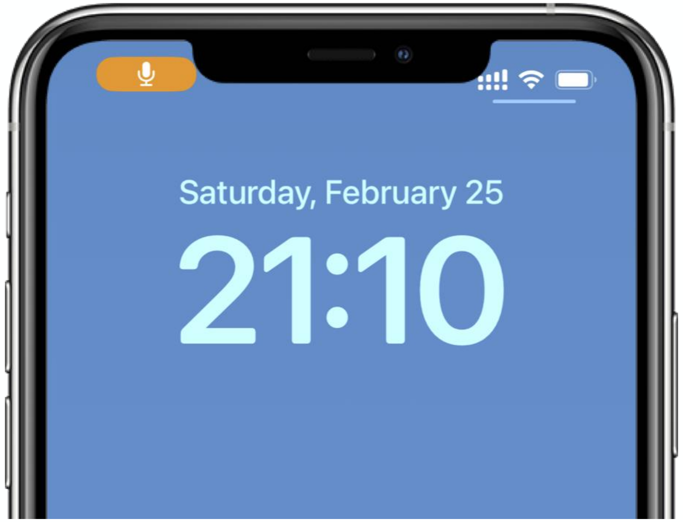
InfoMasker: Preventing Eavesdropping Using Phoneme-Based Noise

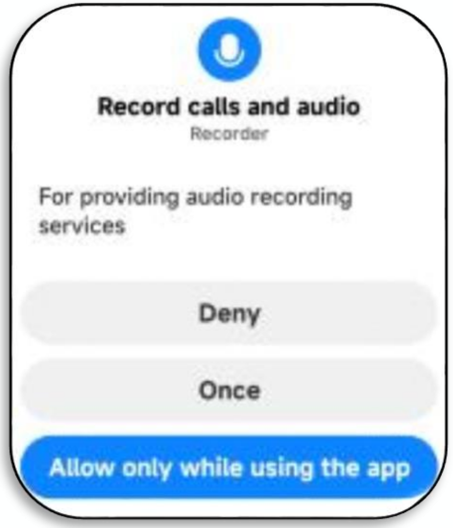
Peng Huang, Yao Wei, Peng Cheng, Zhongjie Ba,
Li Lu, Feng Lin, Fan Zhang, Kui Ren

Eavesdropping with Smart Devices

- Widespread of smart devices equipped with microphone

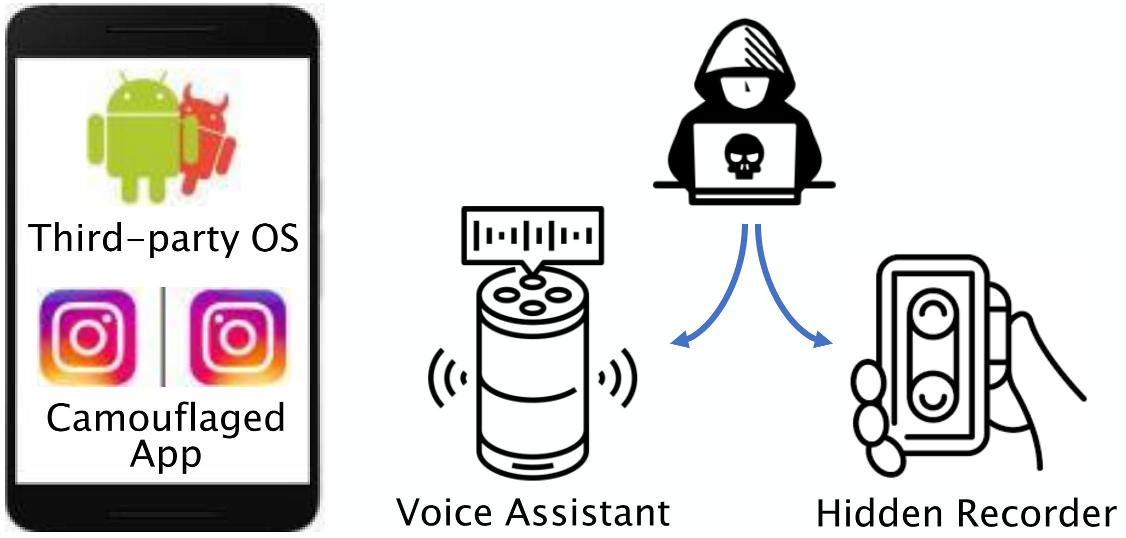
- Developers are committed for privacy protection





Eavesdropping with Smart Devices

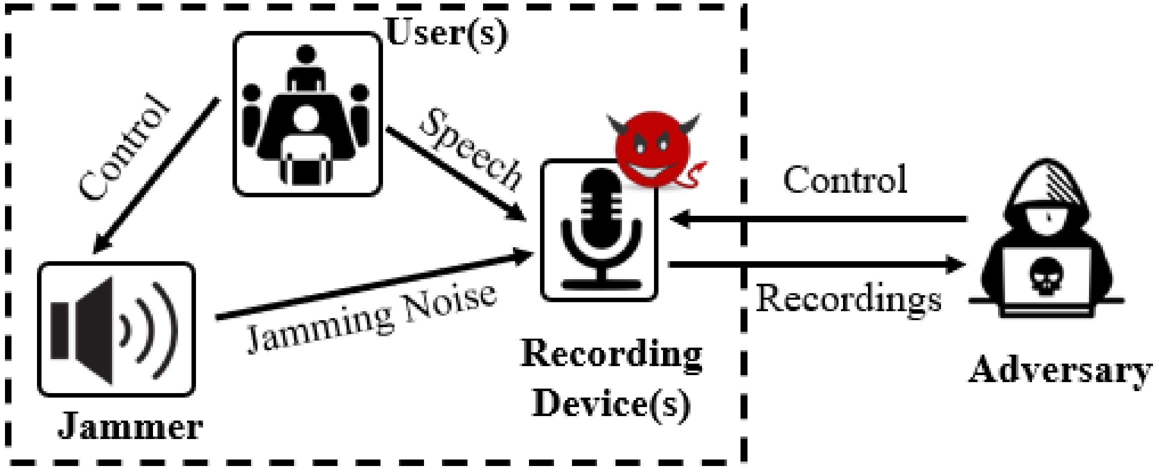
- Still an unsolved problem
  - Third-party operating systems
  - Malicious fake applications
  - Uncontrolled legal recordings
  - Hidden Recorders



- Need to physically block voice eavesdroppers
  - Makes the voice privacy controllable to the users.

Problem Setup

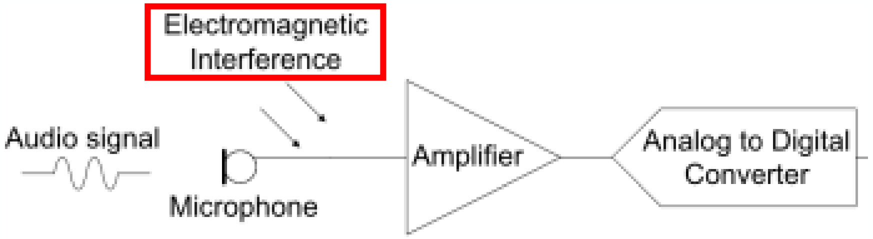
- Application scenario

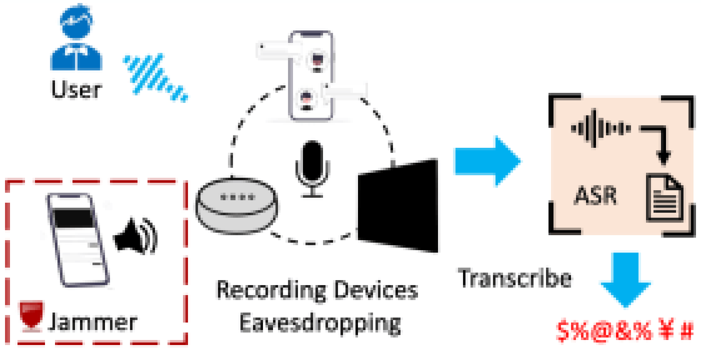


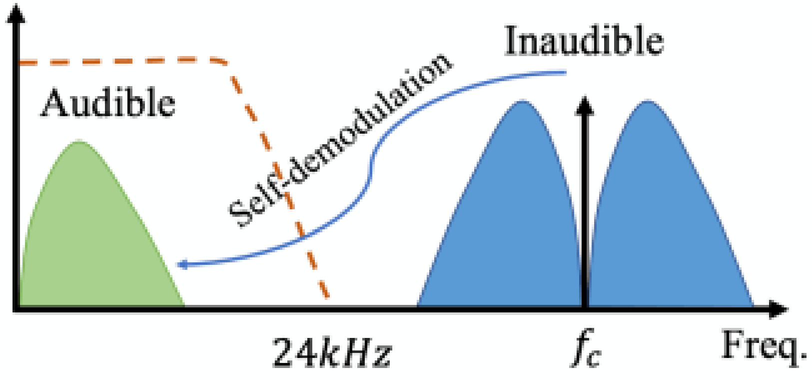
- Design goals
  - Effectiveness
    - Successfully mislead human ears
    - Successfully mislead automatic-speech-recognition tools
  - Robustness
    - Could not be removed by noise reduction methods
  - User-friendly
    - Should not disturb users

Existing Methods to Jam Microphone

- Electromagnetic interference-based jamming
  - **Pros**: No disturbance to users
  - **Cons**: Limited coverage & Affect other devices
- Adversarial example-based jamming
  - **Pros**: No need for special hardware
  - **Cons**: No effect to human ear& generalization ability
- **Ultrasound-based jamming**
  - **Pros**: No disturbance to users & Reasonable coverage



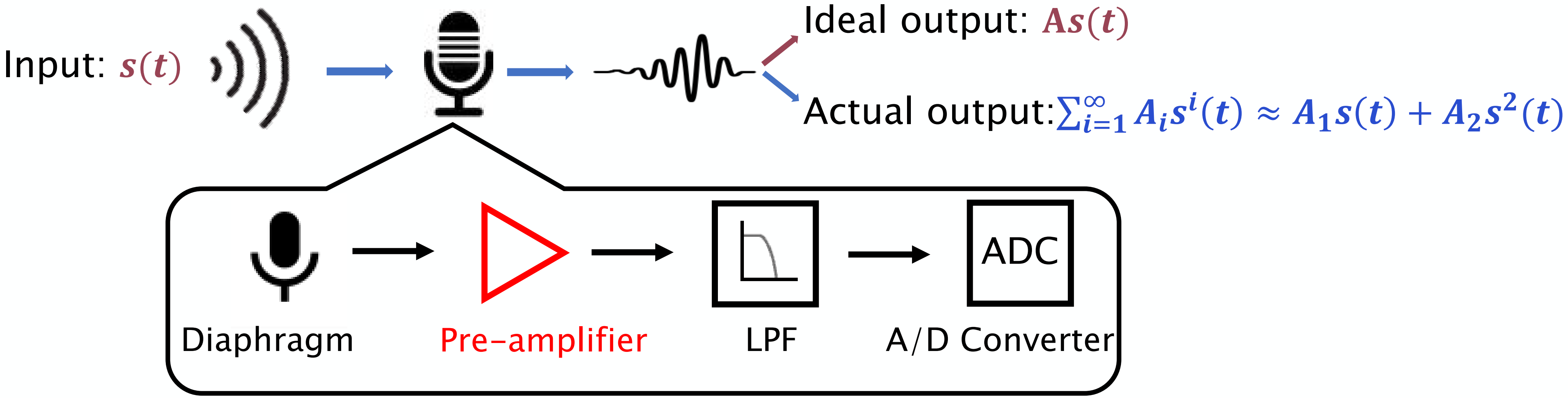




Principle of Ultrasound-Based Microphone Jamming

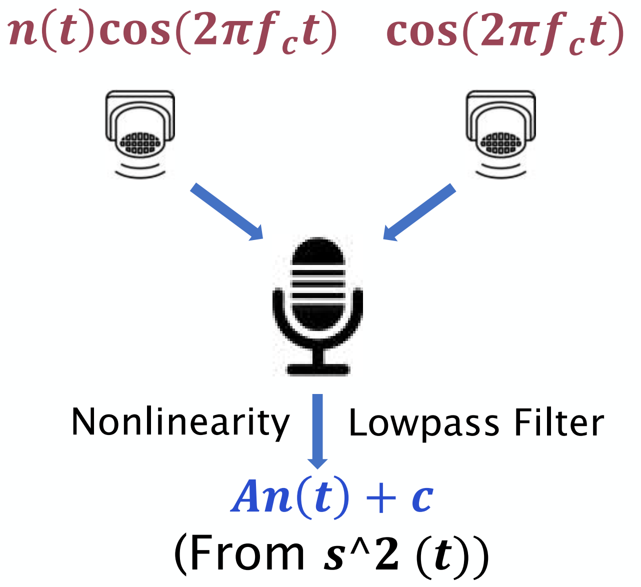
- Nonlinearity in microphone will cause self-demodulation of input signals.
  - Zhang et al. (2017) inject inaudible voice commands to microphone via ultrasound[13]

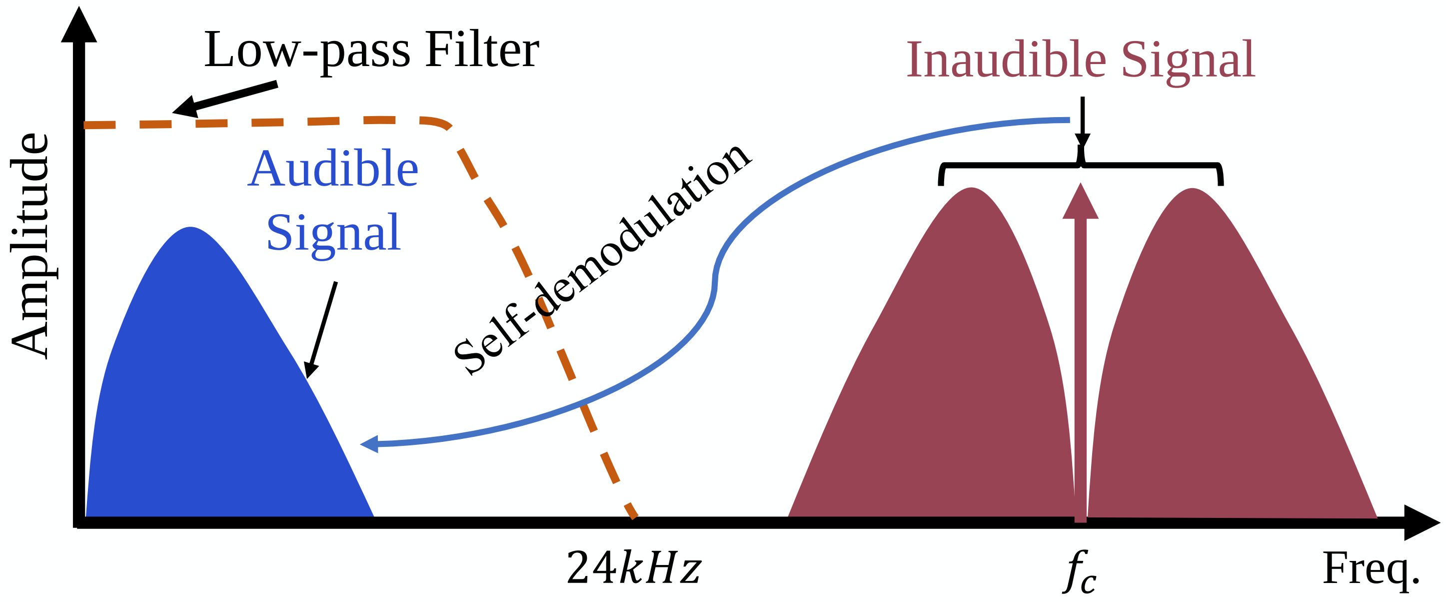
- Nonlinearity in microphone



Principle of Ultrasound-Based Microphone Jamming

- Nonlinearity in microphone will cause self-demodulation of input signals.
  - Zhang et al. (2017) inject inaudible voice commands to microphone via ultrasound[13]
- Inject audible noise $n(t)$ with inaudible ultrasound

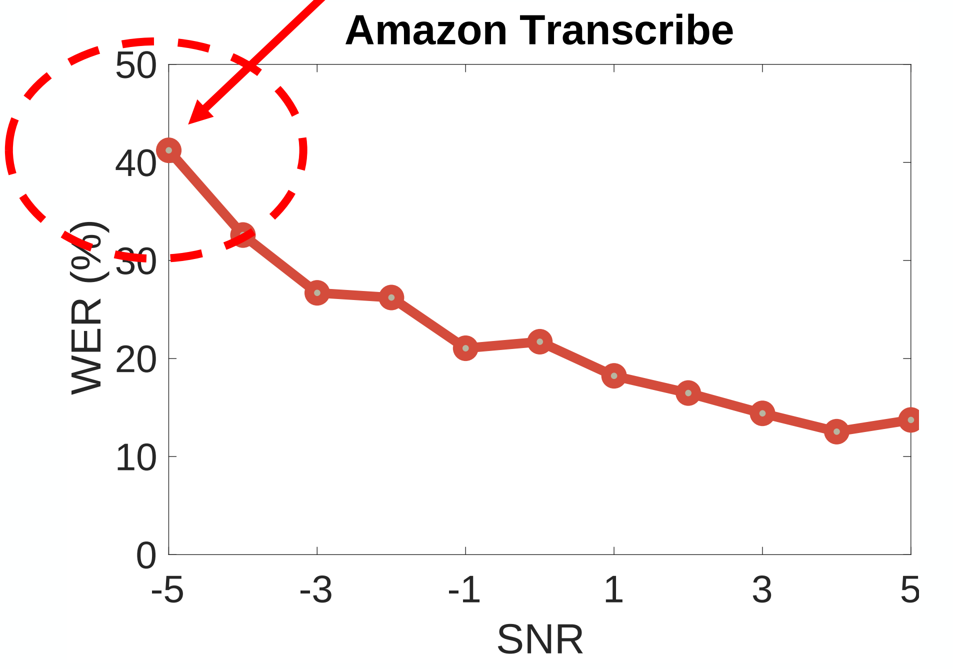




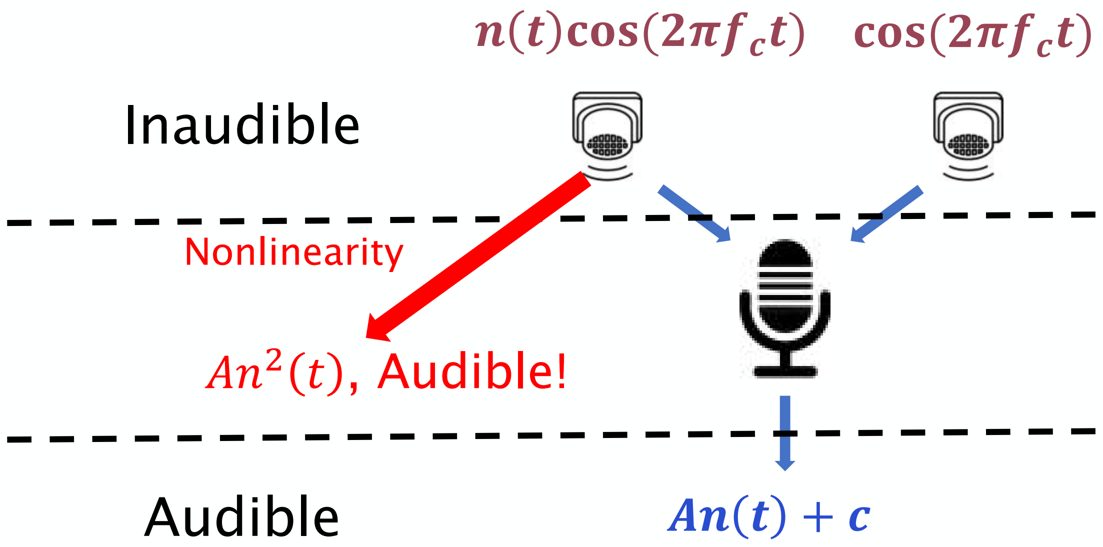
Challenges of Jamming Microphone

1. High demand for noise energy vs. Limited transmission energy

SNR < −5 to achieve a WER higher than 40%



Audible interference generated from transmitter's nonlinearity



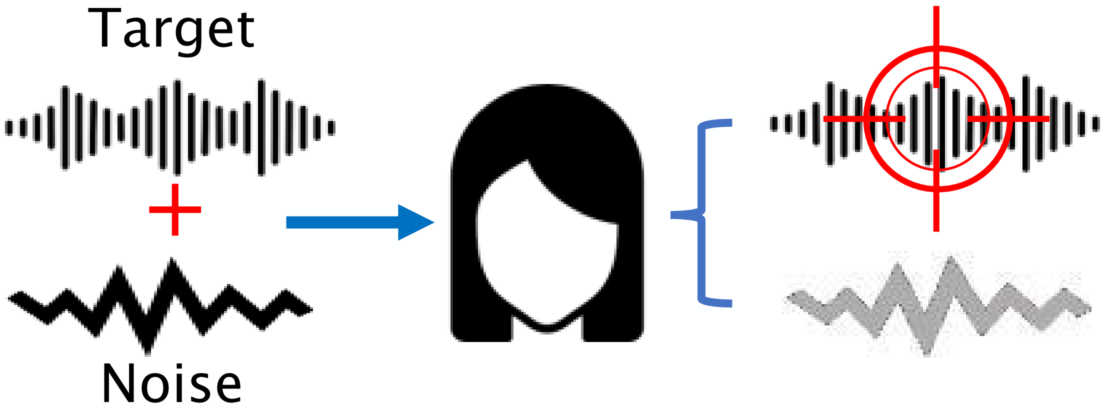
Challenges of Jamming Microphone

2. Target speech recognition tools (human and ASR) have strong denoising ablility

- Common noises with limited energy will be easily removed

- **Cocktail party effect**[4] in human ear

Human brain can easily **focus on the target speech** in a noisy environment

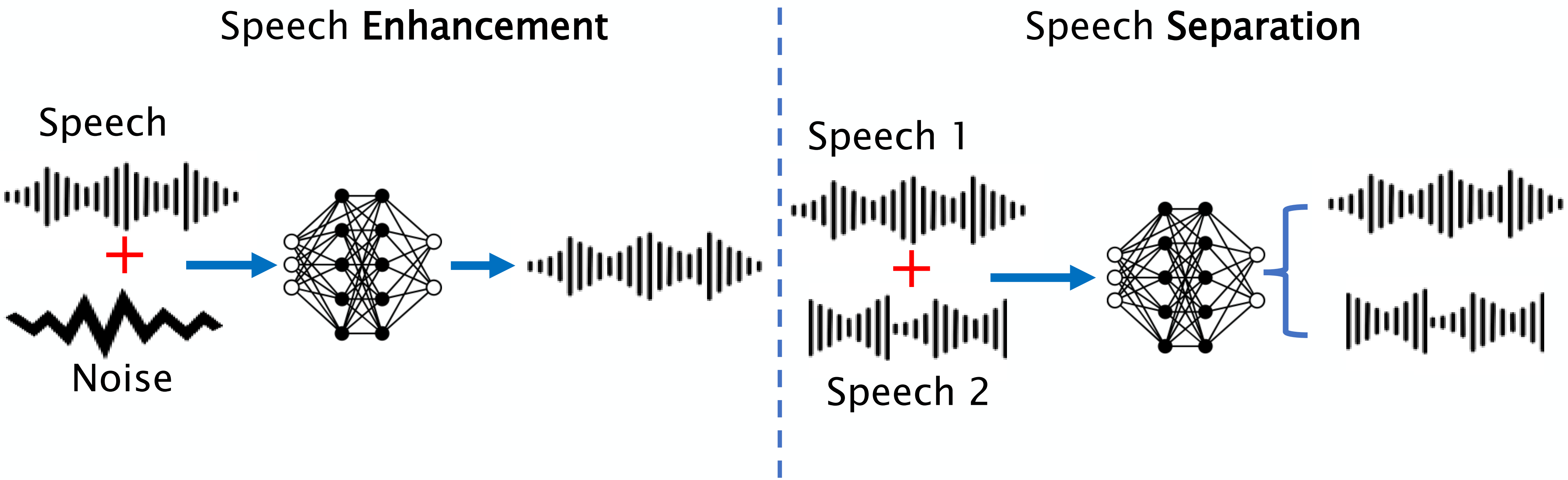


Challenges of Jamming Microphone

2. Target speech recognition tools (human and ASR) have strong denoising ablility

- Common noises with limited energy will be easily removed

- **Noise reduction** methods in ASR

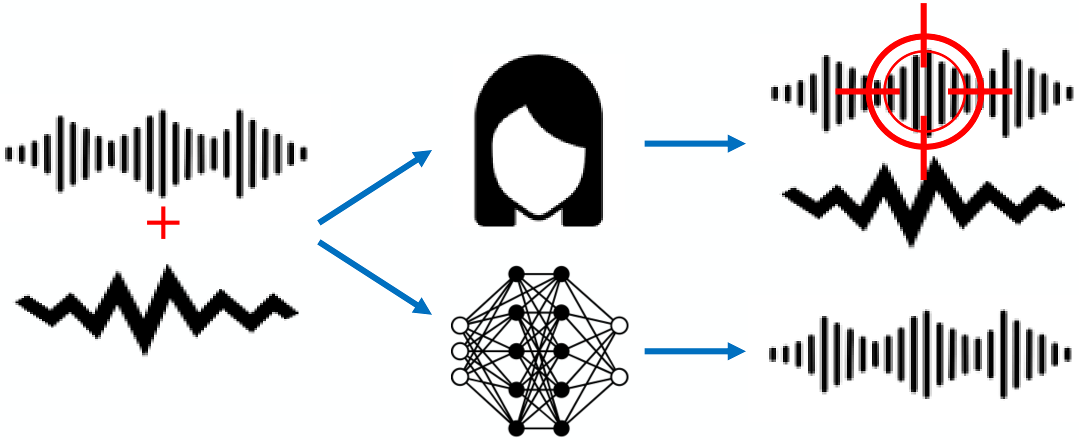


Challenges of Jamming Microphone

2. Target speech recognition tools (human and ASR) have strong denoising ablility

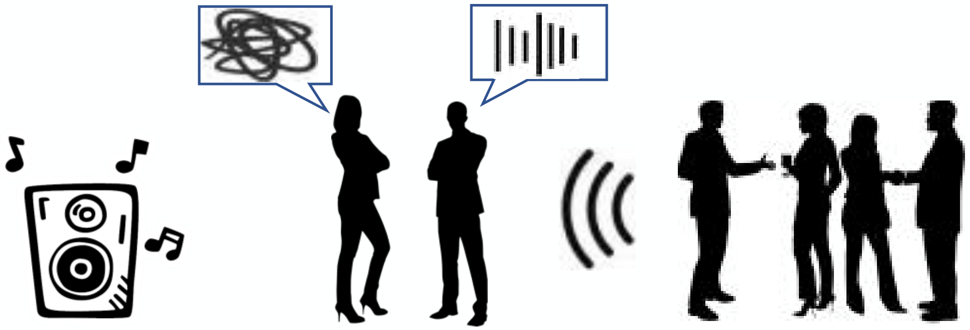
    - Common noises with limited energy will be easily removed

- Both methods rely on the differences between the noise and the speech



Time Domain	Structure
	Speaking Rate
	Random Gaps
	Intensity
Frequency Domain	Fundamental Frequency (F0)
	Timbre
Spatial	–

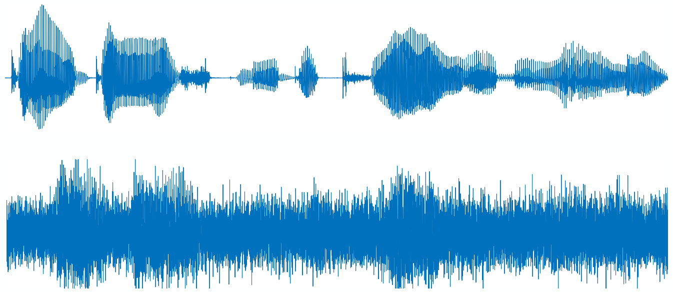
Jamming Strategy: Energetic v.s. Informational



- **Energetic masking: Covering**

Origin wave

Masked wave



- **Informational Masking: Disturbing**

Origin Word: desk

Phonogram: / desk /

Inject / ɪ / / de ɪ sk / desk? disk?

- Characteristics

Pros: No need for prior knowledge

Cons: High energy requirement & Easily to remove

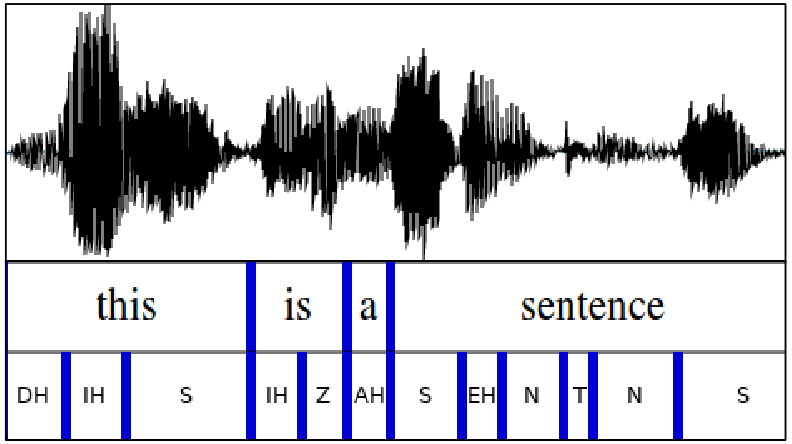
- Characteristics

Pros: Low energy requirement & Hard to remove

Cons: Needs prior knowledge

Informational Masking for Human Speech Jamming

- **Prior knowledge** for jamming human speech
  - Signal structure: a series of phonemes
  - Frequency domain properties: User dependent
    - Fundamental frequency (F0)
    - Timbre
  - Time domain properties : Varying and uncertain

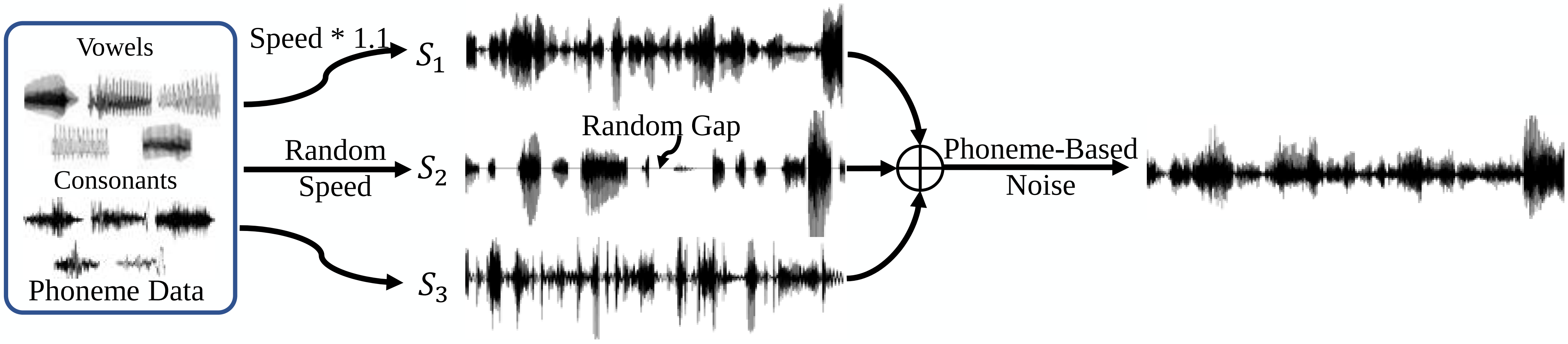


Main idea: Inject phonemes similar to the target speech to disturb it

Phoneme-Based Jamming Noise Design

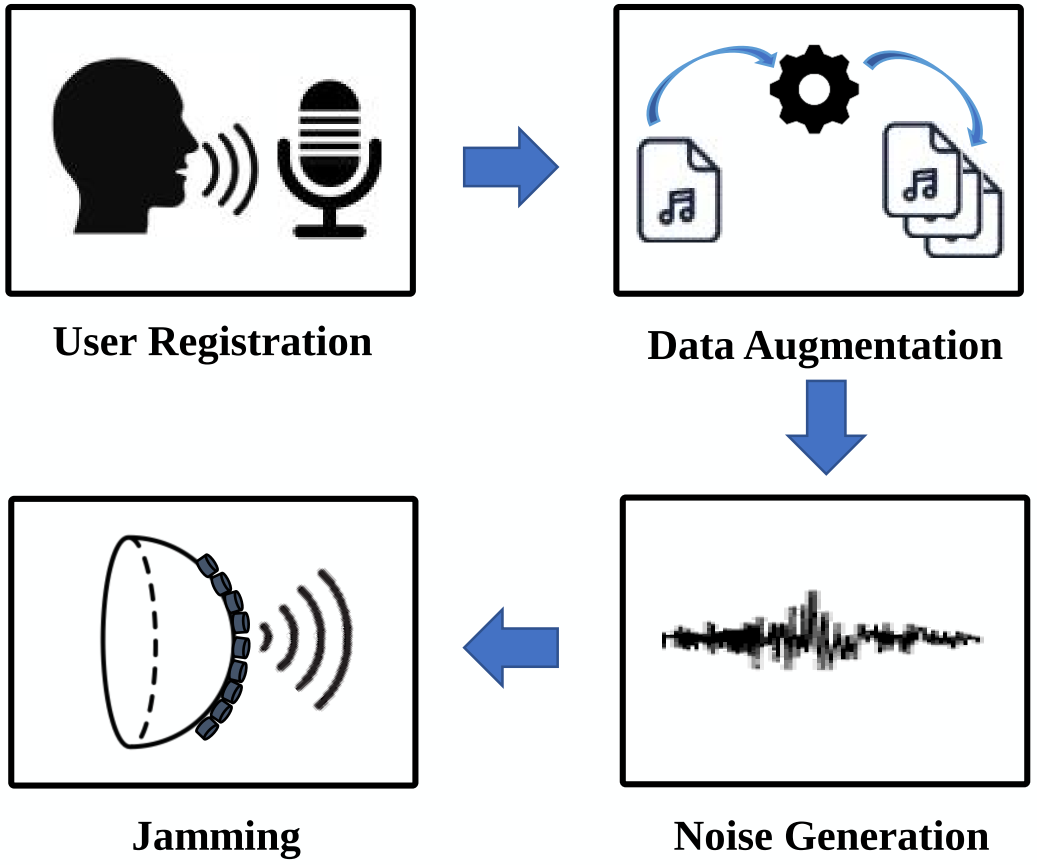
- Noise structure

Noise Series	Function
Ⅰ: **Accelerated** continuous vowels	Inject enough phoneme per unit time
Ⅱ: Vowels with **random speed and gap**	Narrow down the difference in speaking rate
Ⅲ: Continuous consonants	Increase the diversity of the noise



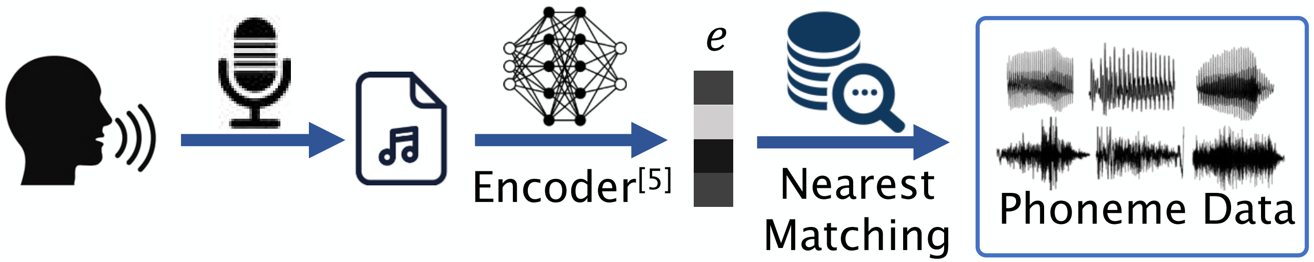
System Workflow

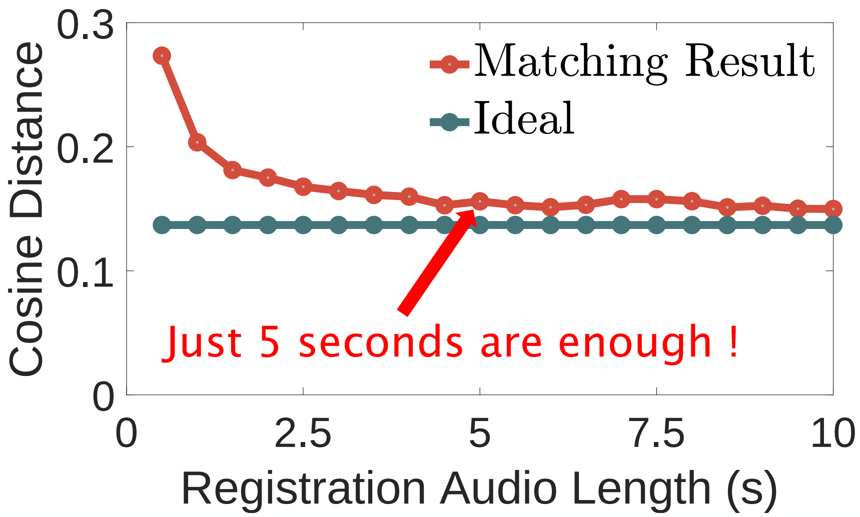
- User Registration
  - Get the user’s voice features
- Data Augmentation
  - Get enough data for noise generation
- Noise Generation
  - Get the noise
- Jamming
  - Inject the noise to microphone



User Registration

- Purpose: Obtain enough phoneme data with similar timbre as the user.
- Extracting from the user's speech is time consuming, and so not practical ❌
- Extract user's voice feature from short registration audios and match speech data from public corpus





Data Augmentation

- Increase the amount of phonemes while retaining similarity with original data

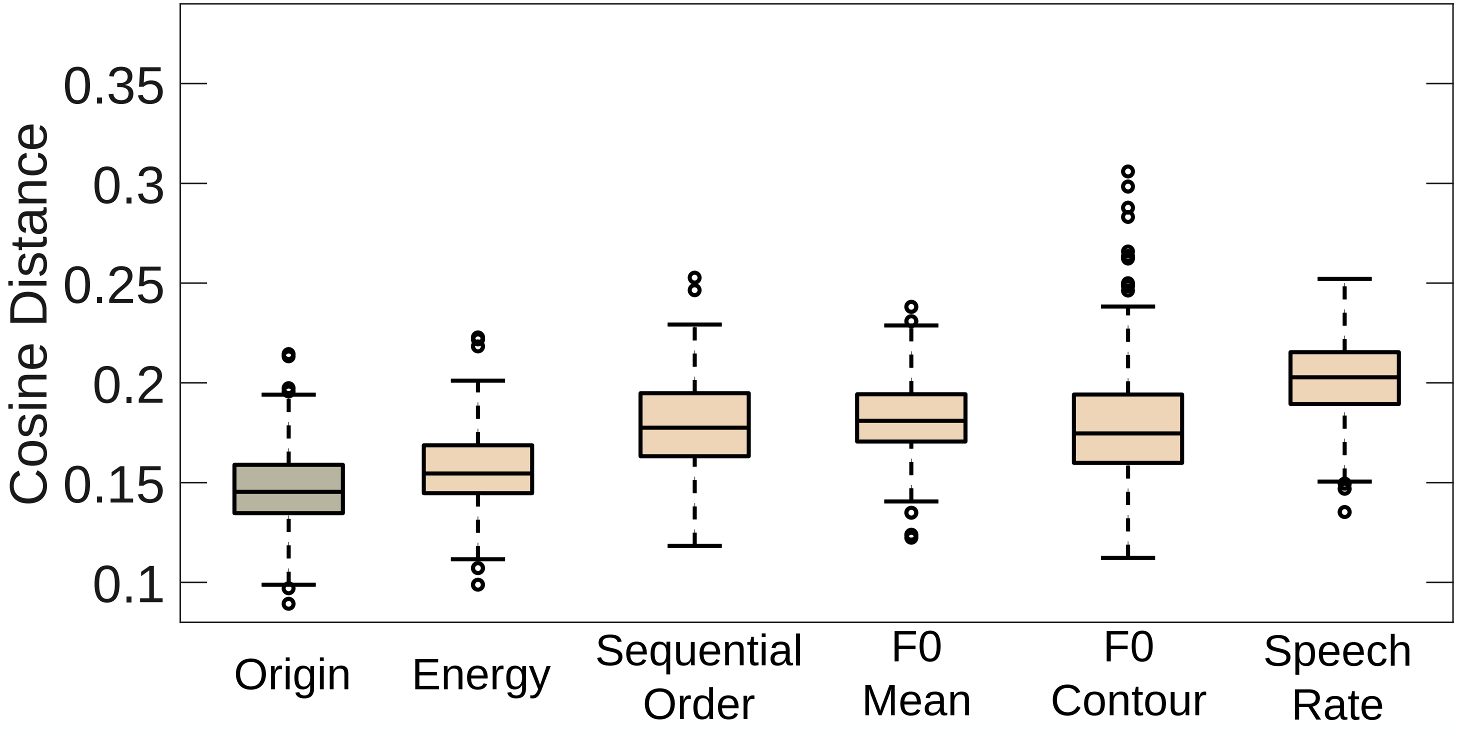
- *Method:* Fine-tune the emotional-related speech properties[6].

Phonetical Properties	Modification Range	Emotional Impact ↑	Emotional Impact ↓
Speech Rate	0.3-1.8	Fear or Disgust	Sadness
F0 Mean	0.9-1.1	Anger or Happiness	Disgust or Sadness
F0 Contour	0.7-1.3	Anger or Happiness	Sadness
Energy	0.5-2.0	-	-
Sequential Order	-	-	-

Data Augmentation

- Increase the amount of phonemes while retaining similarity with original data

- ***Method:*** Fine-tune the emotional-related speech properties[6].



Noise Transmission

- **Lower-sideband modulation** to achieve higher transmission energy

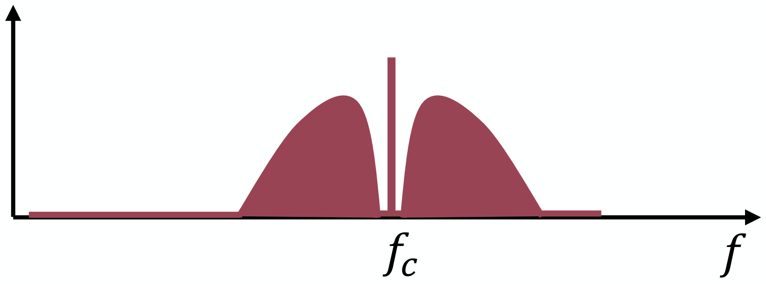
Conventional Modulation

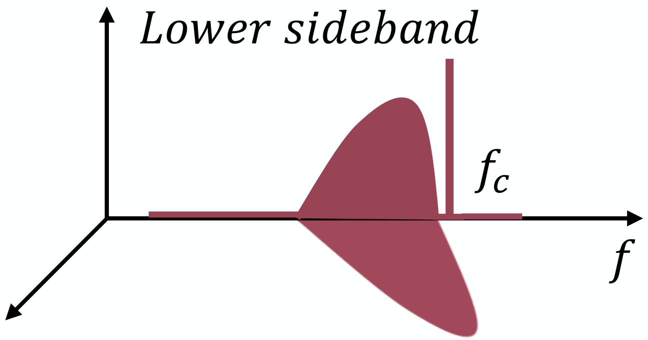
$$s(t) = \sqrt{2}n(t)\cos(2\pi f_c t)$$

Single-sideband Modulation

$$s(t) = n(t)\cos(2\pi f_c t) + \hat{n}(t)\sin(2\pi f_c t)$$

Spectrum:





Audible Signal:

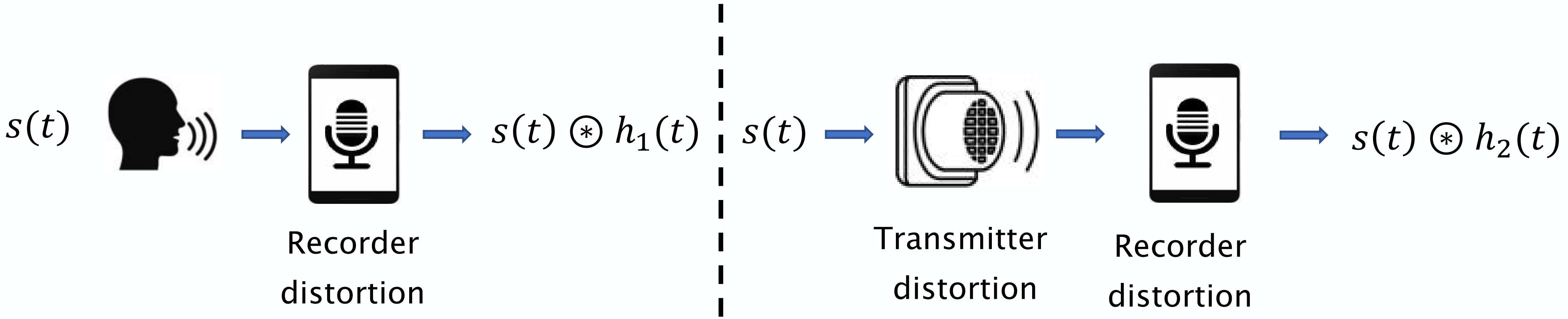
$$|n^2(t)| = \sqrt{2}\left|\frac{1}{2}(n^2(t) + \hat{n}(t))\right|$$

User Study Results:

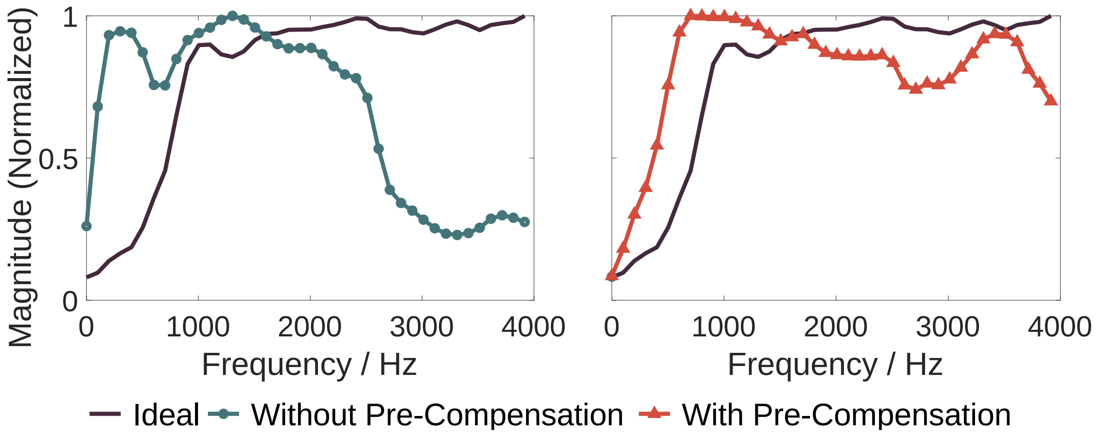
Noise	Normalized Energy		
	DSB-AM	LSB-AM	USB-AM
White Noise	1.00	1.49	1.29
Phoneme-Based Noise	2.77	**4.14**	3.61

Noise Transmission

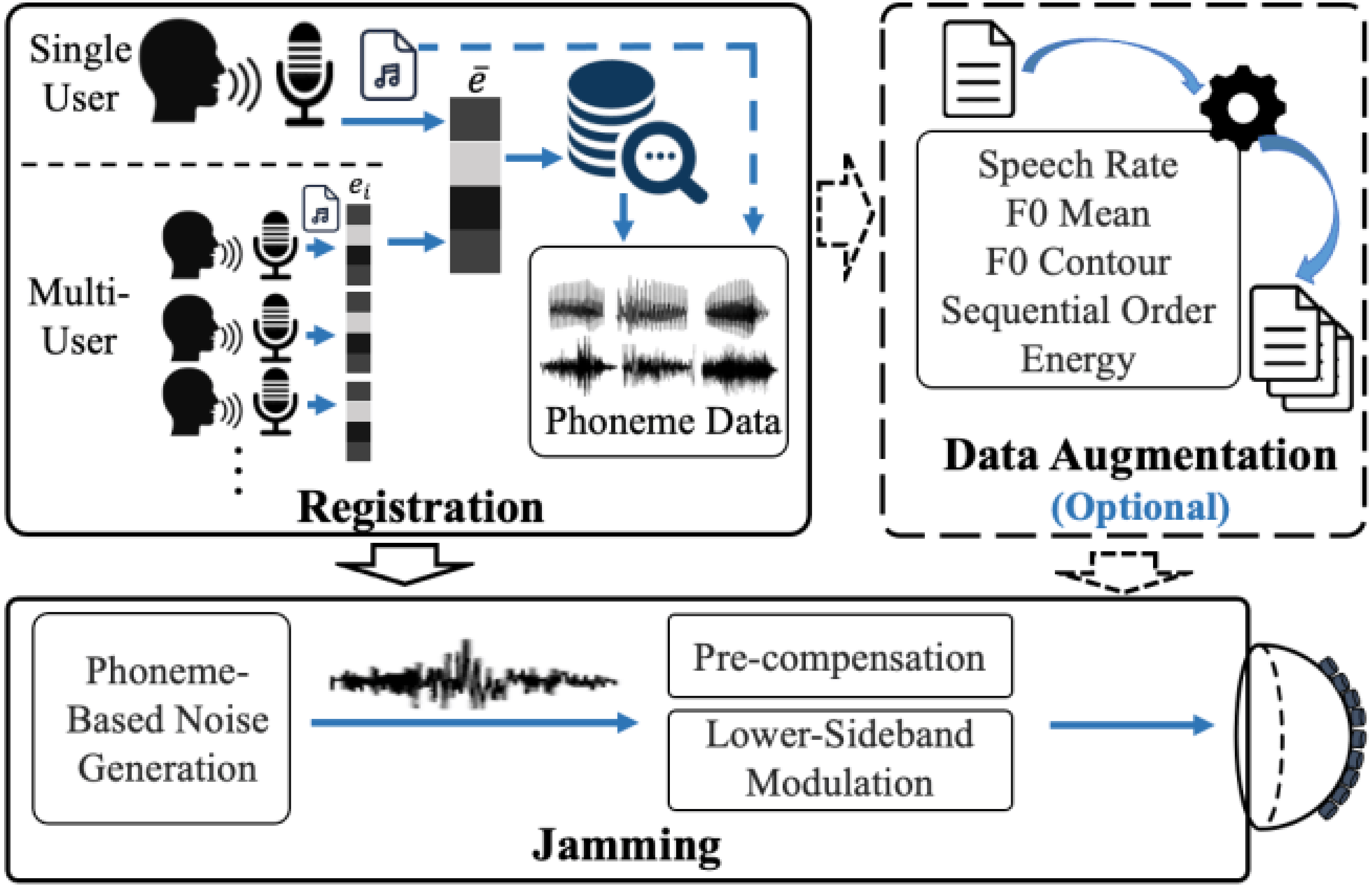
- **Pre-compensation** to reduce distortion during transmission



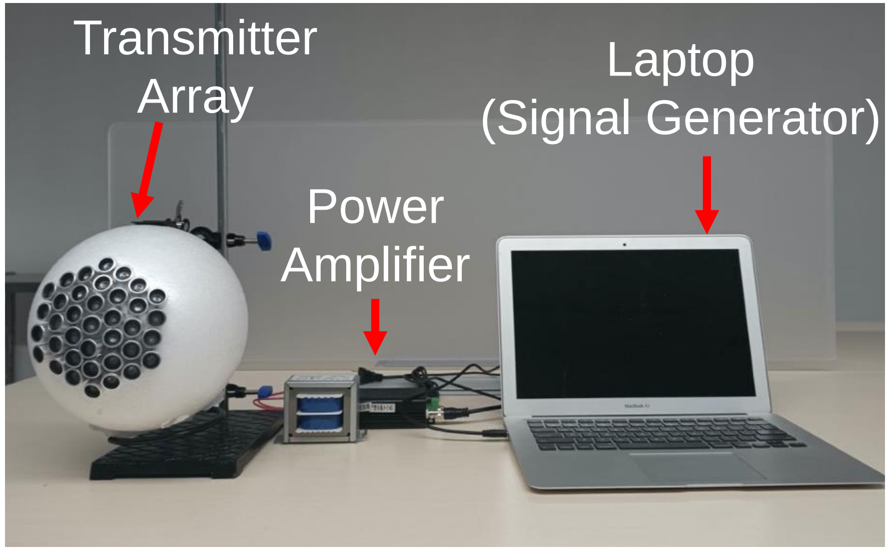
- Estimate $h_1(t)$ and $h_2(t)$, pre-compensate $s(t)$ with $h_1(t) \circledast h_2^{-1}(t)$

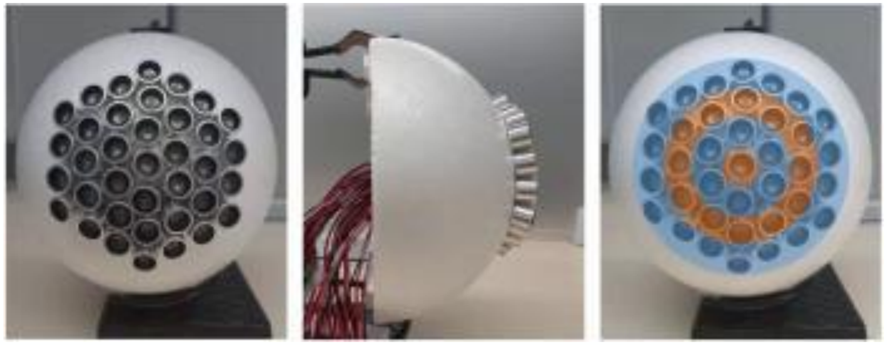


System Overview



System & Hardware



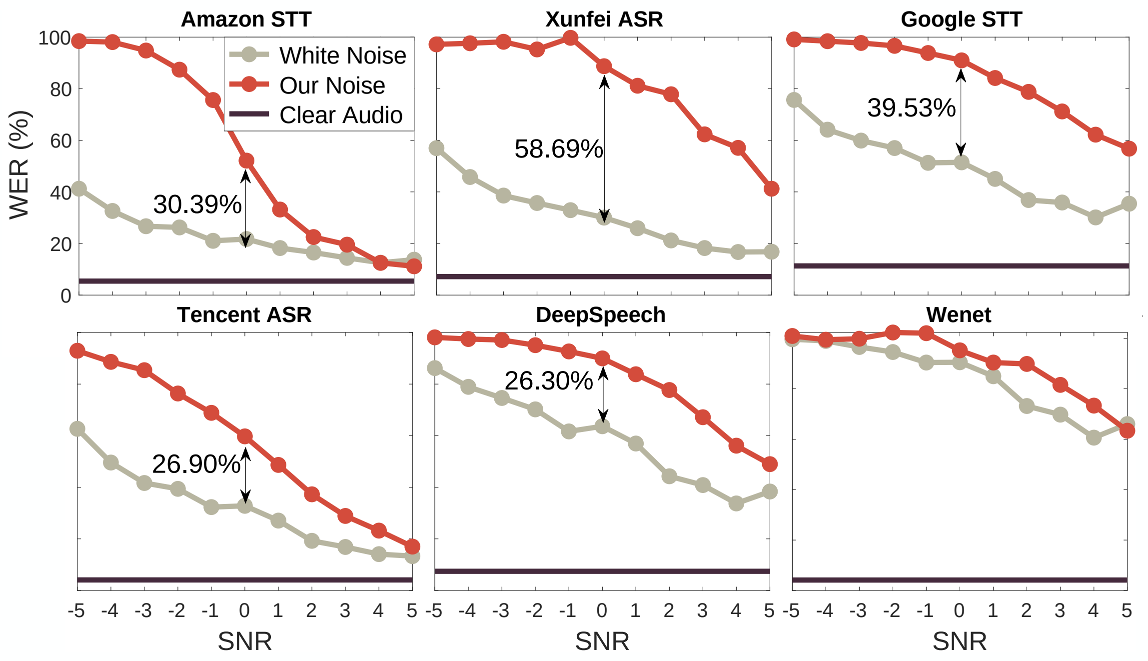


Evaluation: Experimental Setting

- Speech recognition tools
  - 4 Commercial ASR tools
  - 2 Open-Source ASR tools
  - Human recognition
- Datasets
  - LibriSpeech[7] for most experiments
  - TIMIT[8] for training targeted ASRs
  - Harvard Sentences[9] for human recognition
- Evaluate aspects
  - Effectiveness
  - Robustness
- Scenarios
  - Digital domain
  - Real-world jamming
  - Case study: A common office

Evaluation: Effectiveness

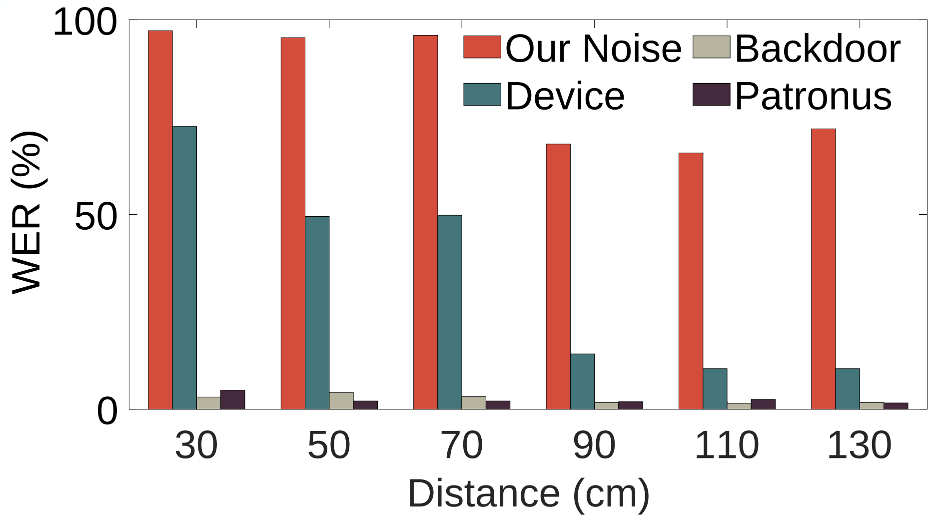
- Digital domain
  - 27000 words for each ASR
  - Compared with [0, 8] kHz bandlimited white noise.
- Real-world jamming
  - 70 hours data



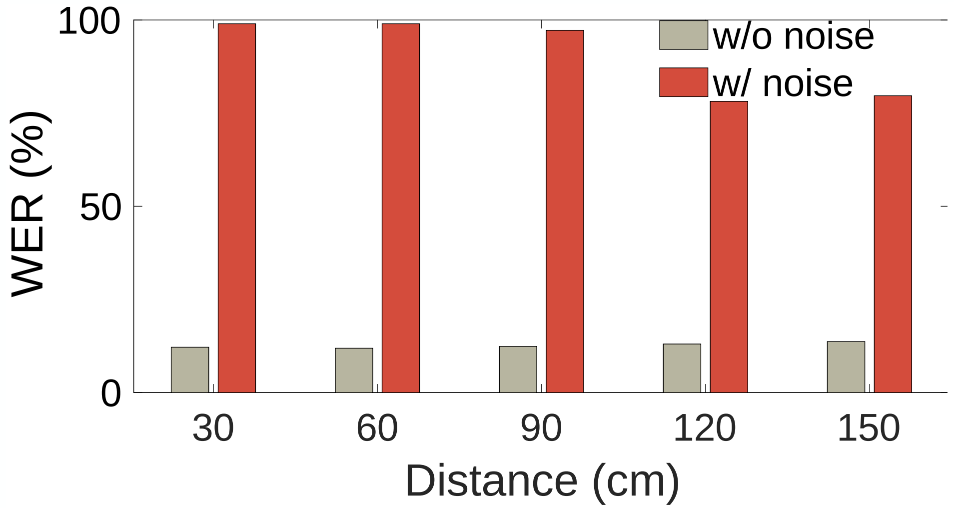
SNR	<-4	[-4,-2]	[-2, 0]	[0,2]	[2,4]	>4	Clear
Avg WER(%)	85.8	81.6	77.6	70.2	56.4	42.3	11.5
Min WER(%)	68.6	77.0	62.4	62.2	45.3	30.3	-
Digital WER(%)	88.6	85.4	68.8	48.67	28.9	17.0	4.1

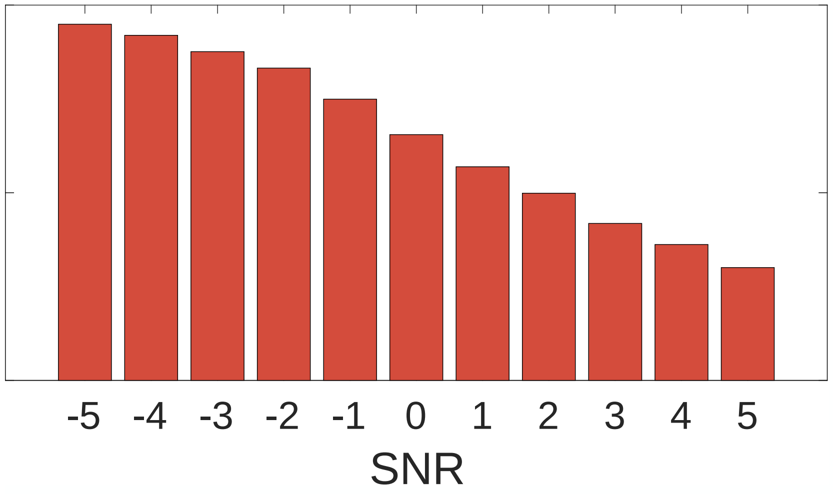
Evaluation: Effectiveness

- Comparisions with existing works
  - Two previous works and one commercial device.
  - With the presence of noise reduction methods



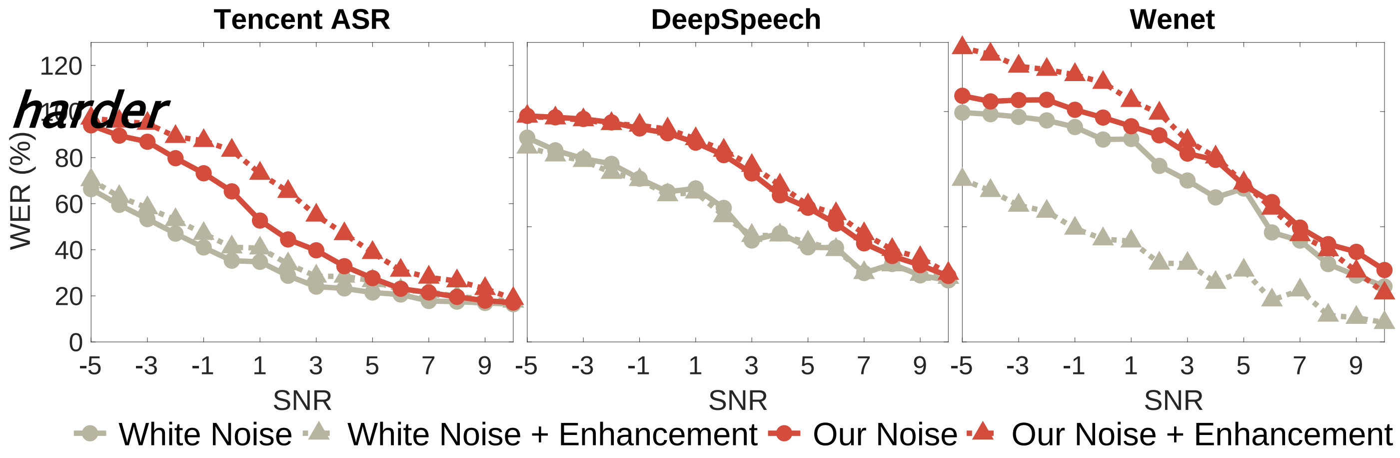
- Real-world end-to-end scenario



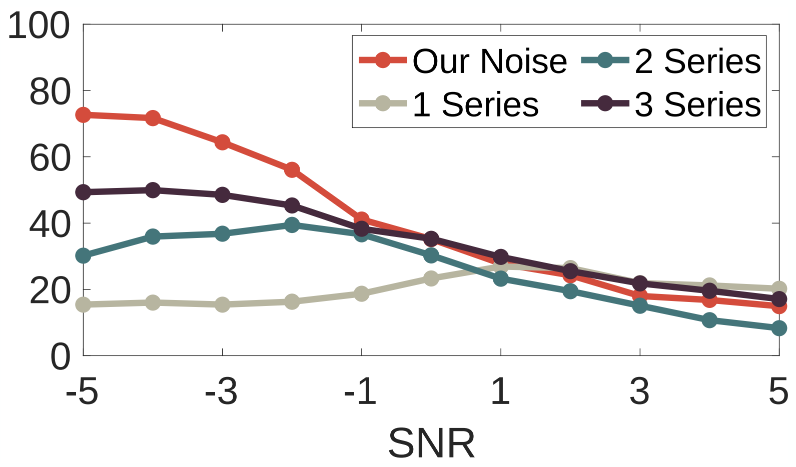


Evaluation: Robustness

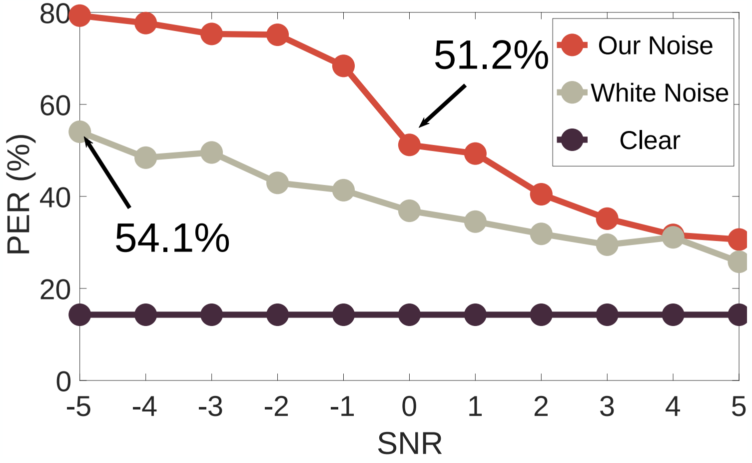
- Speech enhancement method[10]
  - Makes the distrubed speech *harder* to be recognized



- Speech Separation[11]

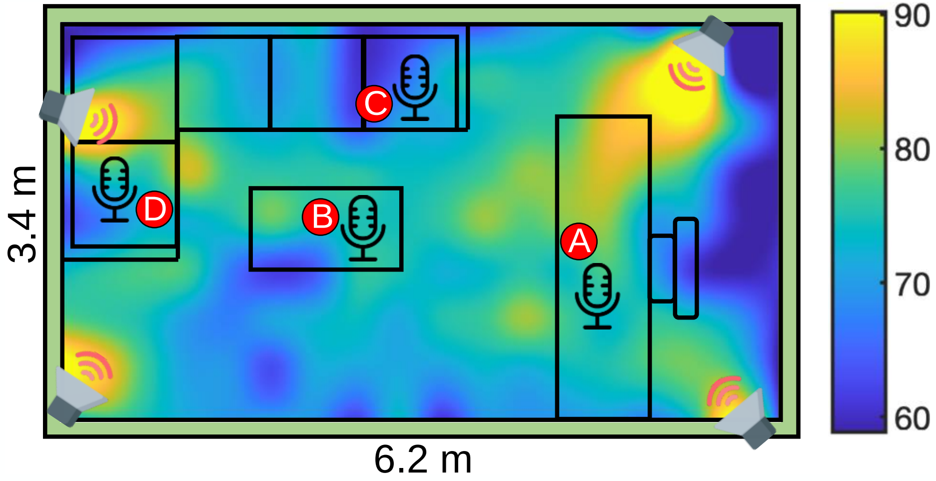


- Specialized ASR



Evluation: Case Study

- Setting



- Results

Types	WER(%)			
	Phone A	Phone B	Laptop	iPad
A	98.0	98.2	95.7	99.3
B	98.8	98.4	88.1	93.8
C	98.5	56.4	95.8	98.6
D	95.7	97.7	97.9	95.3
Amplifiers On	25.8	26.3	32.5	32.0
Clear	16.0	7.1	19.9	15.5

Thank You !

Peng Huang, Yao Wei, Peng Cheng, Zhongjie Ba,
Li Lu, Feng Lin, Fan Zhang, Kui Ren

References

[1] N. Roy, H. Hassanieh, and R. Roy Choudhury, “Backdoor: Making microphones hear inaudible sounds,” in *Proceedings of the 15th An- nual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys ’17. New York, NY, USA: Association for Computing Machinery, 2017, p. 2–14.

[2] L. Li, M. Liu, Y. Yao, F. Dang, Z. Cao, and Y. Liu, “Patronus: Preventing unauthorized speech recordings with support for selective unscrambling,” in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, ser. SenSys ’20. New York, NY, USA: Association for Computing Machinery, 2020, p. 245–257.

[3] Y. Chen, H. Li, S.-Y. Teng, S. Nagels, Z. Li, P. Lopes, B. Y. Zhao, and H. Zheng, “Wearable microphone jamming,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–12.

[4] D. S. Brungart, “Informational and energetic masking effects in the perception of two simultaneous talkers,” *The Journal of the Acoustical Society of America*, vol. 109, no. 3, pp. 1101–1109, Mar. 2001 .

[5] L. Wan, Q. Wang, A. Papir, and I. Lopez-Moreno, “Generalized End-to-End Loss for Speaker Verification,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, 2018, pp. 4879-4883.

[6] R. Cowie *et al.*, "Emotion recognition in human-computer interaction," in *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32-80, Jan 2001.

References

[7] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 5206-5210.*

[8] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, “Timit acoustic-phonetic continuous speech corpus ldc93s1,” 1993.

[9] “IEEE recommended practice for speech quality measurements,” *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225– 246, 1969.

[10] X. Hao, X. Su, R. Horaud, and X. Li, “FullSubNet: A Full-Band and Sub-Band Fusion Model for Real-Time Single-Channel Speech Enhancement,” *arXiv:2010.15508 [cs, eess]*, Jan. 2021.

[11] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, “Attention Is All You Need In Speech Separation,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 21–25.

[12] G. K. C. Chen and J. J. Whalen, “Comparative RFI Performance of Bipolar Operational Amplifiers,” in *1981 IEEE International Symposium on Electromagnetic Compatibility*, Aug. 1981, pp. 1–5.

[13] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, “DolphinAttack: Inaudible Voice Commands,” in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, Dallas, Texas, USA, Oct. 2017, pp. 103–117.