclusion

Conclusions

- ➢ SIEVE: defeat adversarial voice command attacks on voice-controlled vehicles.

- ➢ Distinguish the driver's voice with a three-step scheme.
  - Detecting multiple speakers
  - Identifying human voice
  - Identifying driver's voice

- ➢ Experimental results show our system can achieve a high detection accuracy in real-world situations.

Thank you!

Authors:
Shu Wang, Jiahao Cao, Kun Sun, Qi Li

Questions?
My Email: swang47@gmu.edu

George Mason University

Center for Secure Information Systems CSIS

清華大學 Tsinghua University