ACM MobiSys2025

SecHeadset: A Practical Privacy Protection System for Real-time Voice Communication

Peng Huang[1], Kun Pan[1], Qinglong Wang[1, 2], Peng Cheng[1, 2],
Li Lu[1, 2], Zhongjie Ba[*,1, 2], Kui Ren[1, 2]

[1]The State Key Laboratory of Blockchain and Data Security, Zhejiang University

[2]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security

Paper

Voice Communications are Everywhere

Voice Call

Voice Message

Online Conferencing

Voice Communications are Everywhere

Voice Call

Voice Message

Online Conferencing

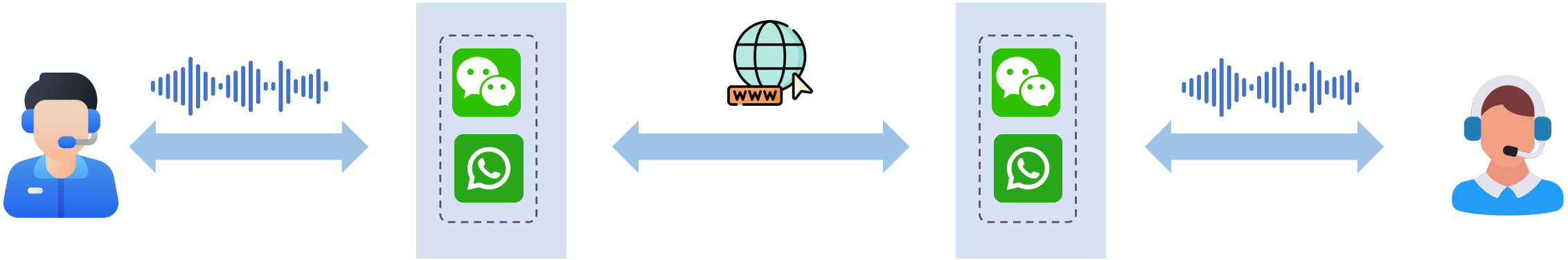
>3 billion

>1 billion

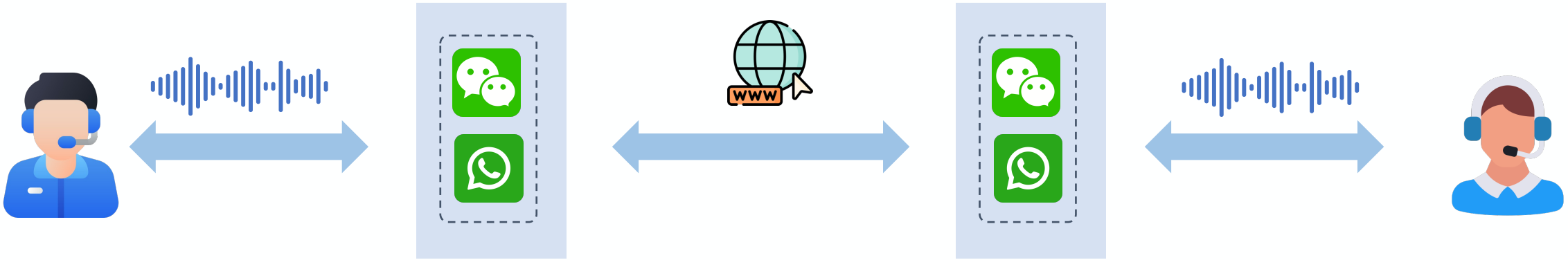
>1 billion

>100 million

Voice Communications are Everywhere



Threats in Voice Communications



Edward Snowden: Leaks that exposed US spy programme

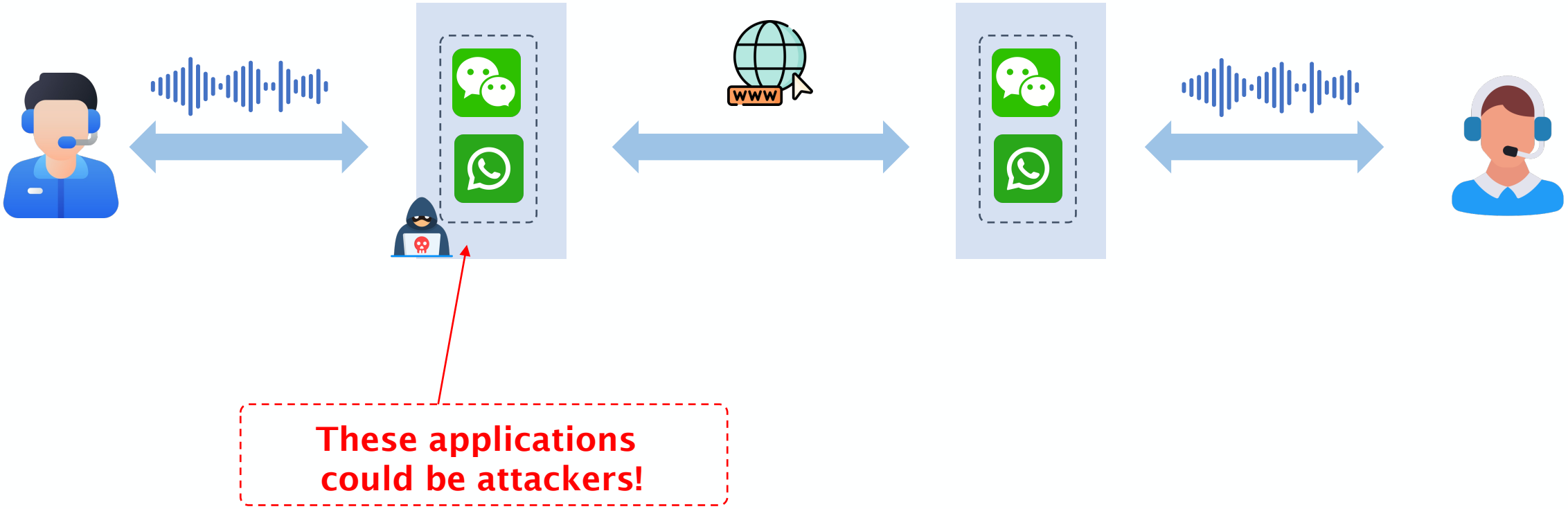
17 January 2014

Hackers could be eavesdropping on your Zoom calls thanks to this flaw

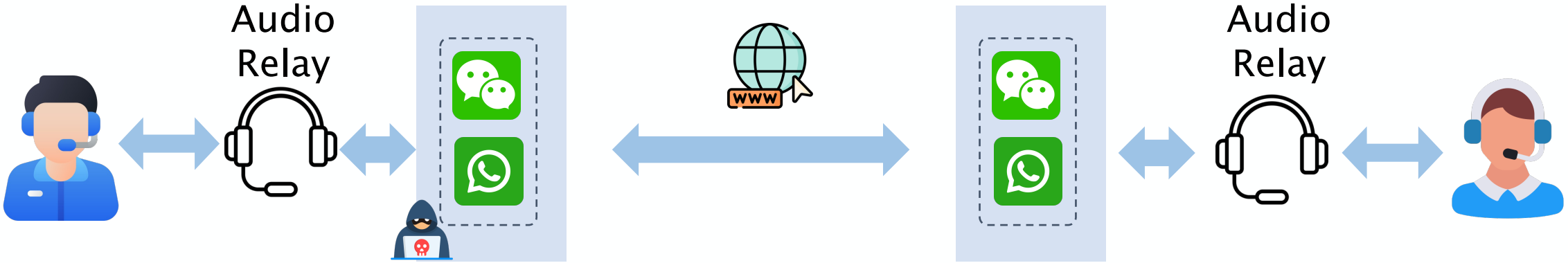
News By Sead Fadilpašić published August 14, 2023

Researchers found worrying new flaws in Zoom

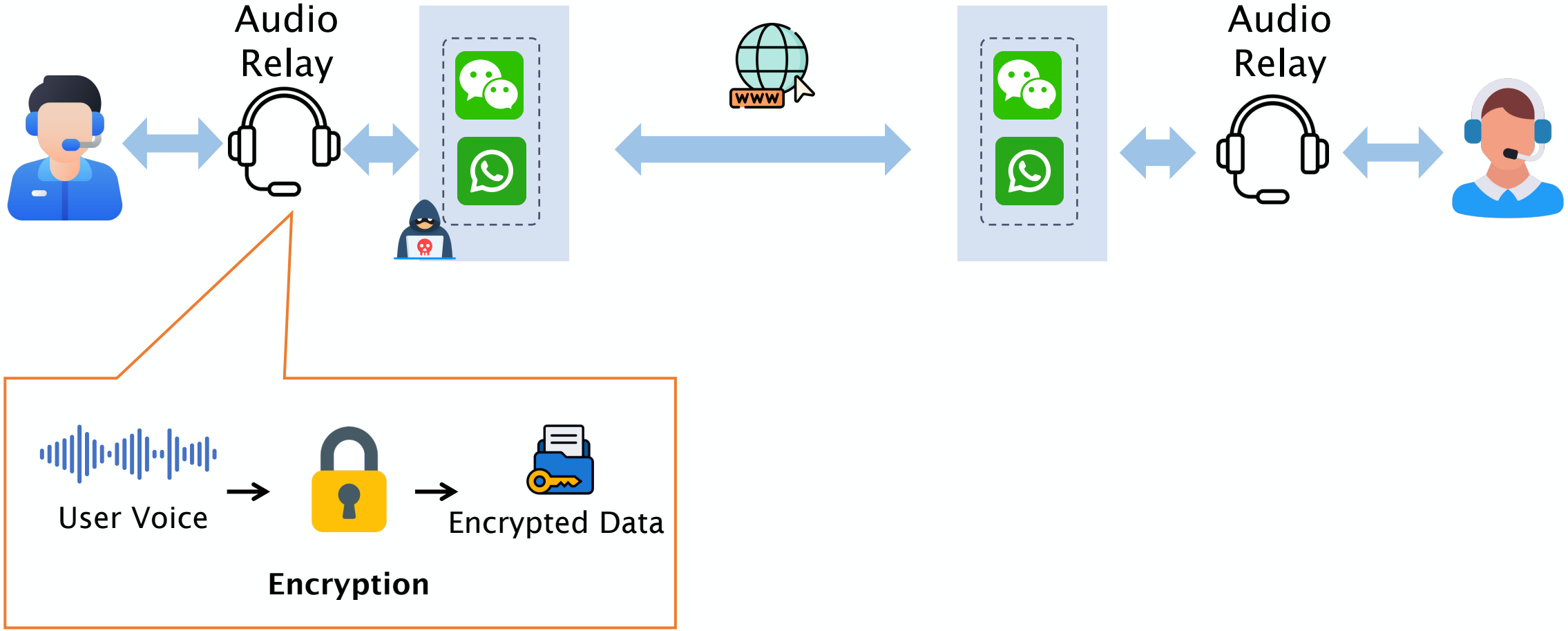
Threats in Voice Communications



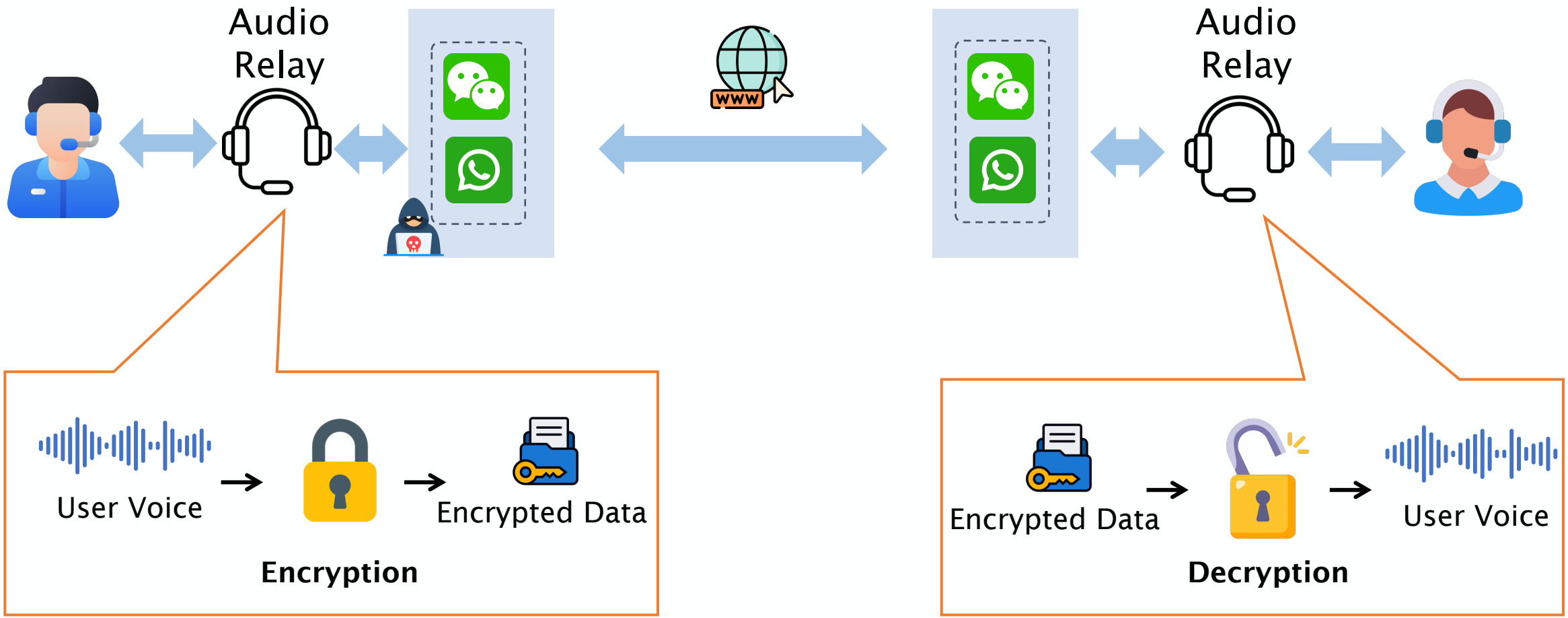
Encryption before Application as Defense



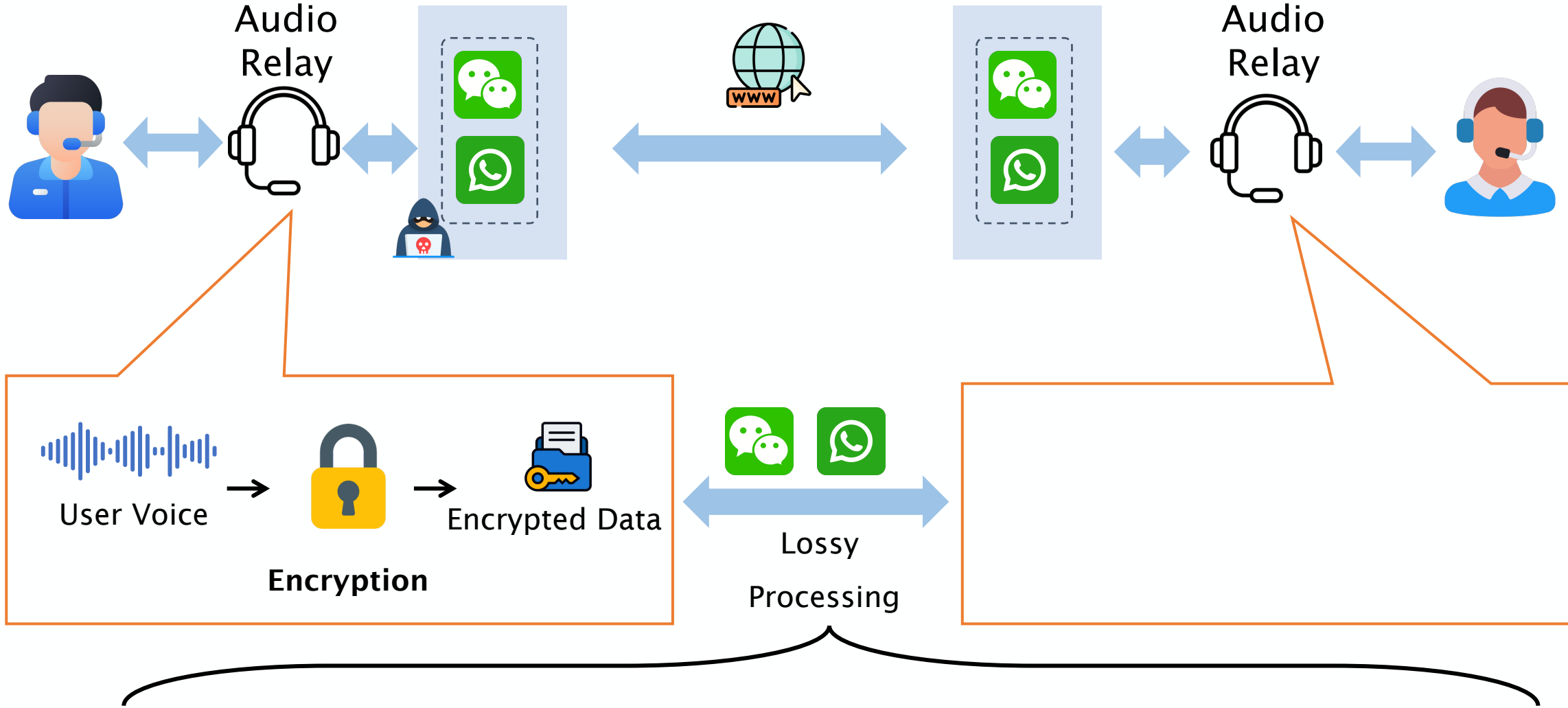
Encryption before Application as Defense



Encryption before Application as Defense

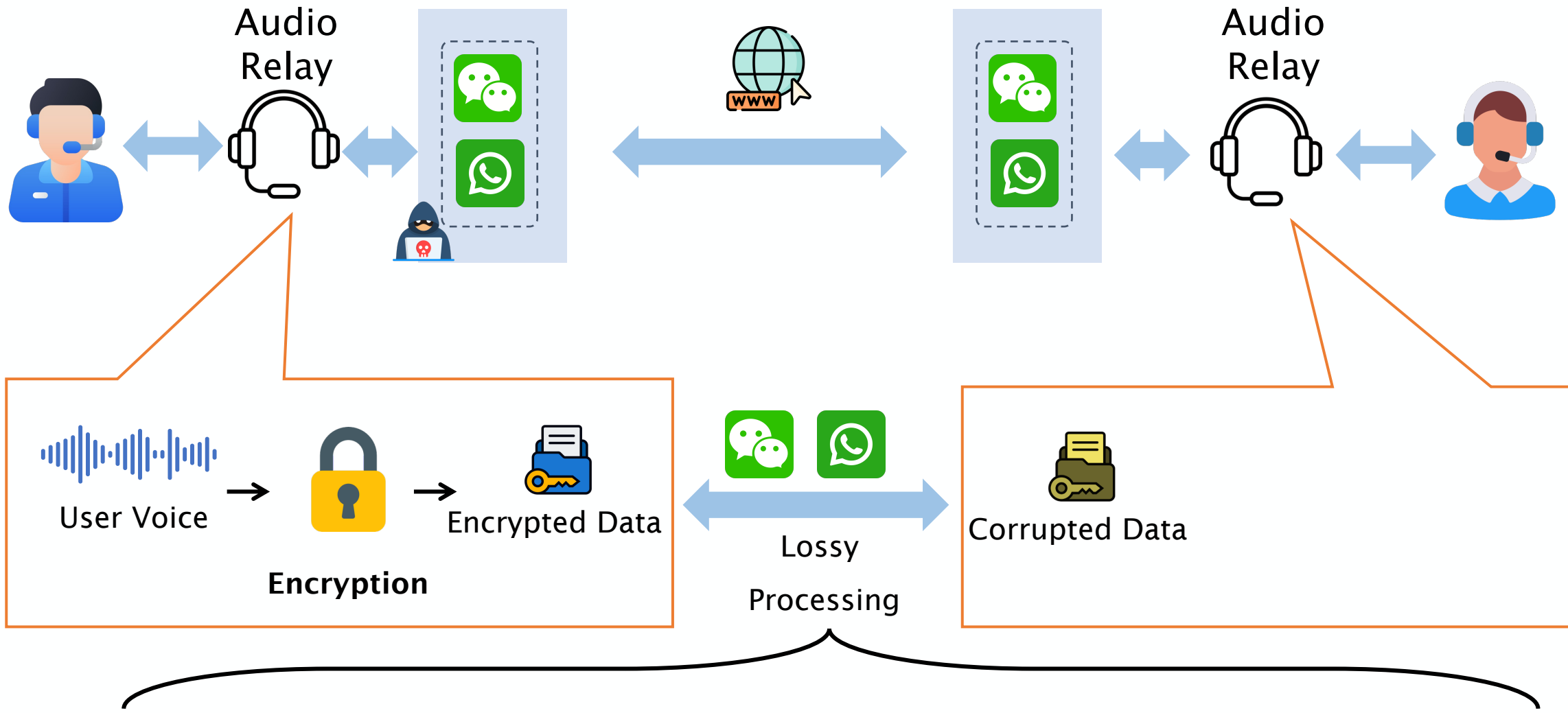


Problem: Distortion during Transmission



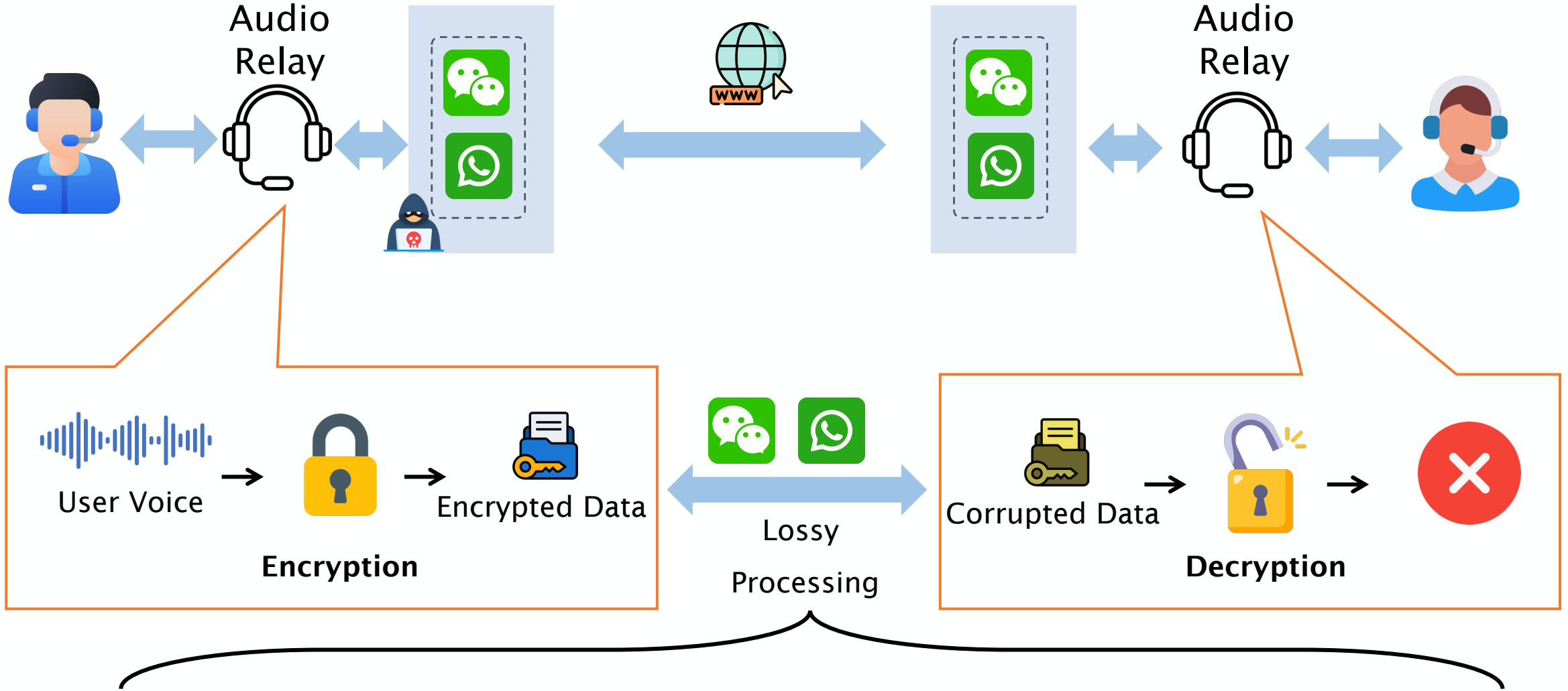
- Noise Suppression
- Audio Compression
- Packet Loss

Problem: Distortion during Transmission



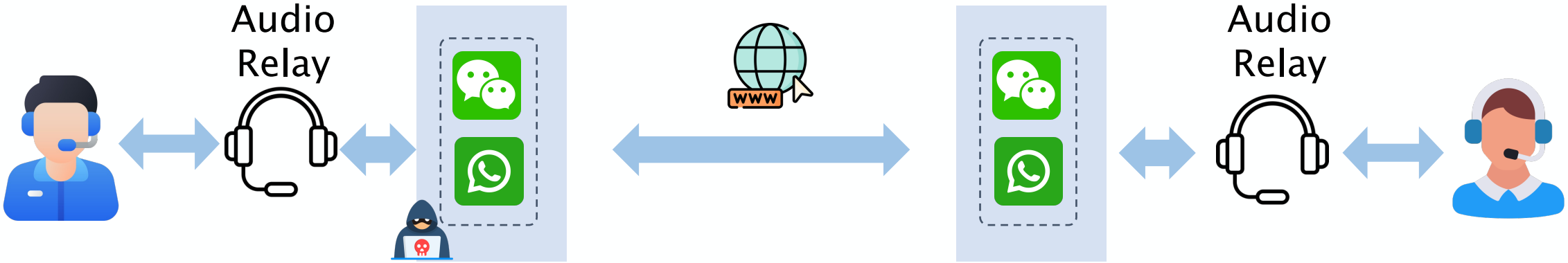
- Noise Suppression
- Audio Compression
- Packet Loss

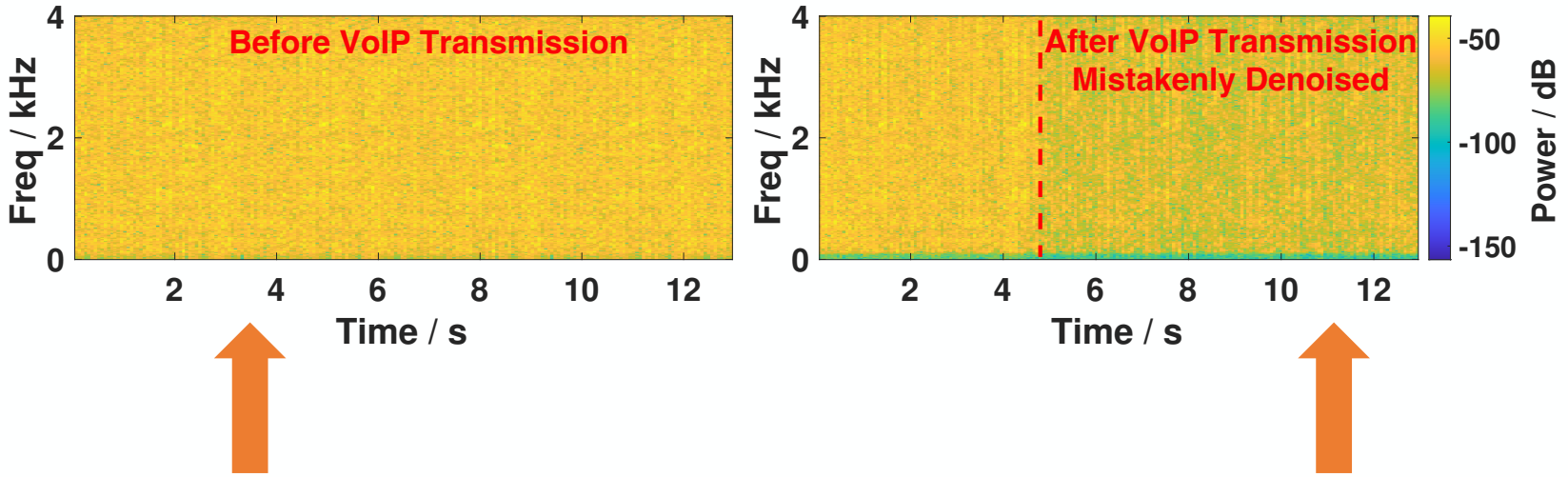
Problem: Distortion during Transmission



- Noise Suppression
- Audio Compression
- Packet Loss

Problem: Distortion during Transmission



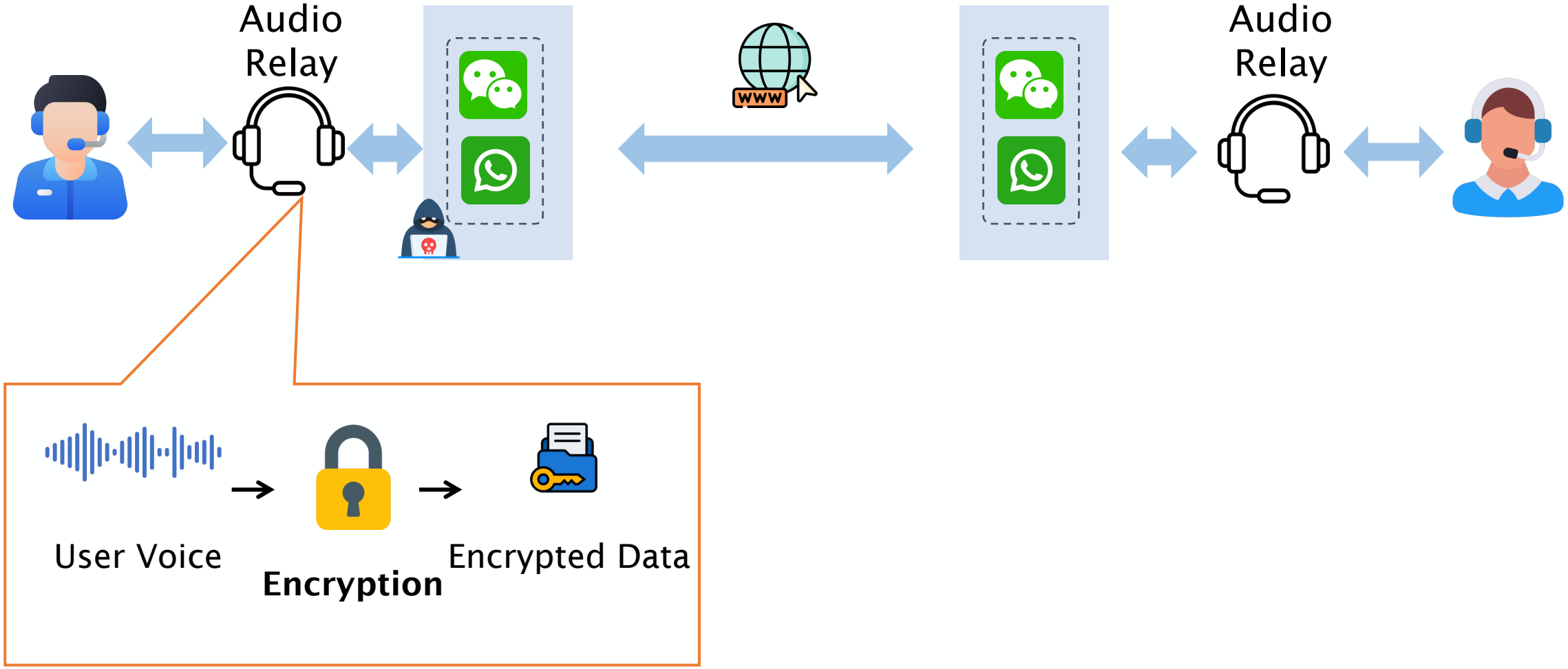


Original encrypted data[1]

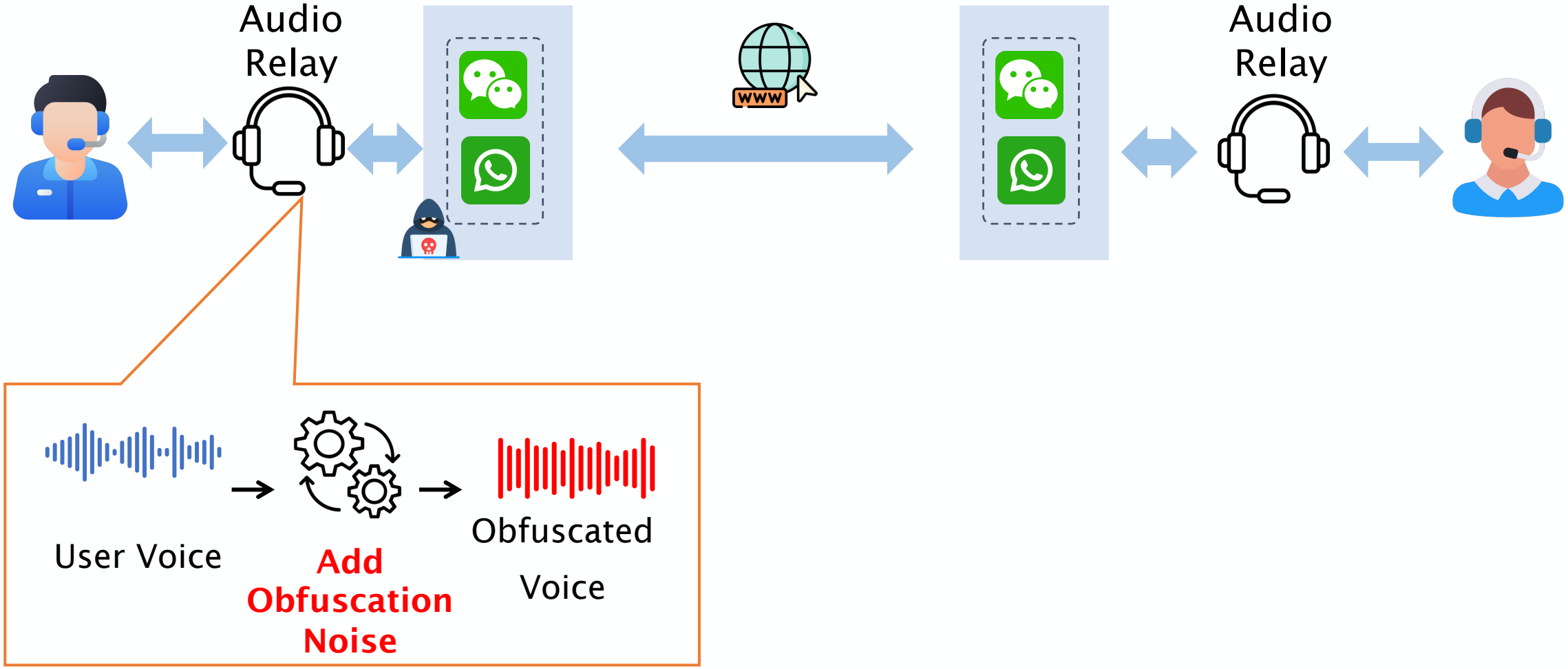
Corrupted encrypted data after lossy transmission

[1] Piotr Krasnowski, Jérôme Lebrun, and Bruno Martin. A novel distortion-tolerant speech encryption scheme for secure voice communication. Speech Commun., 143:57–72, 2022.

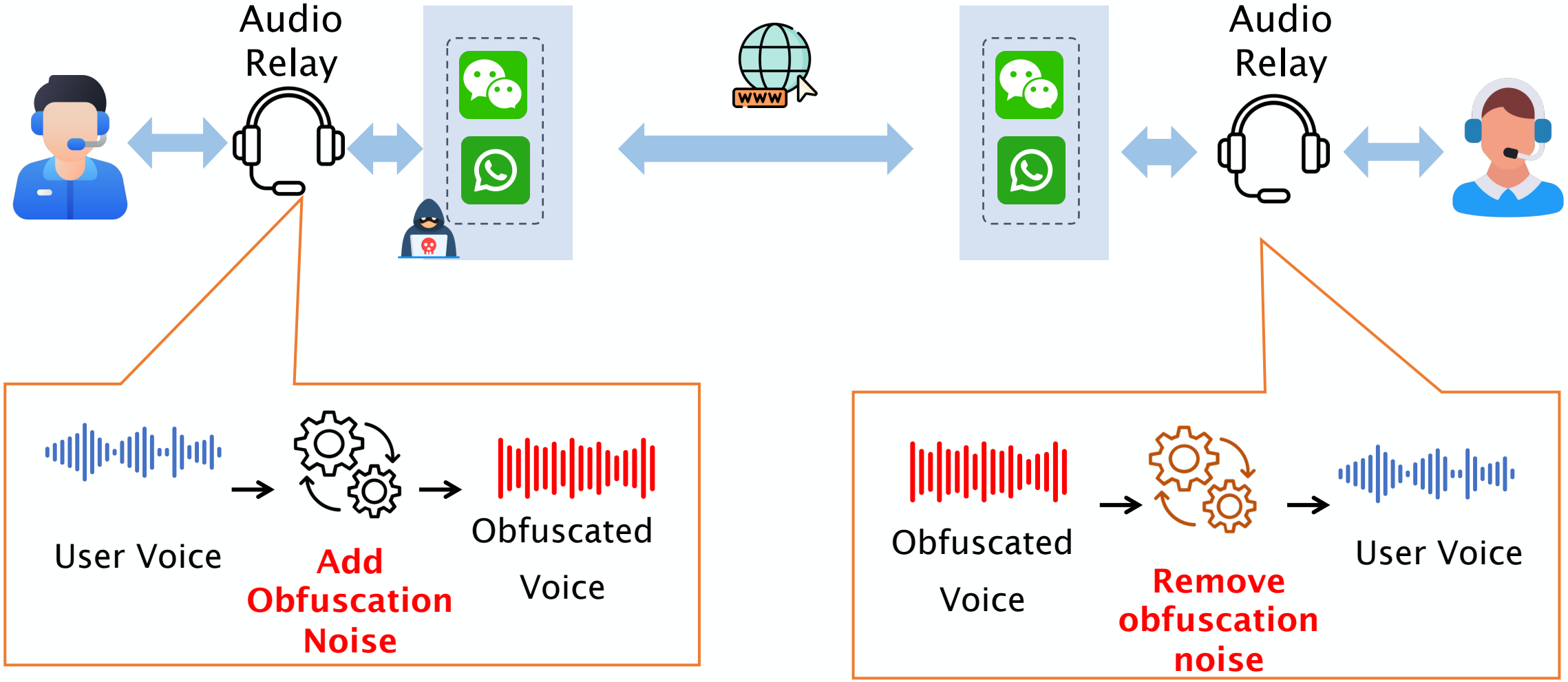
Our Idea: Voice Obfuscation Instead of Encryption



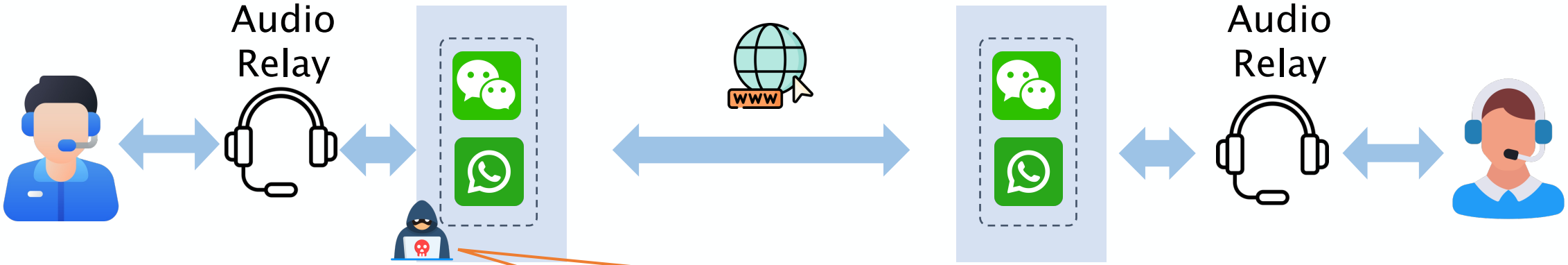
Our Idea: Voice Obfuscation Instead of Encryption



Our Idea: Voice Obfuscation Instead of Encryption

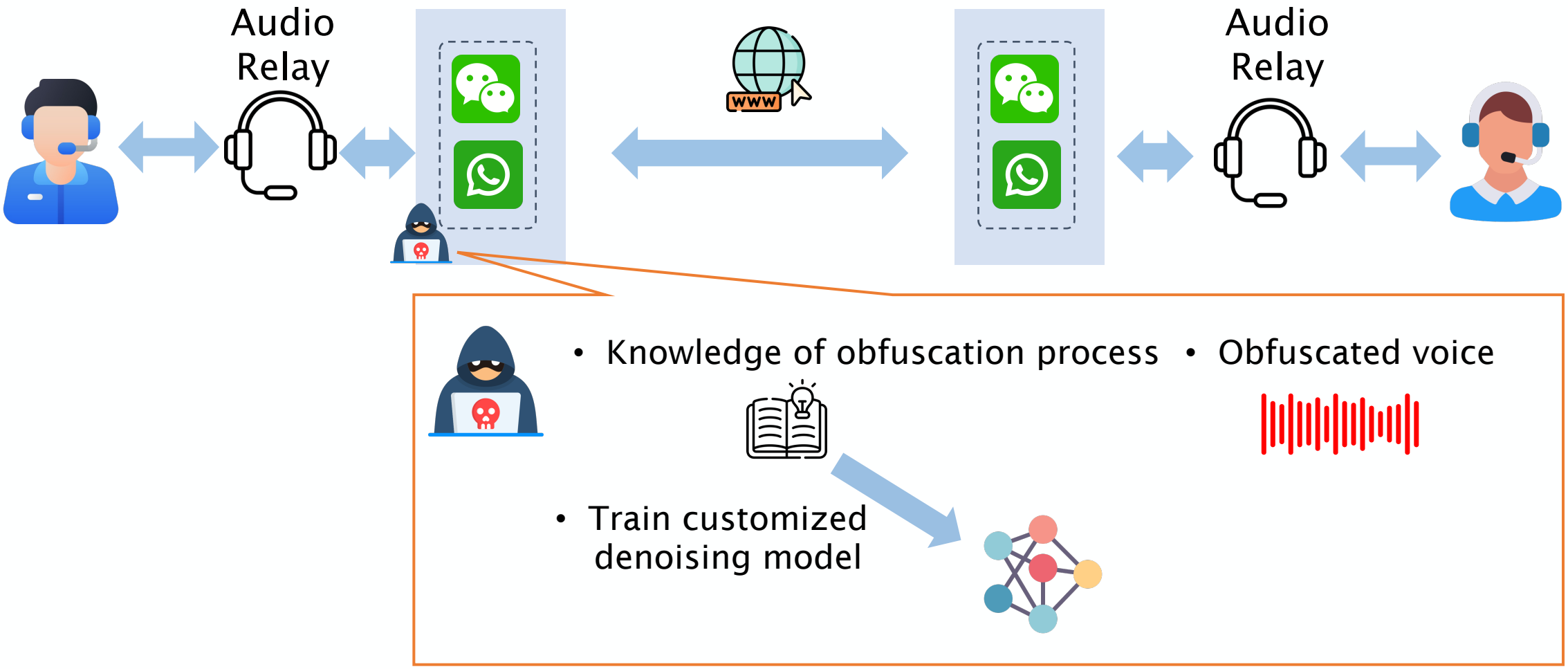


Challenge with Voice Obfuscation

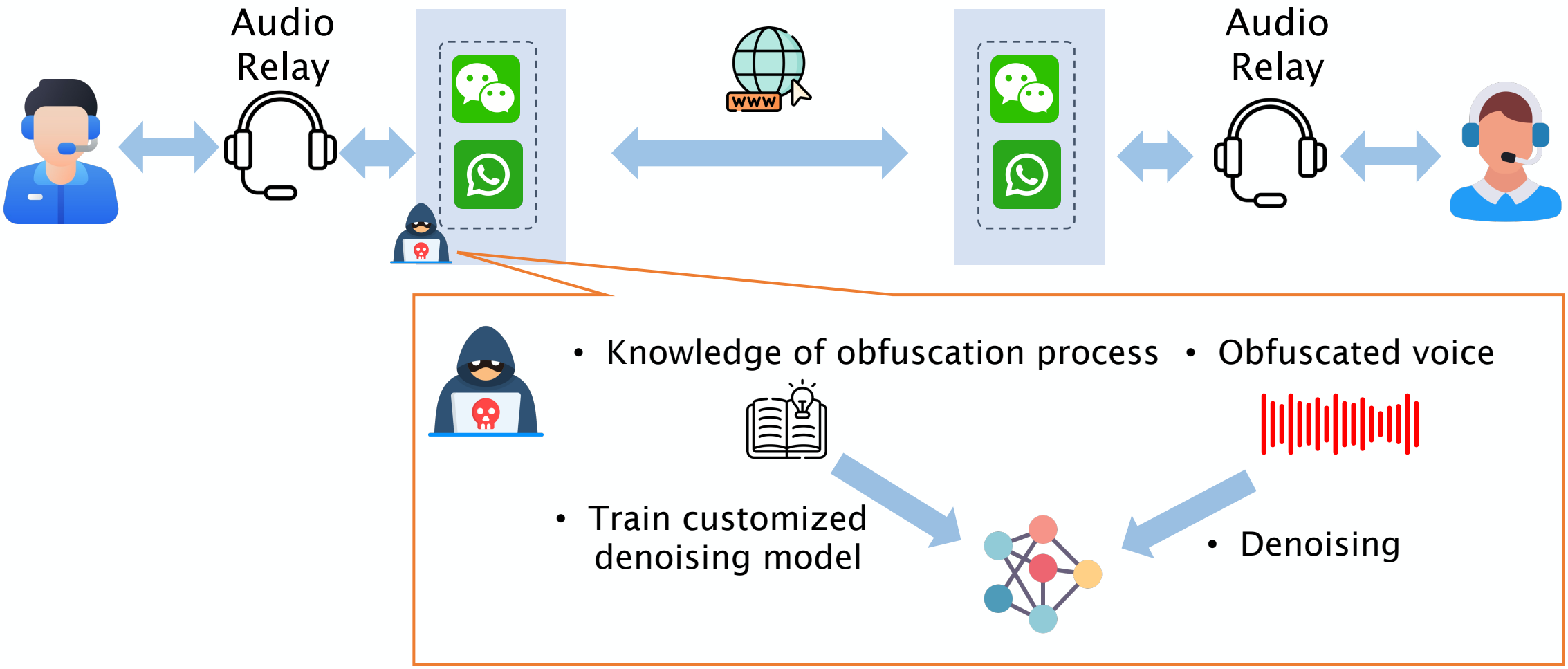


- Knowledge of obfuscation process
- Obfuscated voice

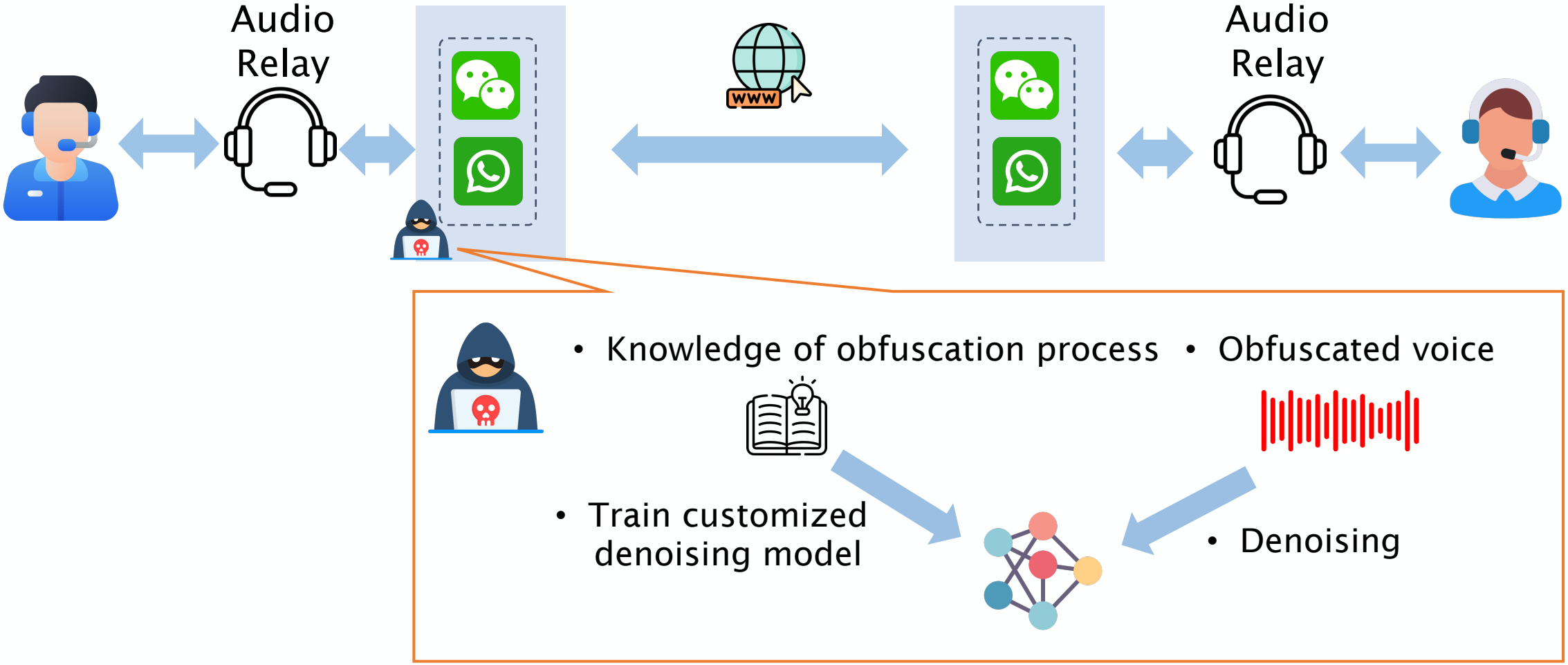
Challenge with Voice Obfuscation



Challenge with Voice Obfuscation



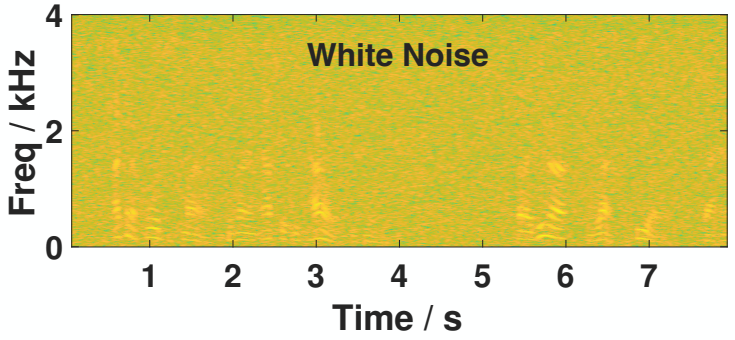
Challenge with Voice Obfuscation



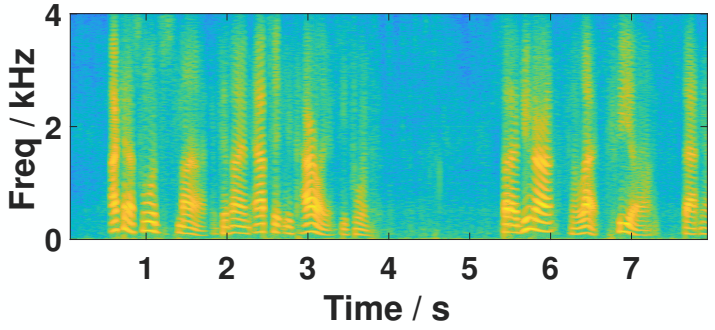
- The obfuscation noise should **be robust against customized denoising model.**

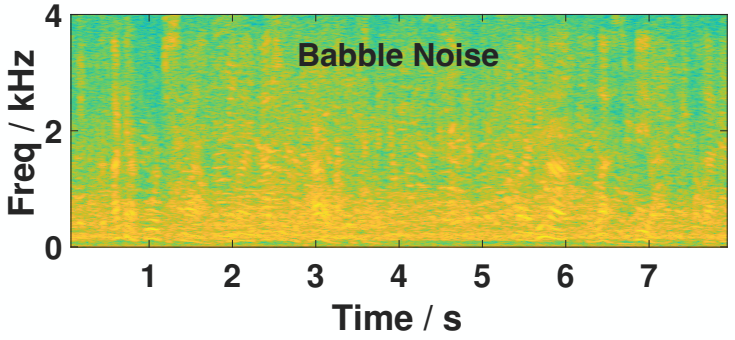
Choice of Obfuscation Noise

- Traditional noises could be easily removed by deep-learning model

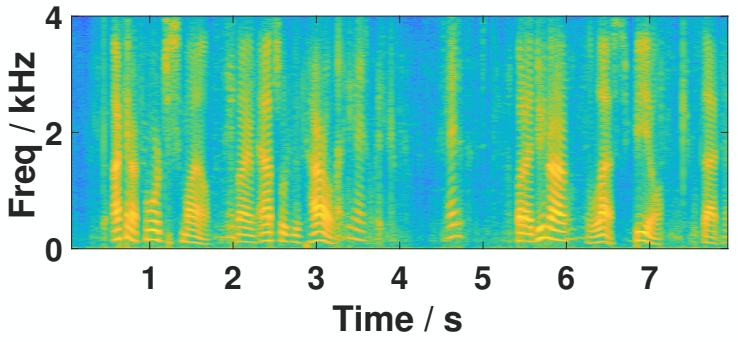


Denoised by [1]





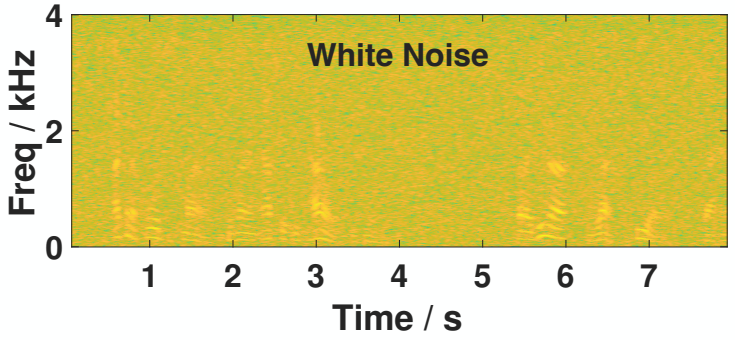
Denoised by [1]



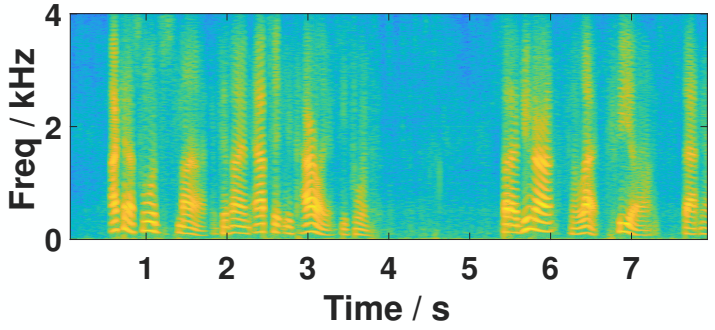
[1] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is all you need in speech separation. ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021, pages 21–25. IEEE, 2021.

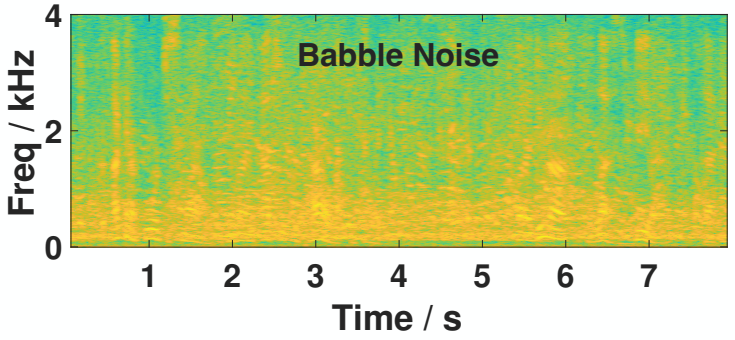
Choice of Obfuscation Noise

- Traditional noises could be easily removed by deep-learning model
  - Noises are not coupled with voices

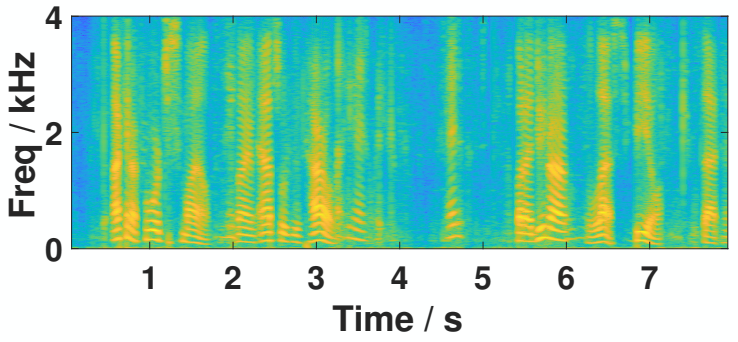


Denoised by [1]





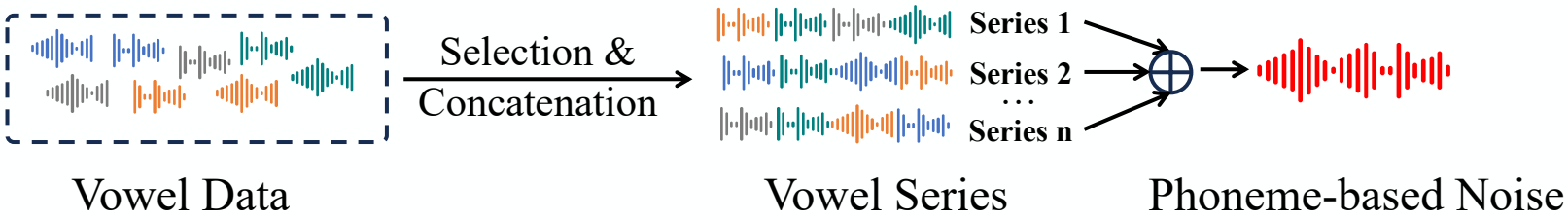
Denoised by [1]



[1] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is all you need in speech separation. ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021, pages 21–25. IEEE, 2021.

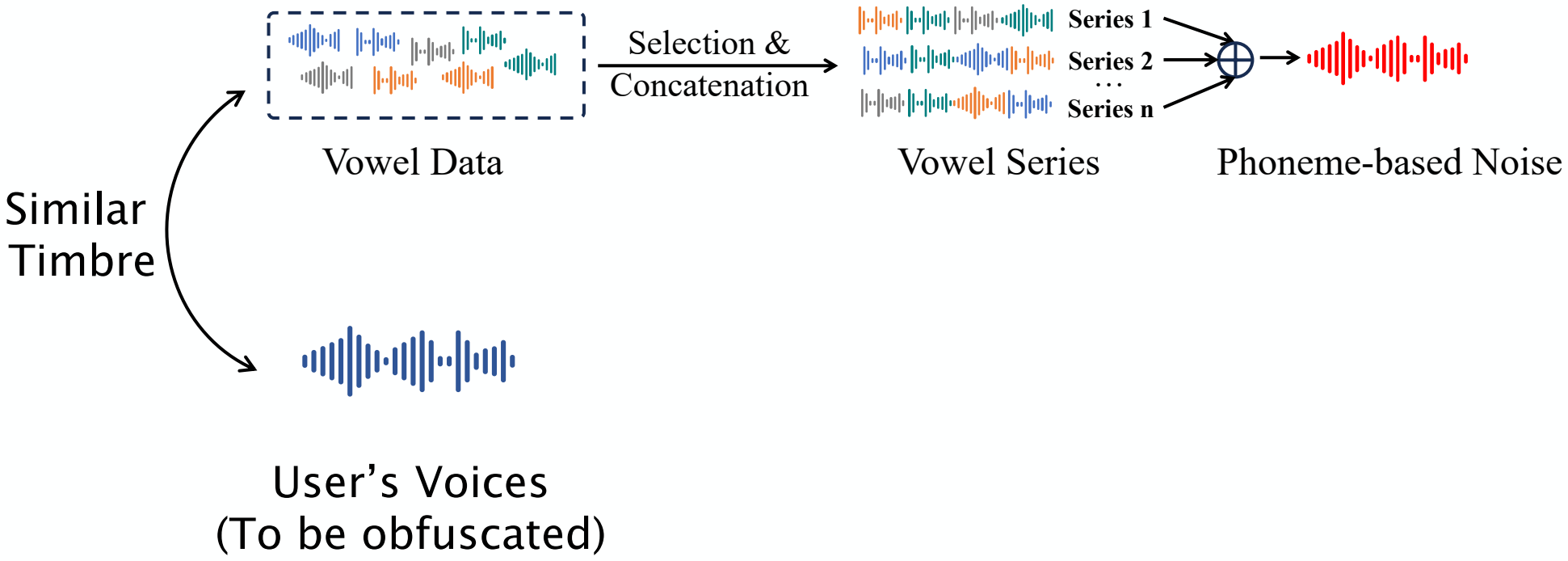
Choice of Obfuscation Noise

- Traditional noises could be easily removed by deep-learning model
  - Noises are not coupled with voices
- Our Design: Phoneme-based Denoising-Resistant Noise



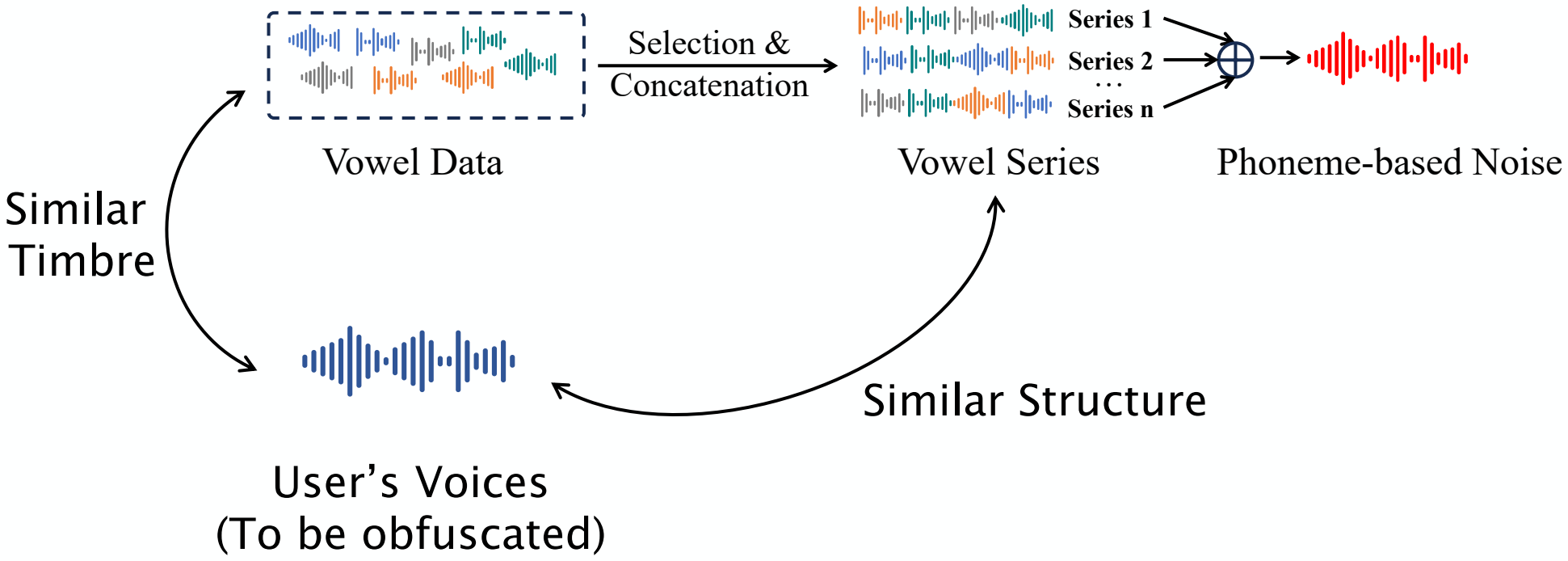
Choice of Obfuscation Noise

- Traditional noises could be easily removed by deep-learning model
  - Noises are not coupled with voices
- Our Design: Phoneme-based Denoising-Resistant Noise



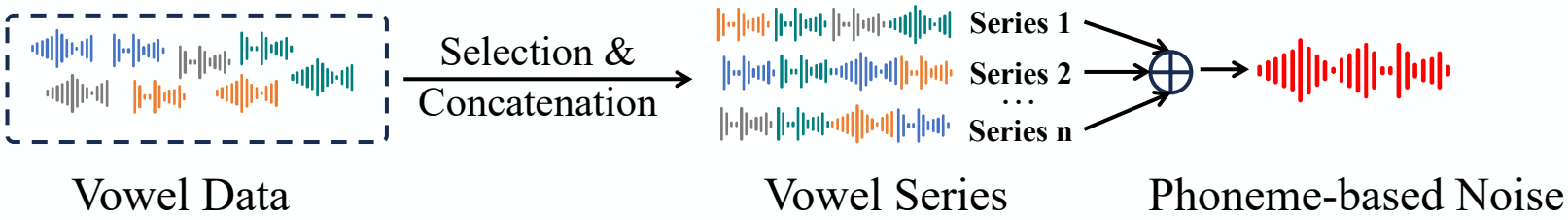
Choice of Obfuscation Noise

- Traditional noises could be easily removed by deep-learning model
  - Noises are not coupled with voices
- Our Design: Phoneme-based Denoising-Resistant Noise



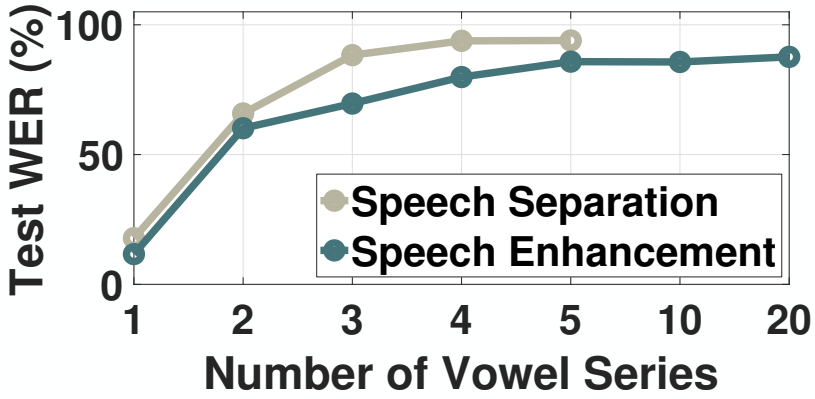
Choice of Obfuscation Noise

- Traditional noises could be easily removed by deep-learning model
  - Noises are not coupled with voices
- Our Design: Phoneme-based Denoising-Resistant Noise

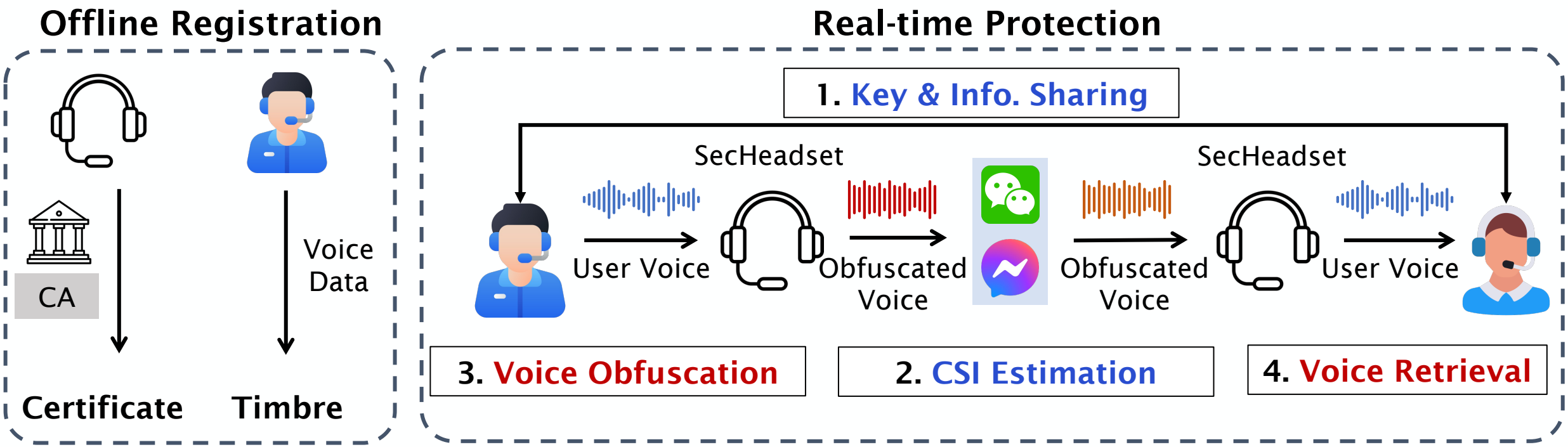


- Training customized denoising models to validate noise robustness

When the number of vowel series >=3, even customized denoising models could not remove the noise (WER > 70%)

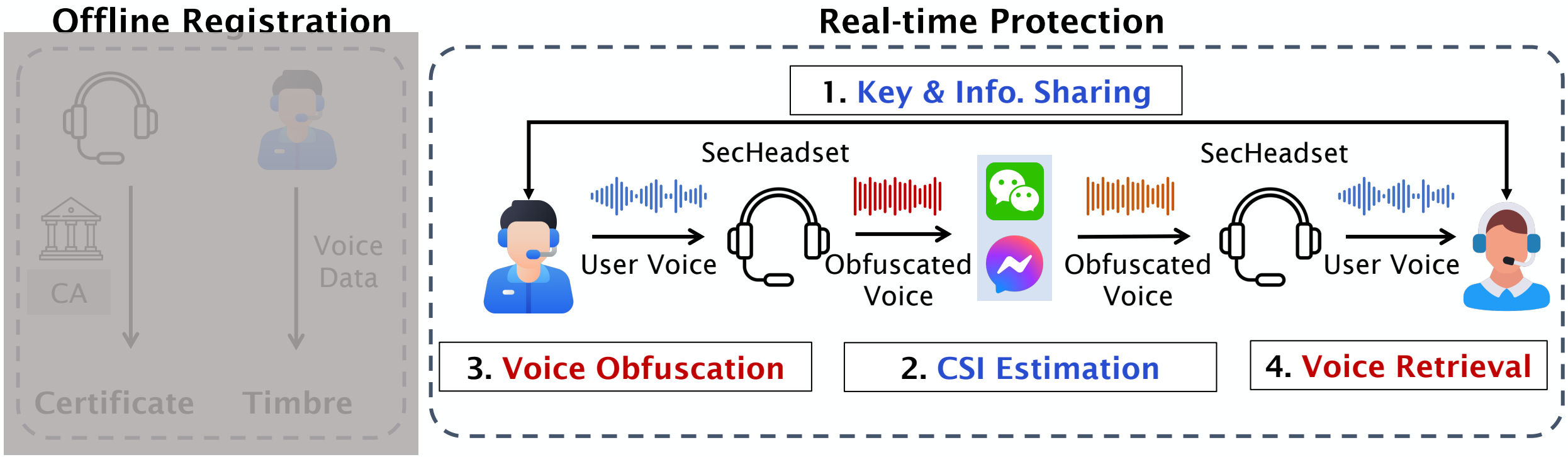


System Overview

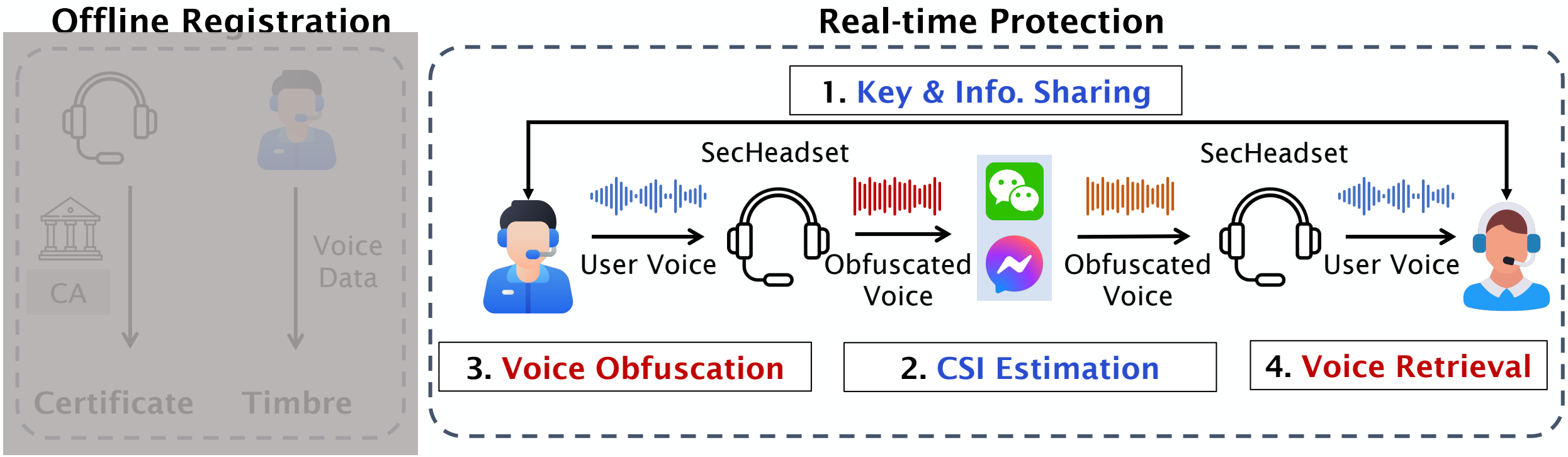


System Overview

System Overview



System Overview

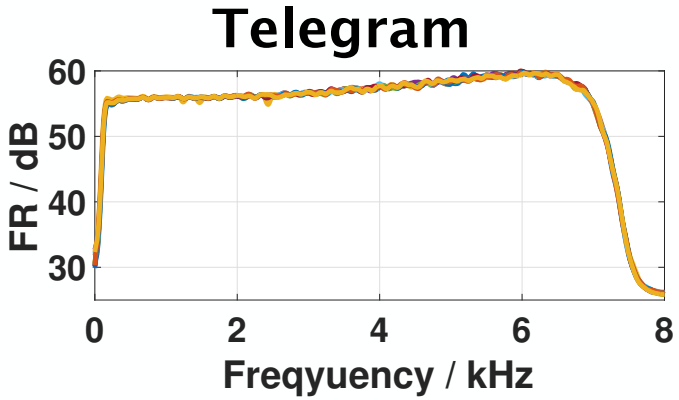


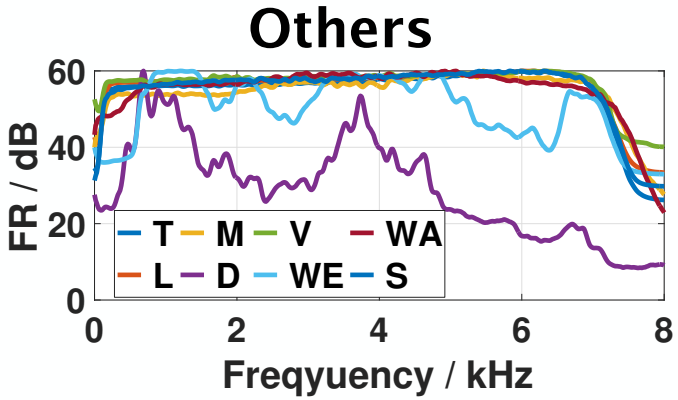
1. **Key & Info. Sharing:** public key, user timbre, session random number *(RN)*

3. **Voice Obfuscation:** Generate noise based on timbre and *RN*, add to voices

CSI Estimation

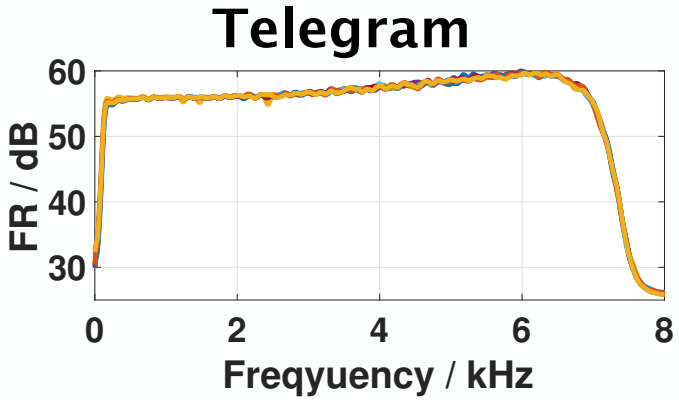
- Frequency responses are stable in each application, while different across them

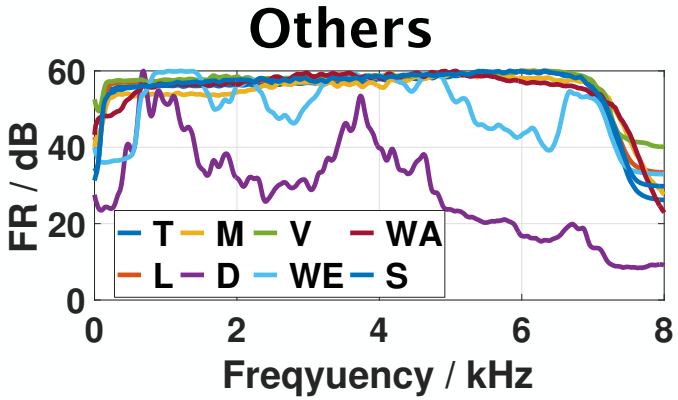




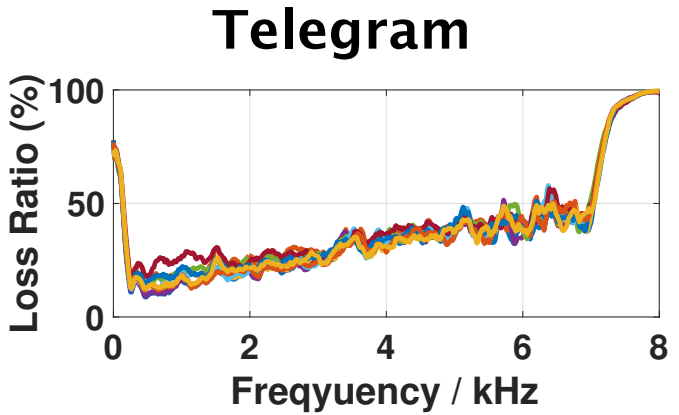
CSI Estimation

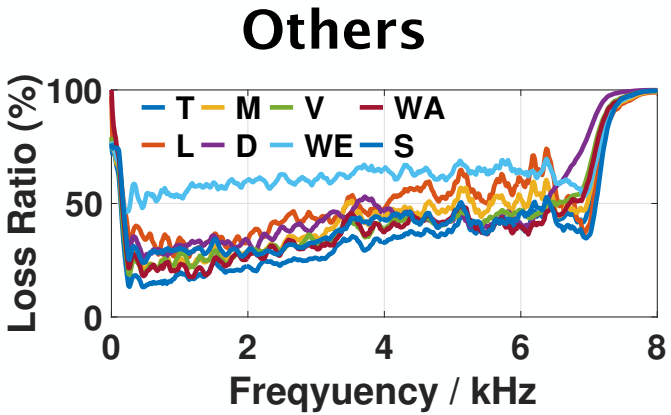
- Frequency responses are stable in each application, while different across them





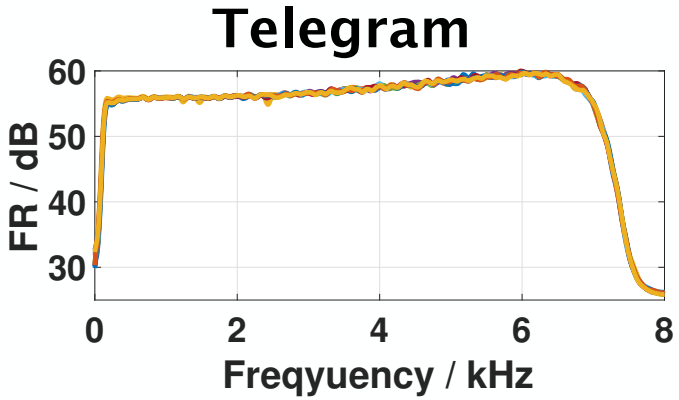
- So as the compression loss ratio

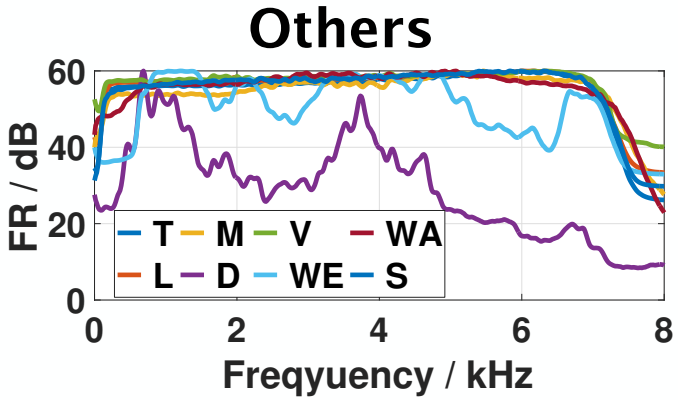




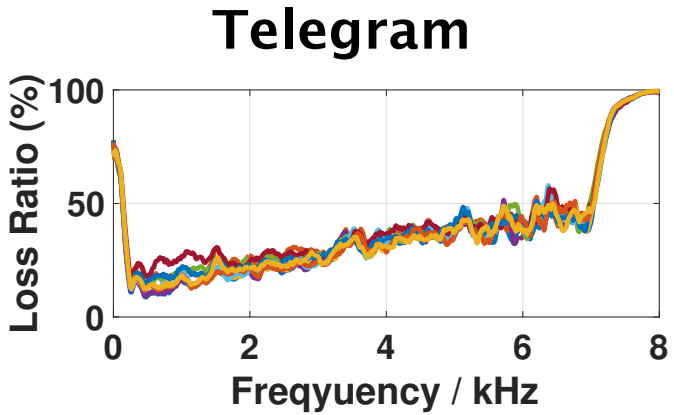
CSI Estimation

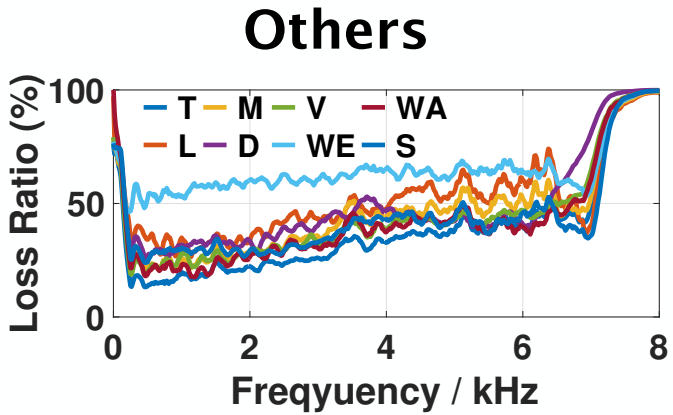
- Frequency responses are stable in each application, while different across them





- So as the compression loss ratio



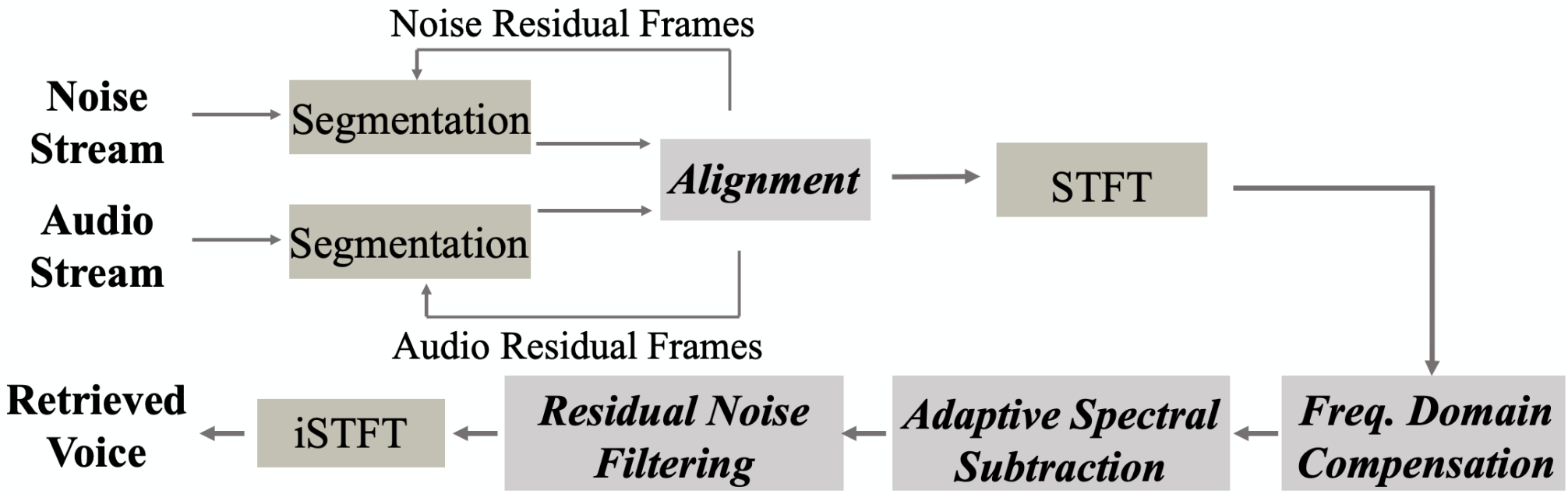


Estimate CSI with a one-time probe exchange at the start of communication

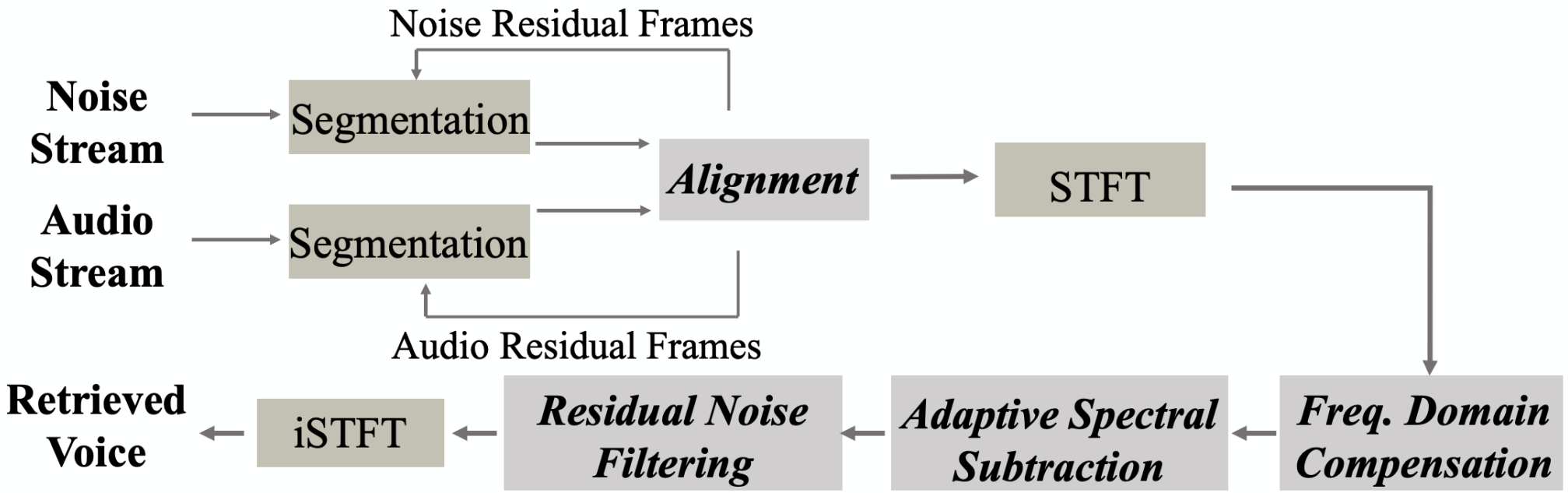
Voice Retrieval with Adaptive Spectral Subtraction

- Deep learning-based models could be effective, but hard to achieve real-time in edge devices

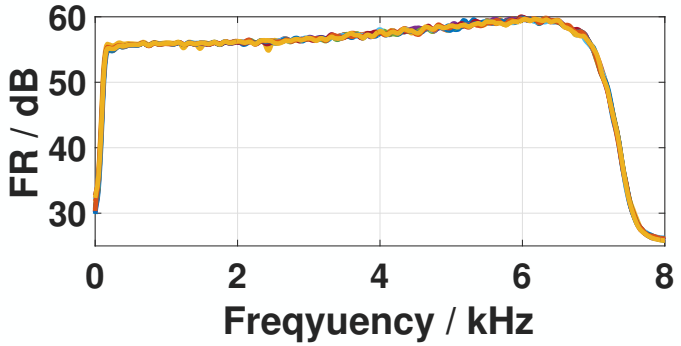
Voice Retrieval with Adaptive Spectral Subtraction



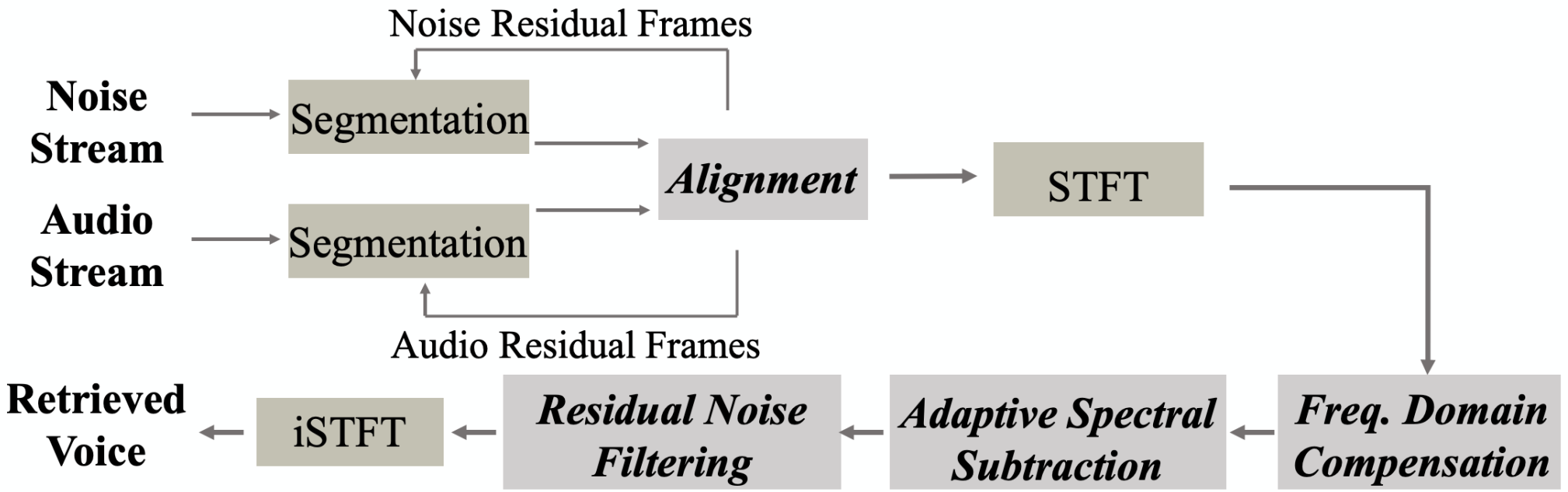
Voice Retrieval with Adaptive Spectral Subtraction



Frequency Response



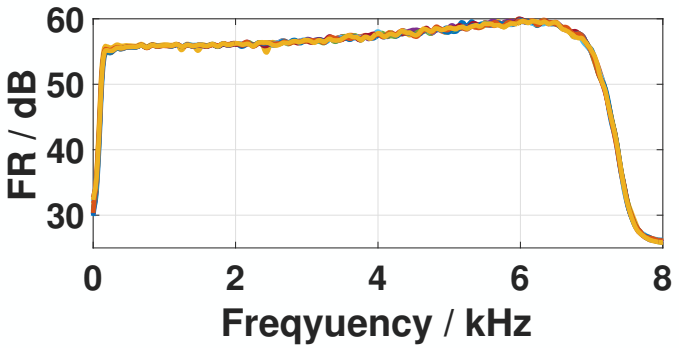
Voice Retrieval with Adaptive Spectral Subtraction



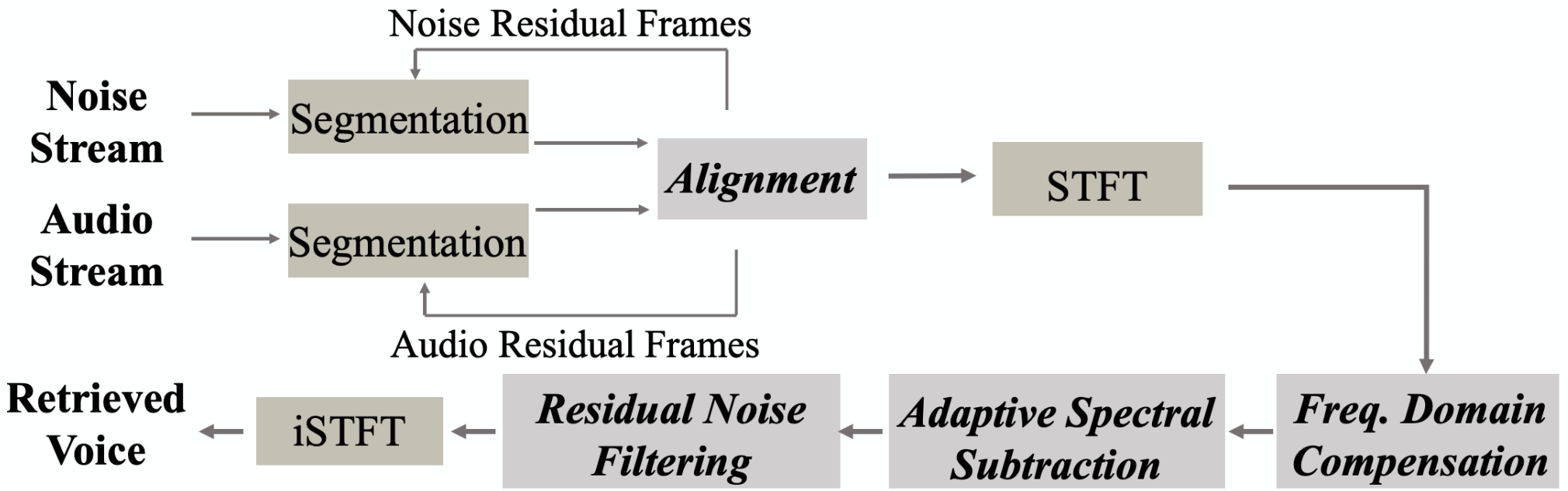
$$\hat{S}(f,t) = Y(f,t) \cdot \left(1 - \frac{||\alpha(f) \cdot N(f,t)||}{||Y(f,t)||}\right)$$

$$\min_{\alpha(f)} L = \sum_f \sum_t ||\hat{S}(f,t)||$$

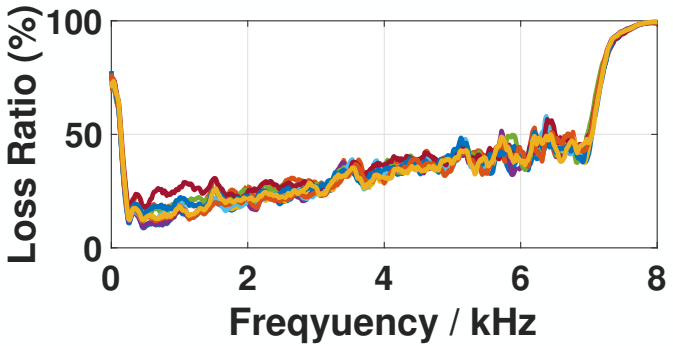
Frequency Response



Voice Retrieval with Adaptive Spectral Subtraction



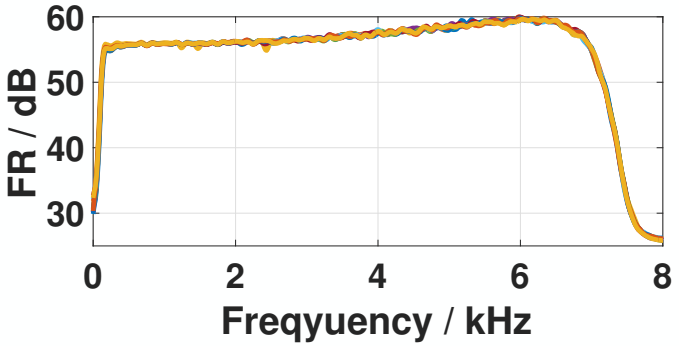
Loss Ratio



$$\hat{S}(f,t) = Y(f,t) \cdot \left(1 - \frac{||\alpha(f) \cdot N(f,t)||}{||Y(f,t)||}\right)$$

$$\min_{\alpha(f)} L = \sum_f \sum_t ||\hat{S}(f,t)||$$

Frequency Response



Evaluation Setup

- Tested Channels: VoIP and Voice Message
- Tested Applications:


WeChat


Messenger

Line


DingTalk


Telegram


Viber


WhatsAPP


Skype

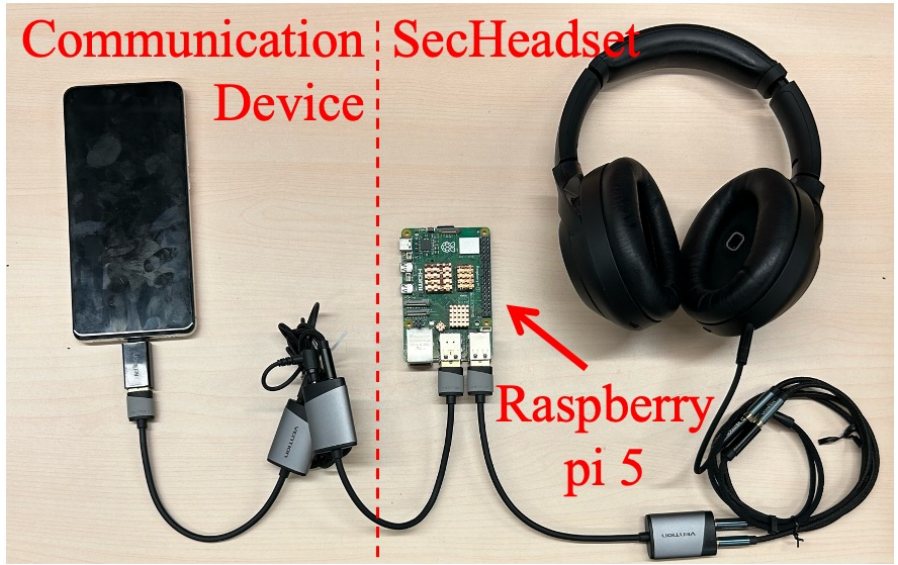
Evaluation Setup

- Tested Channels: VoIP and Voice Message
- Tested Applications:

WeChat Messenger Line DingTalk

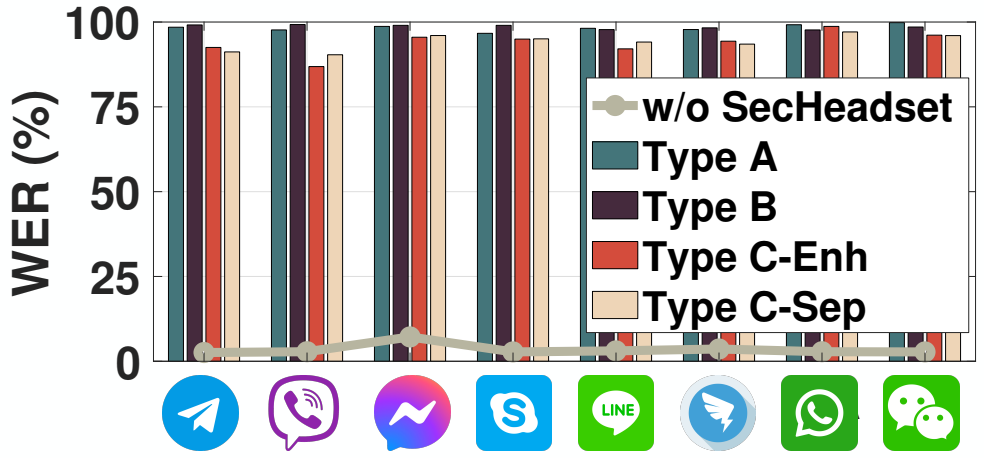
Telegram Viber WhatsAPP Skype

- Software: Python 3.10
- Hardware Prototype:

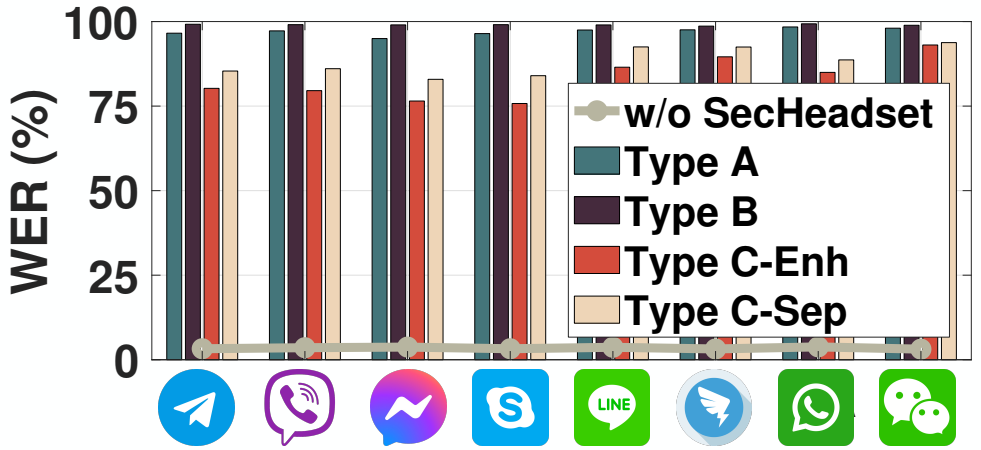


Evaluation Results

- Attacker's recognition error rate of obfuscated voices, **higher is better**
- Type A & B: Attackers **without** knowledge of our system
- Type C: Attackers **with** knowledge of our system



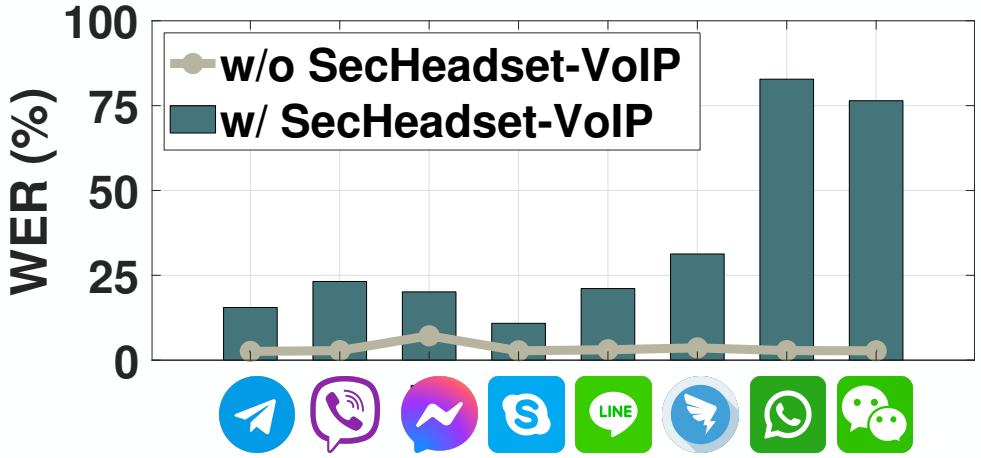
Attacker in VoIP Channel



Attacker in Voice Message Channel

Evaluation Results

- User's recognition error rate of obfuscated voices, **lower is better**

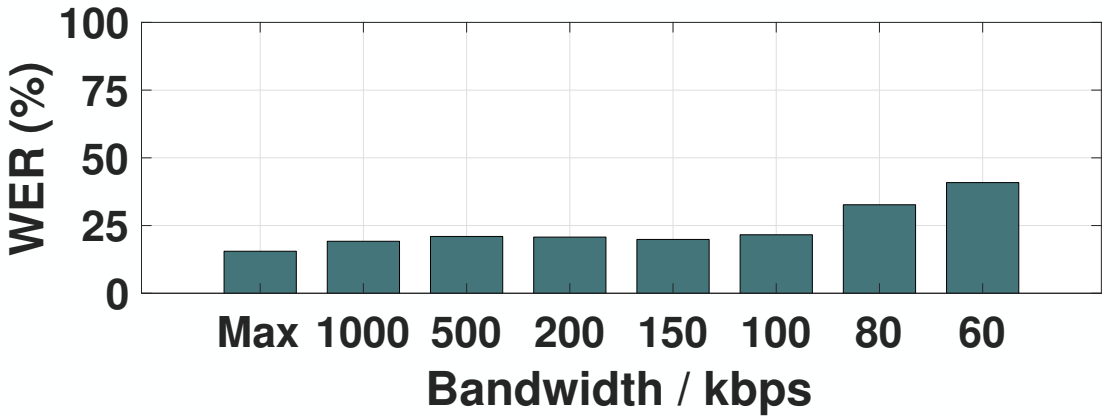


Users in VoIP Channel

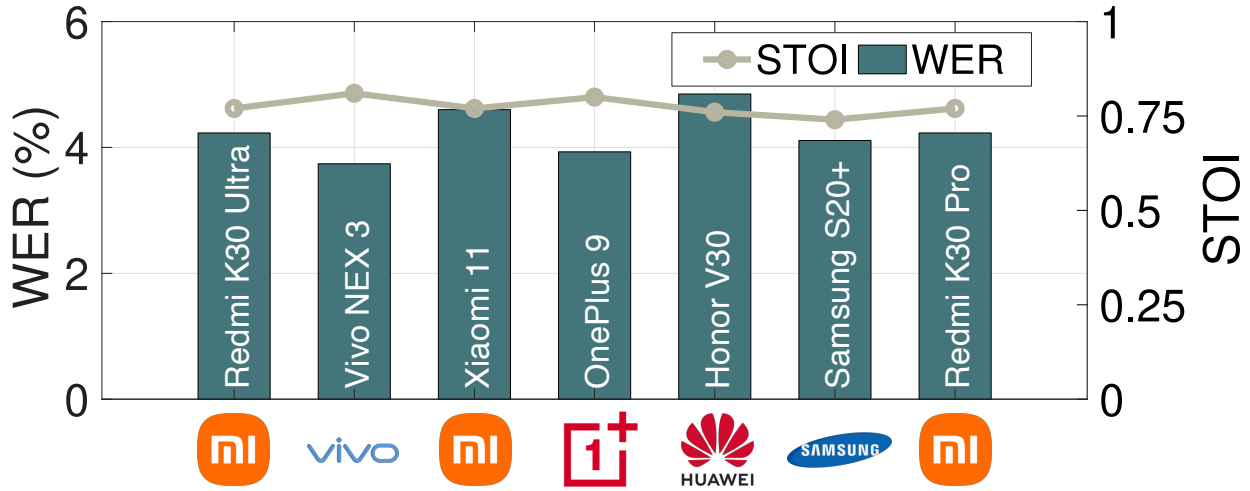
Users in Voice Message Channel

Evaluation Results

- Impact of network bandwidth



- Impact of device



Evaluation Results

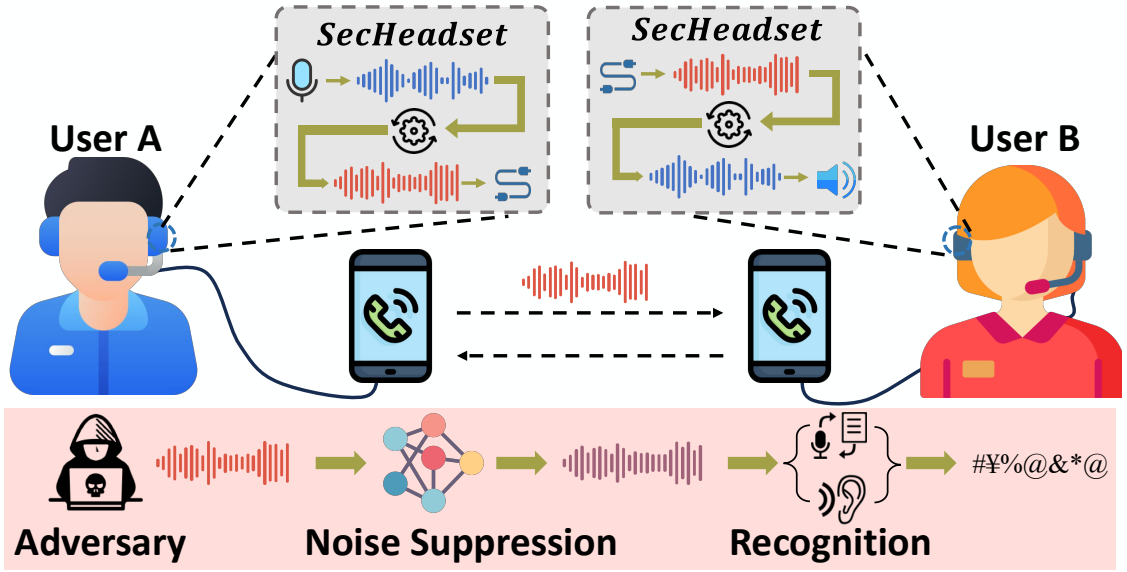
- Time consumption in voice retrieval
  - 17.6 ms for each audio block (64 ms), achieves real-time processing
  - Total delay: 64 ms + 17.6 ms = 81.6 ms

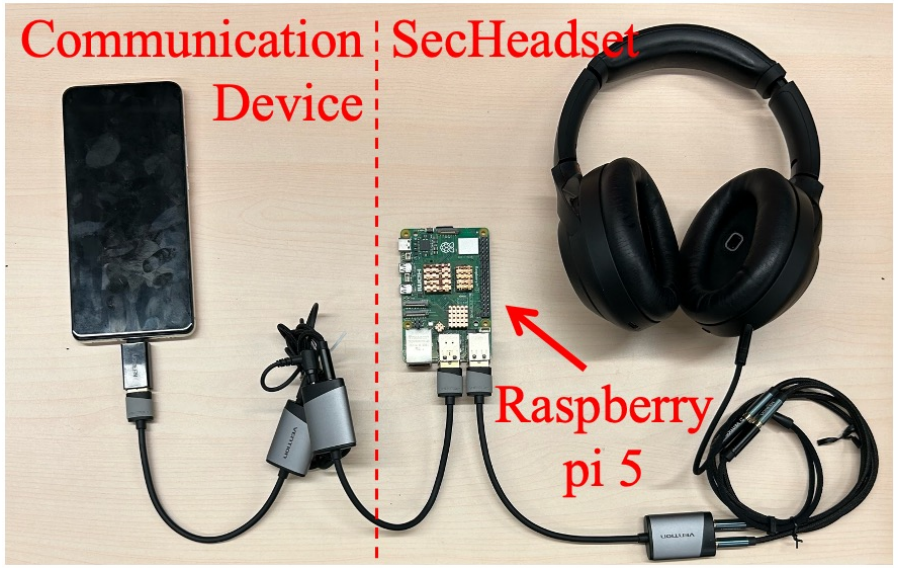
Module	Alignment	STFT	Spec. Sub.	iSTFT	Res. Filter
Times(ms)	0.28	1.00	12.30	0.65	3.60

Conclusion

- Introduce **SecHeadset** to protect voice privacy with COTS edge devices

- SecHeadset works well in **commonly used voice communication applications**





Thank You!

Paper

Peng Huang, Kun Pan, Qinglong Wang, Peng Cheng,
Li Lu, Zhongjie Ba, Kui Ren
penghuang@zju.edu.cn

I'm looking for postdoc position!

My Homepage