MicPro: Microphone-based Voice Privacy Protection

Shilin Xiao, Xiaoyu Ji*, Chen Yan, Zhicong Zheng, Wenyuan Xu
Ubiquitous System Security Lab. (USSLAB), Zhejiang University





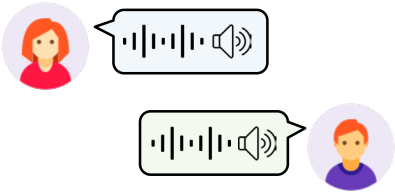
Millions of Voices are Recorded Every Minute

Content: required by apps

Voiceprint: may not required

⇩

Privacy leakage



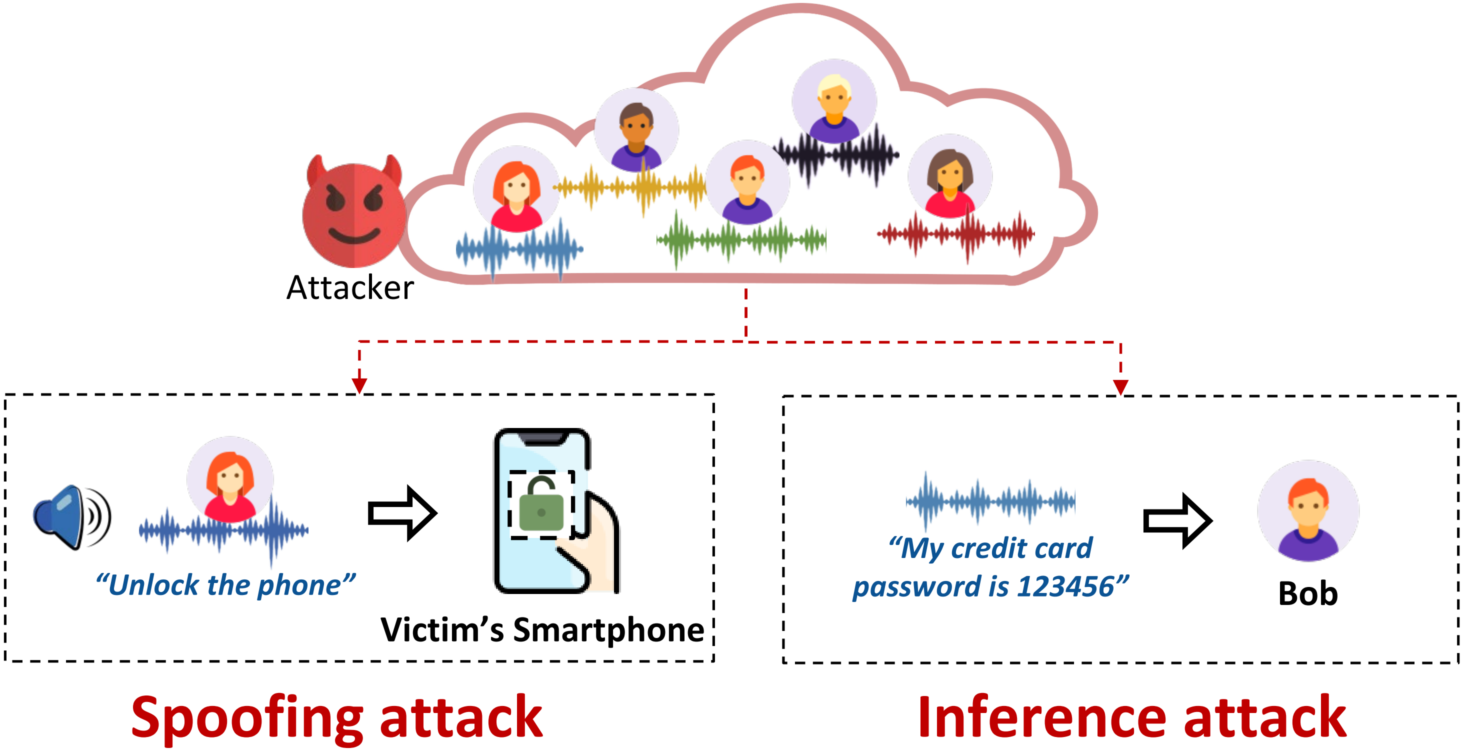
"Hey Siri"

"Hey Google"

Virtual Meetings **Voice Messages** **Voice Assistants**

Voiceprints are inevitably leaked along with these voice clips!

Two Types of Attacks Utilizing Voiceprints

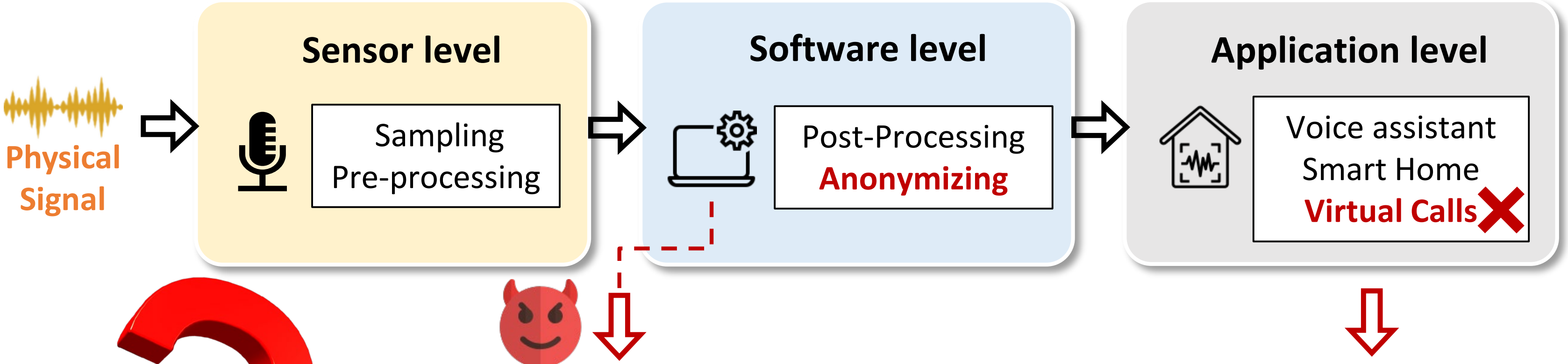


Voiceprint Protection: Speech Anonymization

- Existing anonymization methods use *audio clips* at the *software level*

Voiceprint Protection: Speech Anonymization

- Limitations of existing methods

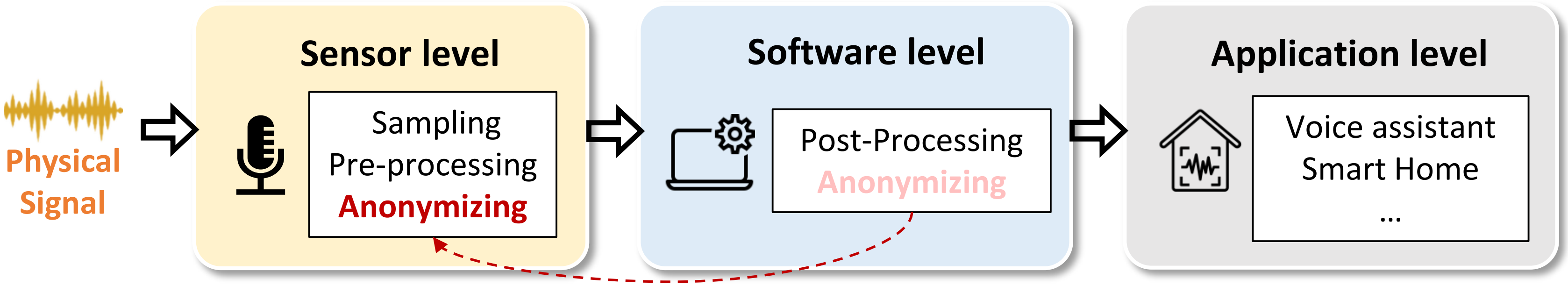


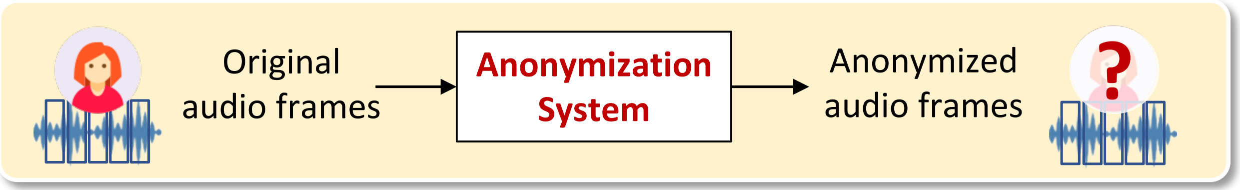
1. Security: the original audio can be stolen

2. Usability: real-time constrains (need audio frames)

Sensor-level Anonymization

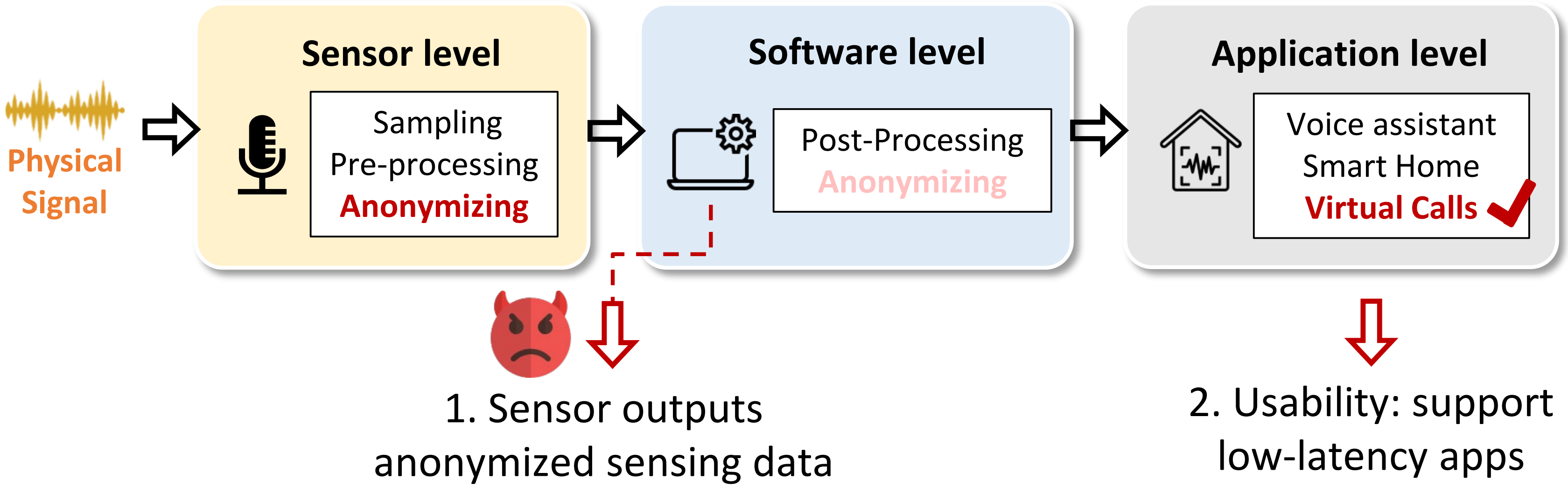
- We can anonymize *audio frames* at the *sensor level*



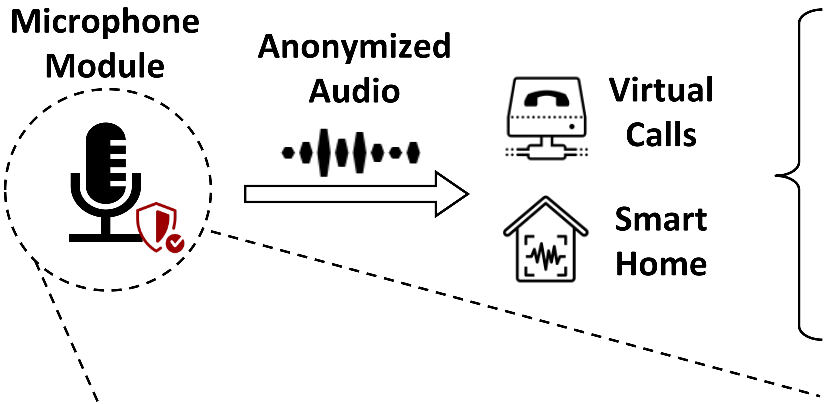


Sensor-level Anonymization

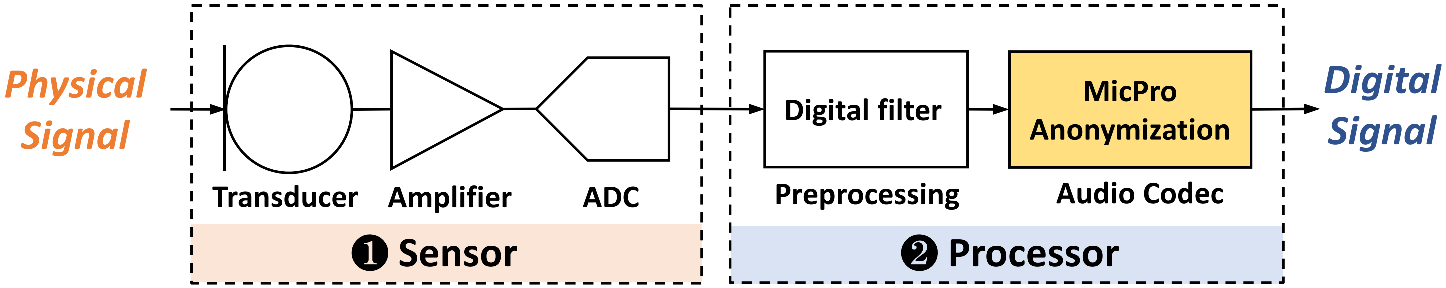
- What's the benefit?



MicPro: Privacy-by-Design Microphone



1. **Anonymity**
   - *Spoofing attack*
   - *Inference attack*
2. **Usability**
   - *Intelligibility & Naturalness for listeners*
   - *Low latency for transmissions*
   - *Lightweight for implementations*



Key Challenges for the Design

- A privacy-by-design microphone module requires:

1. No hardware modification
2. Low computational overhead

Q1: How to achieve anonymity without hardware modifications?

A1: Utilize the built-in parameters, e.g. line spectral frequency, in a popular *audio codec*

Q2: How to achieve anonymity and usability at the same time?

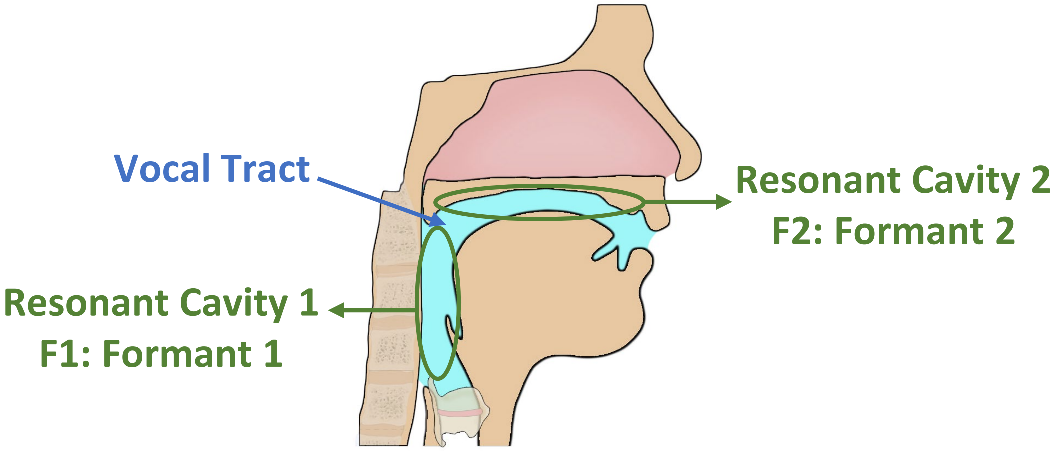
A2: Formulate *multi-objective optimization* problems solved by a genetic algorithm

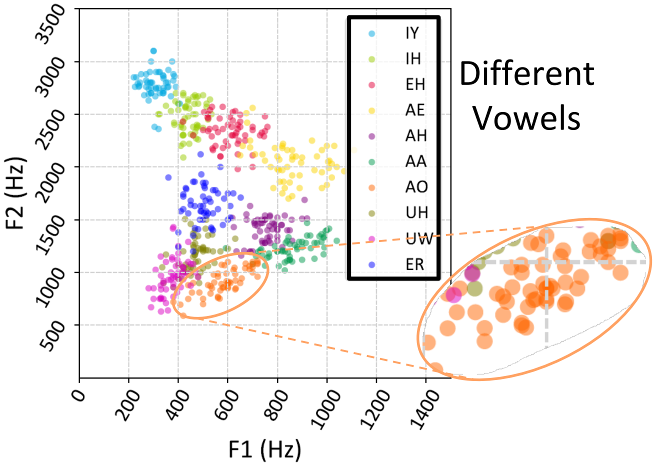
Which feature of voiceprint to modify for anonymization?

Formant

Formant

- Formants are resonant frequencies and map to identities
  1. Formants represent the shape of the vocal tract
  2. The shape of the vocal tract is unique for everyone





Formants ⇨ Voiceprints

Formants distribution differs among people

How to Change Formants?

- **Linear Prediction Coding (LPC)** can model the shape of vocal tract
- Audio codecs use **Line Spectral Frequency (LSF)** as LPC's equivalent representations

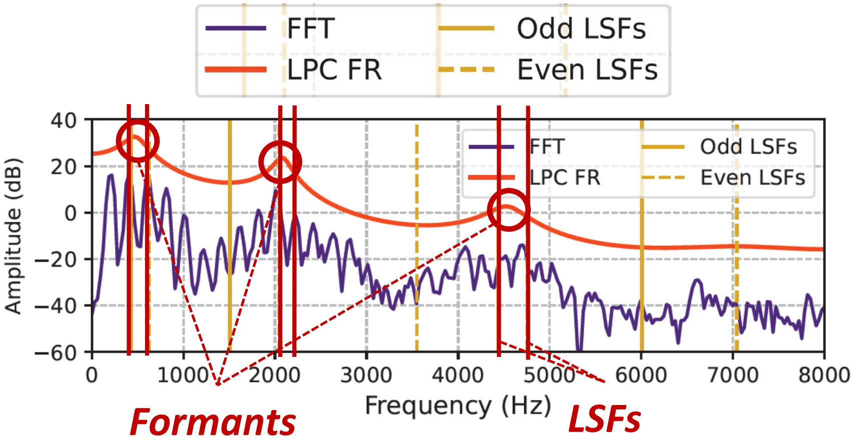
$$\hat{x}(n) = -\sum_{k=1}^{p} \boxed{a_k} x(n-k) + e(n)$$

LPC coefficients

Equivalent Representations

Line spectral frequencies (LSFs)

$$0 < \boxed{\omega_1 < \theta_1 < \cdots < \omega_{p/2} < \theta_{p/2}} < \pi$$



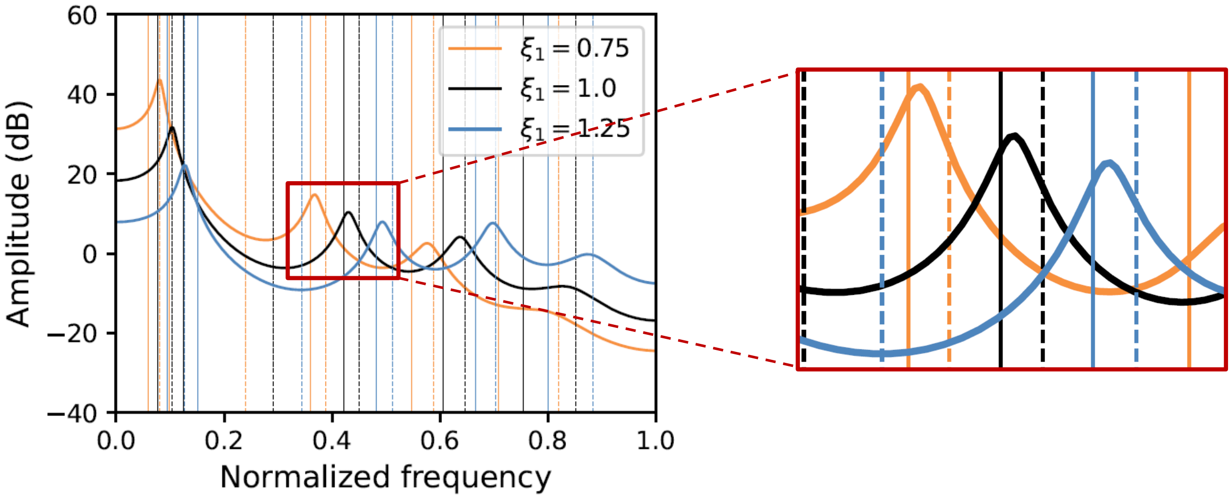
LSFs represent the positions of formants

Change LSFs ⇨ Change the Formants

How to modify LSFs?

Formant Transformations

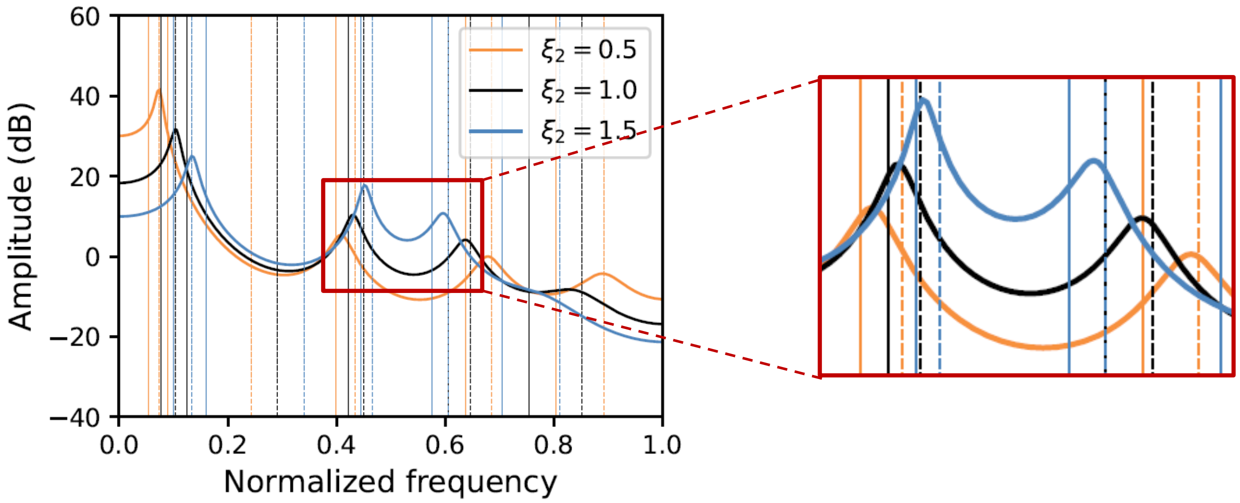
- **Func 1: Shifting formants** $\tilde{\omega}_i = F_1(\omega_i, \xi_1) = \omega_i + \omega_i(\xi_1 - 1)(1 - \omega_i) \quad i = 1, \cdots, p$



***ξ_1>1 (ξ_1<1)* shifts the formants towards higher (lower) frequencies**

Formant Transformations

- **Func 2: Spreading formants** $\tilde{\omega_i} = F_2(\omega_i, \xi_2) = \omega_i + (\xi_2 - 1)\sin(2\pi\omega_i)/p \quad i = 1, \cdots, p$

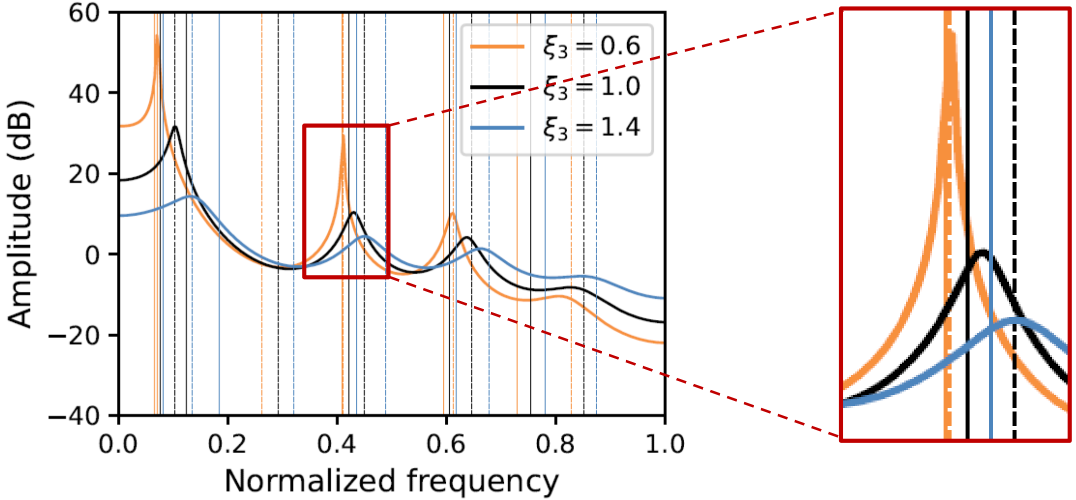


ξ_2>1 (ξ_2<1) means to gather (spread) the formants

Formant Transformations

- **Func 3: Adjusting bandwidths**

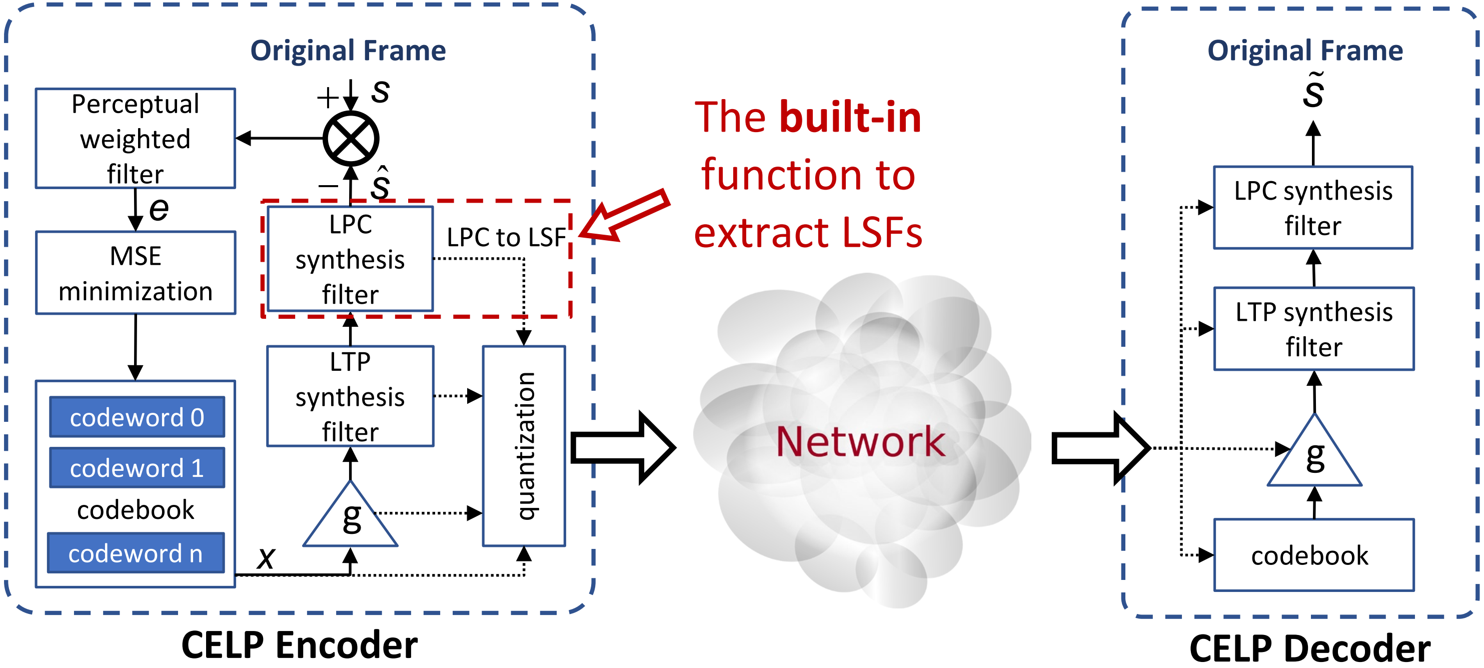
$$\tilde{\omega}_i = F_3(\omega_i, \xi_3) = \sum_{k=0}^{i-1} \{\omega_{k+1} - \omega_k + (\xi_3 - 1)[\frac{1}{p+1} - \omega_{k+1} + \omega_k]\}$$



ξ_3>1 (ξ_3<1) means to expand (shrink) the formants bandwidth

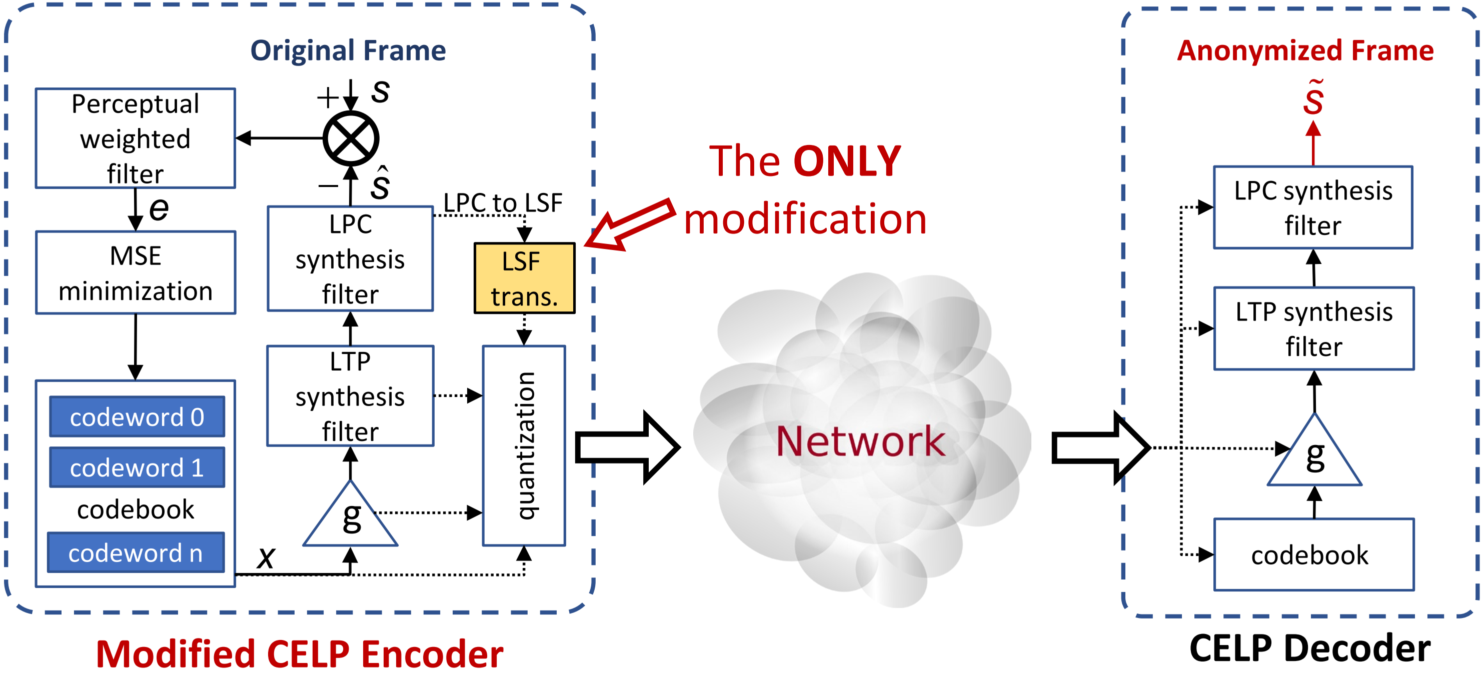
CELP Modification for Formant Transformations

- **CELP**: **C**ode **E**xcitation **L**inear **P**rediction codec (based on LPC)



CELP Modification for Formant Transformations

- **CELP**: **C**ode **E**xcitation **L**inear **P**rediction codec (based on LPC)



How to determine the coefficients of formant transformations?

Objective Function Formulation

- Multi-Objective Function

We anonymize audios and preserve usability for two objectives:

Objective 1: for human

$$\mathbf{T1}: \min_{\xi} \quad S_{\mathrm{ASV}}[v(x), v(\tilde{x})], S_{\mathrm{pept}}(x, \tilde{x})$$
$$\text{s.t.} \quad x, \tilde{x} \in [-1, 1] \quad \text{and} \quad \xi \in [0, 2]$$

Objective 2: for ASRs

$$\mathbf{T2}: \min_{\xi} \quad S_{\mathrm{ASV}}[v(x), v(\tilde{x})], S_{\mathrm{ASR}}(x, \tilde{x})$$
$$\text{s.t.} \quad x, \tilde{x} \in [-1, 1] \quad \text{and} \quad \xi \in [0, 2]$$

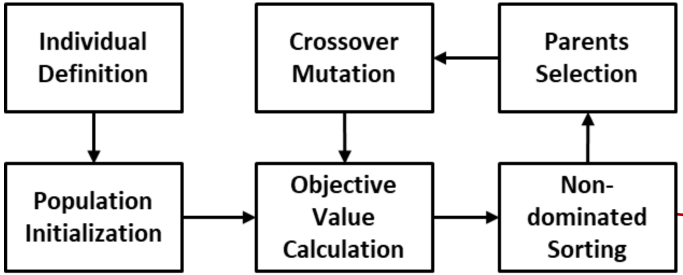
$S_{\mathrm{ASV}}[v(x), v(\tilde{x})]$	*Cosine distance*
$S_{\mathrm{pept}}(x, \tilde{x})$	*Perception score (STOI)*
$S_{\mathrm{ASR}}(x, \tilde{x})$	*Word Error Rate*

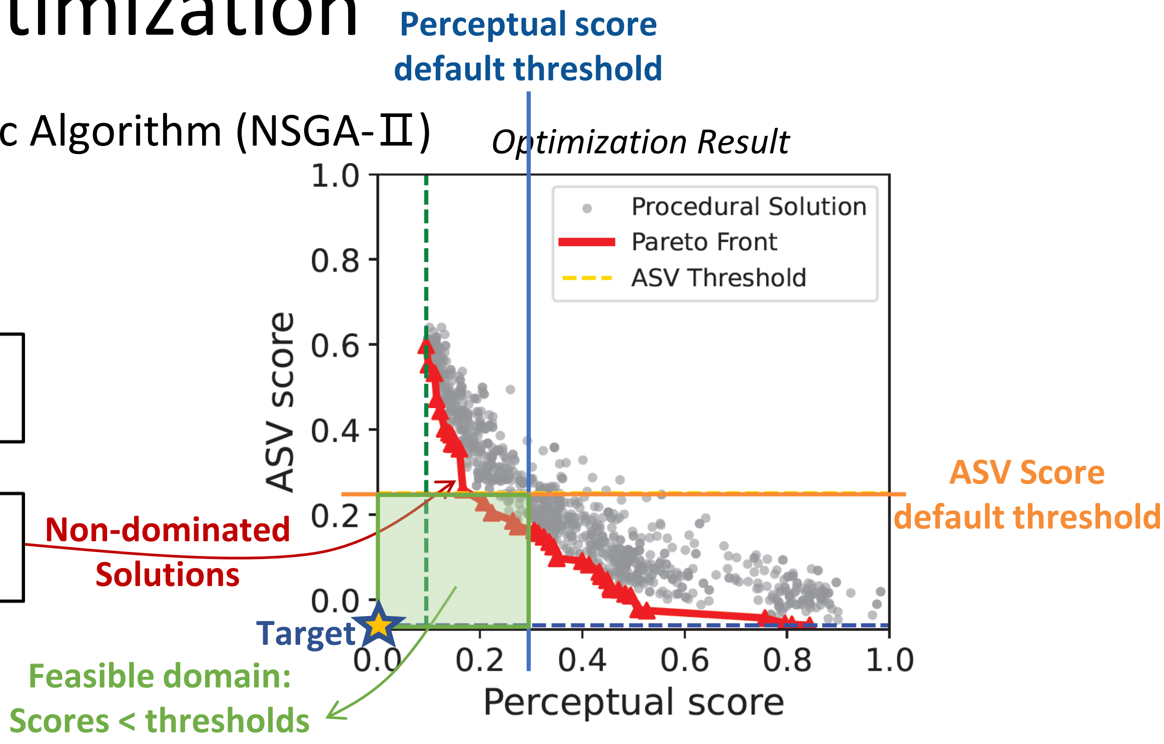
x	*Original signal*
$v(x)$	*Voiceprint embeddings of original signal*
$\tilde{x}$	*Anonymized signal*
$v(\tilde{x})$	*Voiceprint embeddings of anonymized signal*

Multi-objective Optimization

- ☐ Non-dominated Sorting Genetic Algorithm (NSGA-Ⅱ)

Flow chart of NSGA-Ⅱ





Coefficients of feasible solutions are used for anonymization

Evaluation: Setup

Datasets

- **6 datasets (subsets)**
  VoxCeleb1, LibriSpeech, VCTK, AISHELL
- **2272 speakers**
- **262,790 utterances**
- **2 Language**
  English & Chinese

Dataset	Subset	#Speaker	#Utterance	Duration (s)
VoxCeleb1 (E)	*dev*	1,211	148,642	3.9 ~ 144.9
LibriSpeech (E)	*train-clean-360*	921	104,014	1.1 ~ 29.7
VoxCeleb1 (E)	*test*	40	4,874	3.9 ~ 69.1
LibriSpeech (E)	*test-clean*	40	2,260	1.3 ~ 35
VCTK (E)	*wav48*	40*	2,000†	2.1 ~ 15.1
AISHELL (C)	*test*	20	1,000†	1.9 ~ 14.7

ASVs & ASRs

- **3 ASV models, EER < 2.8%**
  ECAPA-TDNN, X-Vector, I-Vector
- **3 ASR models, WER < 3.9%**
  transformer, wav2wec, crdnn-rnn
- **2 Language**
  English & Chinese

ASV Model	Catagory	EER	ASR Model	Language	WER
ECAPA-TDNN	DNN-based	0.7%	transformer	E&C	2.27%
X-Vector	DNN-based	2.5%	wav2vec2	E	1.90%
I-Vector	Statistic	2.8%	crdnn-rnn	E	3.90%

Evaluation: Setup

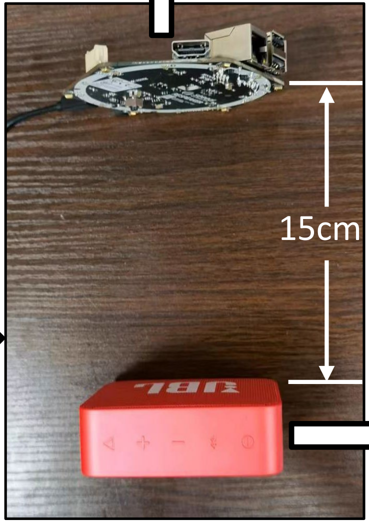
☐ Physical Setup

Microphone module: Respeaker Core V2

- RK3229 MCU with Linux system

MicPro Microphone:
Records and **anonymizes** audio

Laptop:
Illustrates the result



Represents a human speaker

Evaluation: Setup

Baselines, two existing anonymization methods based on signal processing

1. McAdam Transformation (MT) [1]
2. VoiceMask (VM) [2]

Evaluation Metrics

Anonymity

1. **Miss-Match Rate (MMR):** the rate anonymized audio mismatched with the correct speaker;
2. **Equal Error Rate (EER):** the rate when False Accept Rate = False Rejection Rate;

Usability

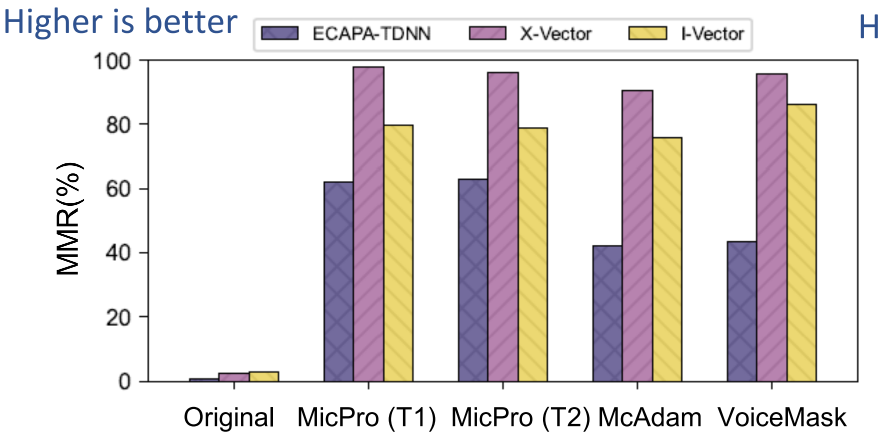
3. **Latency:** the delay of the codec;
4. **Short-Time Objective Intelligibility (STOI).** STOI indicates speech intelligibility;
5. **Subjective quality:** clearness, naturalness, similarity, and acceptability;
6. **Word Error Rate (WER):** the dissimilarity of ASR results between original and anonymized audio

[1] Jose Patino, Natalia Tomashenko, Massimiliano Todisco, et.al. Speaker Anonymisation Using the McAdams Coefficient. In Interspeech 2021.
[2] Jianwei Qian, Haohua Du, Jiahui Hou, et.al. 2017. Voicemask: Anonymize and sanitize voice input on mobile devices. arXiv preprint.

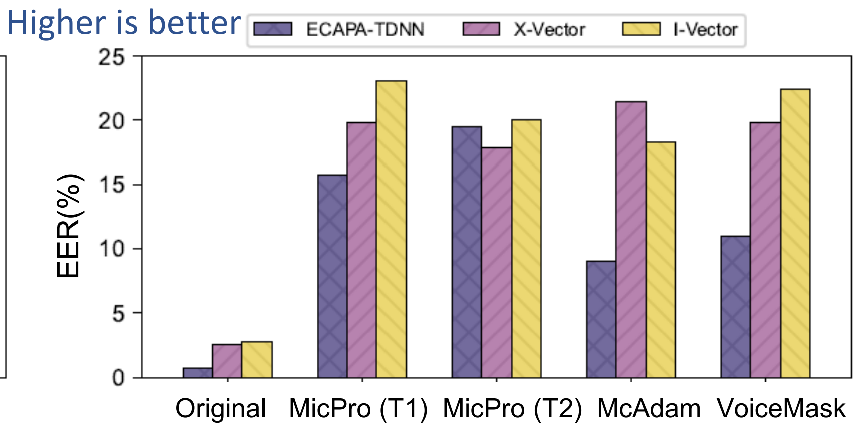
Evaluation: Result

☐ Anonymity performance

Mis-Match Rate



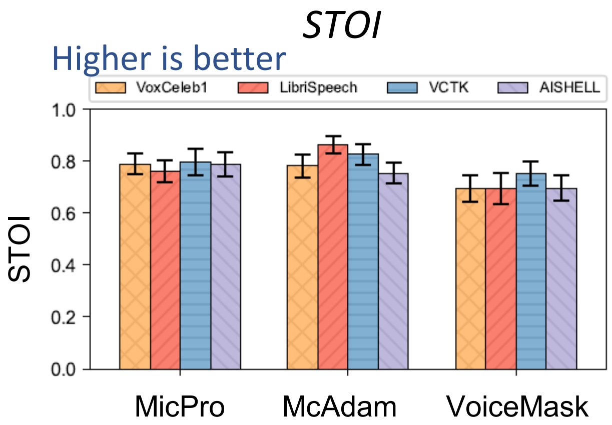
Equal Error Rate

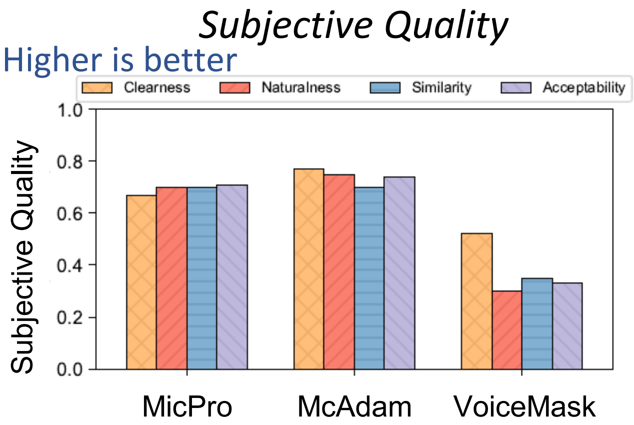


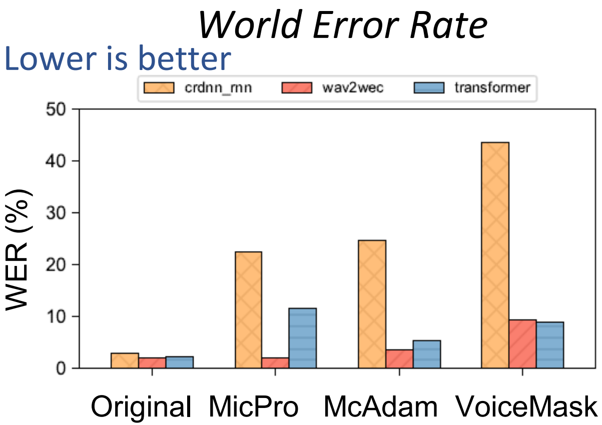
MicPro anonymity outperforms baseline methods in SOTA ASV

Evaluation: Result

☐ Usability performance







MicPro usability outperforms VM and is comparable with MT

Evaluation: Result

☐ Usability performance

Latency increase after modifying the CELP codec

t_{dur} (s)	t_{enc} (ms)	$\tilde{t}_{enc}$ (ms)	l (ms)	$\tilde{l}$ (ms)	Δl (ms)	$\delta l(\%)$
5	683 ± 18	685 ± 10	16.366	16.370	0.004	0.02
30	$3,864 \pm 22$	$3,868 \pm 24$	16.288	16.289	0.001	0.01
120	$15,289 \pm 45$	$15,293 \pm 32$	16.274	16.274	0.000	0.00
Avg.	-	-	16.309	16.311	0.002	**0.01**

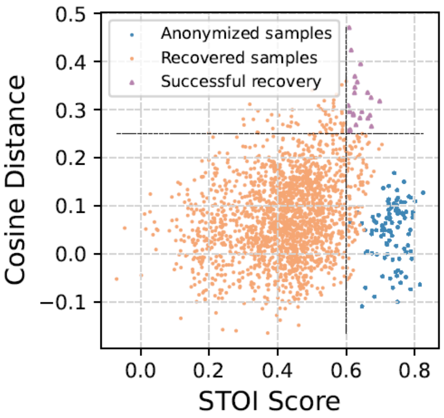
MicPro has latency lower than 17ms

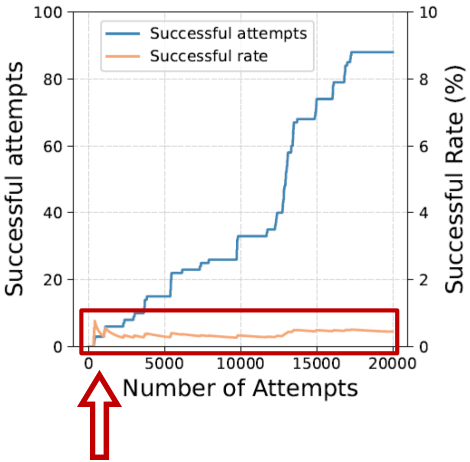
The latency increase is only 0.01%

Evaluation: Result

□ Resistance to attacks

Successful rate of spoofing attack

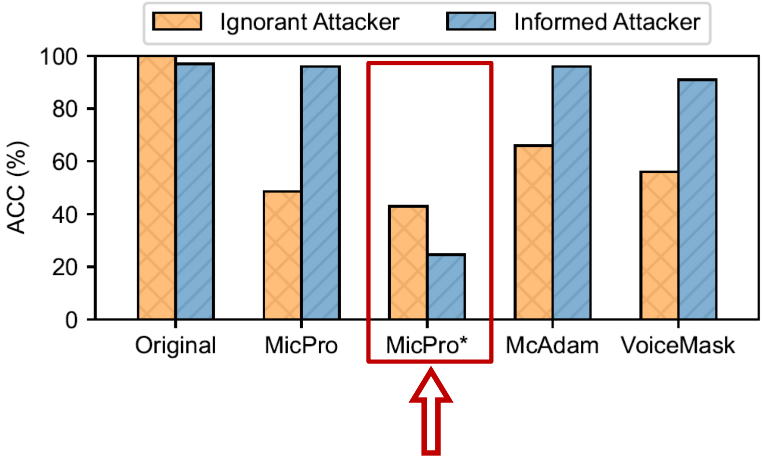




The attack successful rate is only 0.44%

Accuracy of inference attack

Lower is better



MicPro performs best against inference attack

Conclusion

1. The first **privacy-by-design microphone modules** which can produce anonymous recordings

2. We design **formant transformations** within a CELP codec and formulate optimization problems to determine the coefficients

3. We implement MicPro on an off-the-shelf microphone, **validate the performance and resistance to attacks**

MicPro: Microphone-based Voice Privacy Protection

Find our demo and code at:

https://github.com/USSLab/MicPro

Contact the authors at:

xshilin@zju.edu.cn

xji@zju.edu.cn

yanchen@zju.edu.cn

zheng_zhicong@zju.edu.cn

wyxu@zju.edu.cn

USSLAB

Homepage: www.usslab.org