V-CLOAK: Intelligibility-, Naturalness- & Timbre-Preserving Real-Time Voice Anonymization

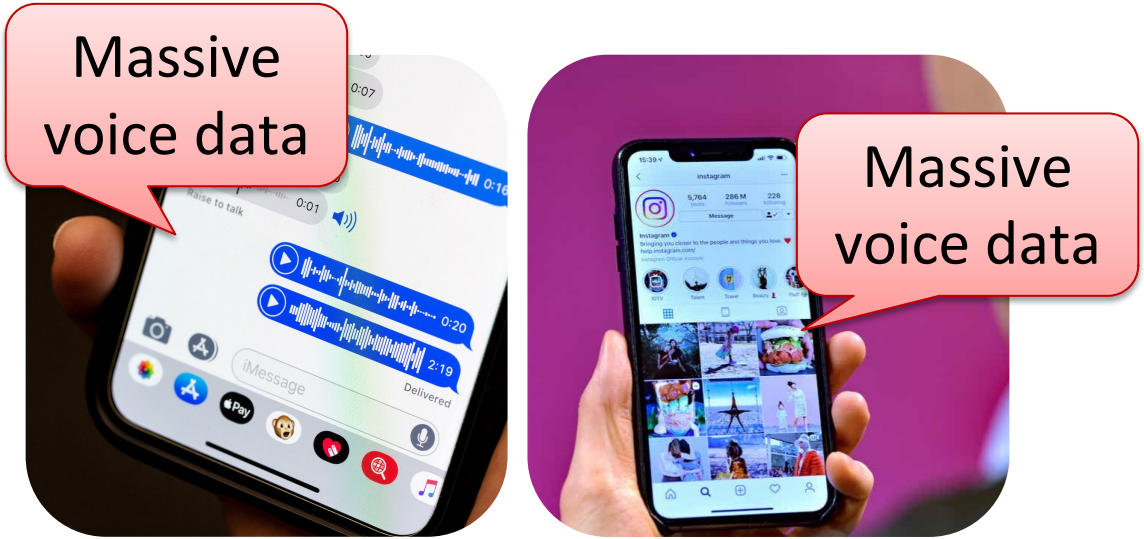
Jiangyi Deng[1], Fei Teng[1], Yanjiao Chen*[1], Xiaofu Chen[2], Zhaohui Wang[2], Wenyuan Xu[1]

[1]Ubiquitous System Security Lab (USSLAB), Zhejiang University

[2]Wuhan University

{jydeng, chenyanjiao, wyxu}@zju.edu.cn

Your Voice Data May be Abused!



Voice Message

Social Media

Your Voice Data May be Abused!

Voice Message

Social Media

Adversary **Identity leakage**

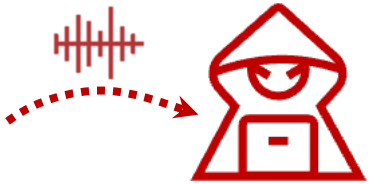
Privacy inference: *user tracking …*

Identity theft: *fake voice synthesis …*

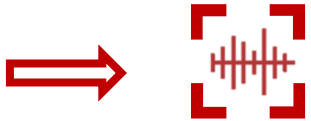
Your Voice Data May be Abused!

Voice Message

Social Media



Adversary



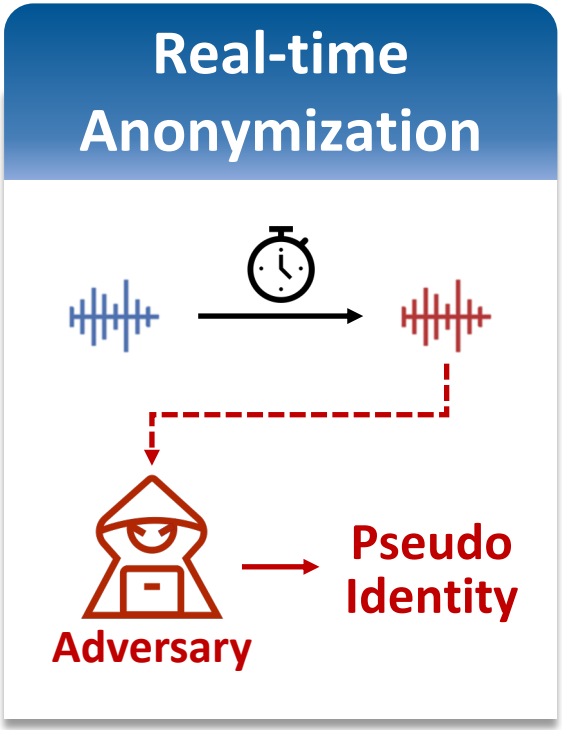
Identity leakage

Privacy inference: *user tracking …*

Identity theft: *fake voice synthesis …*

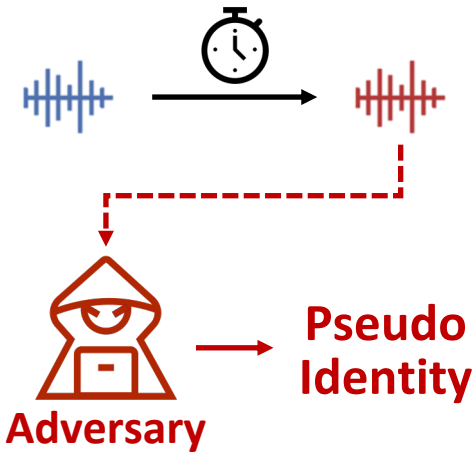
How can we protect our voice in social media scenarios?

Protection Method Requirements



Protection Method Requirements

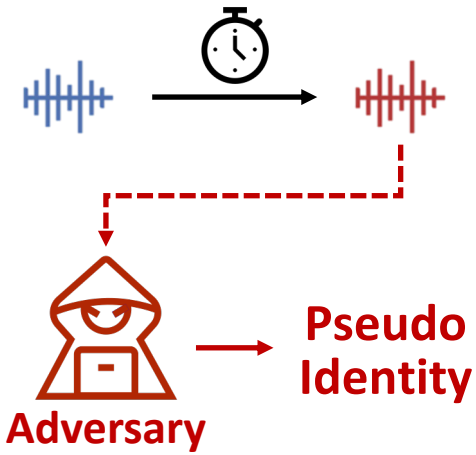
Real-time Anonymization



Timbre-preserving & Naturalness

Protection Method Requirements

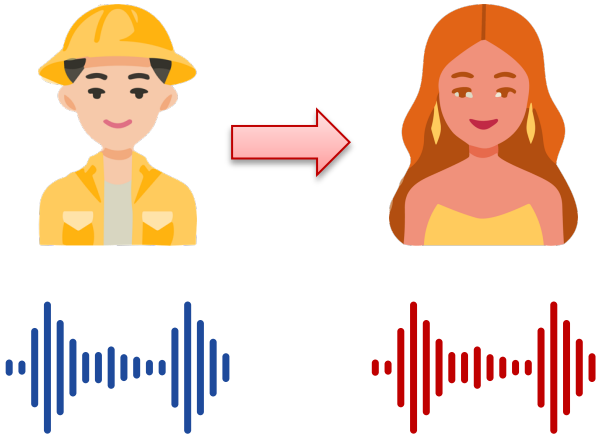
Real-time Anonymization



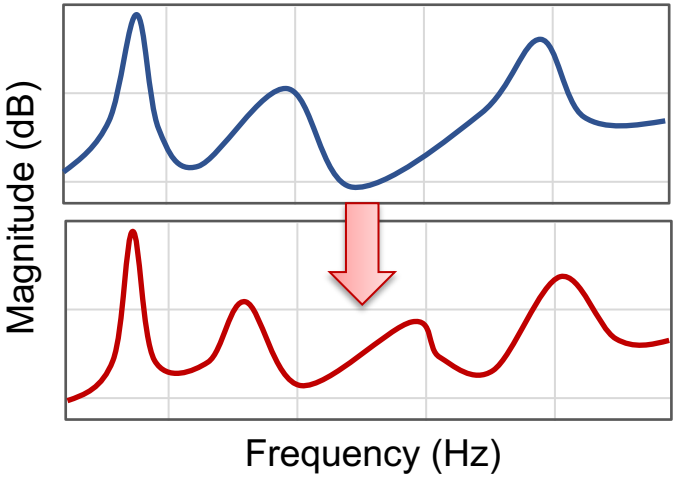
Timbre-preserving & Naturalness

Transcription Perfomance

CONFLICT: Anonymity VS Timbre-preservation

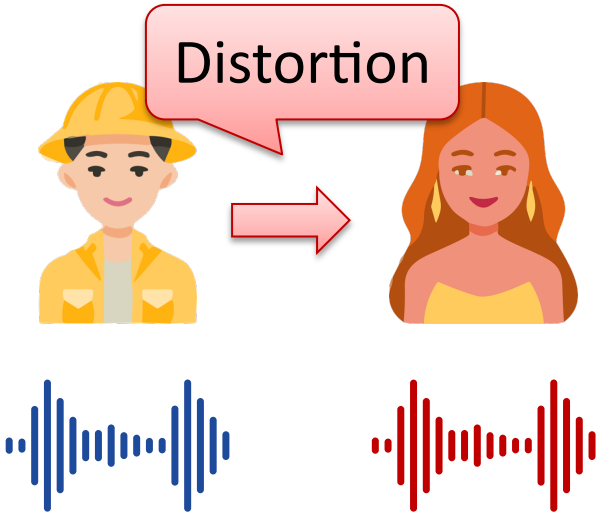


Voice Conversion/Synthesis

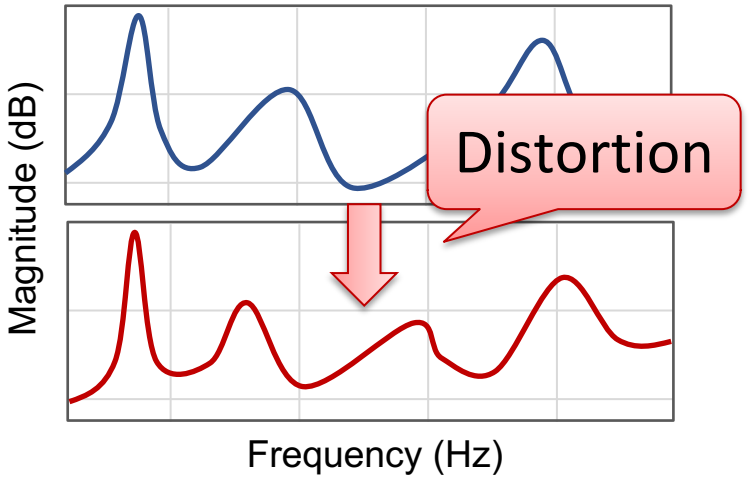


Signal Processing

CONFLICT: Anonymity VS Timbre-preservation

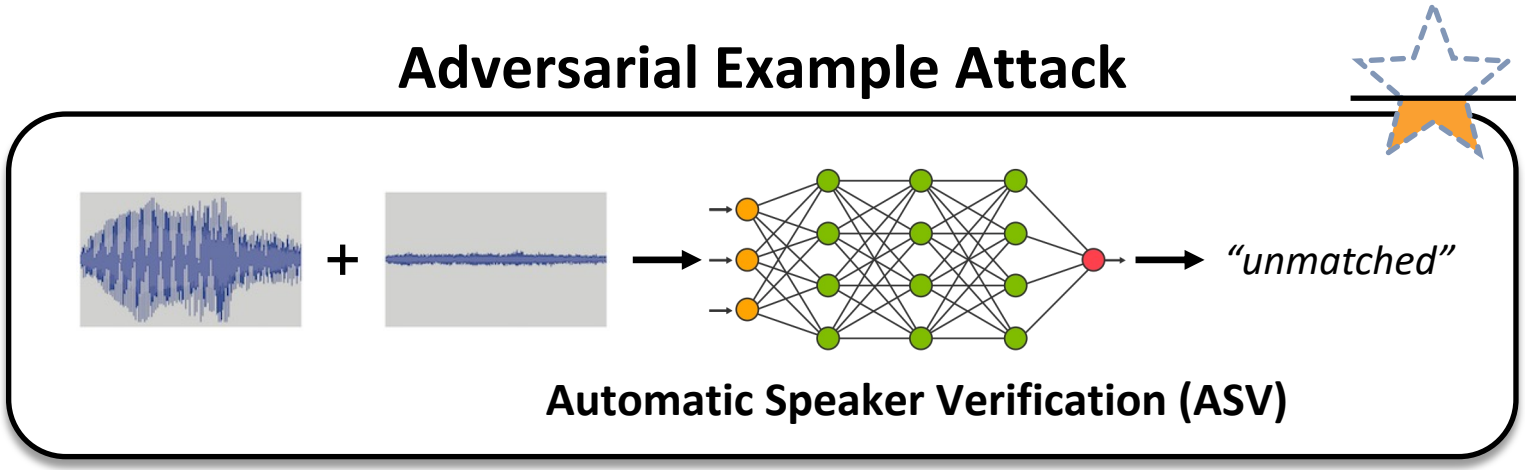


Voice Conversion/Synthesis



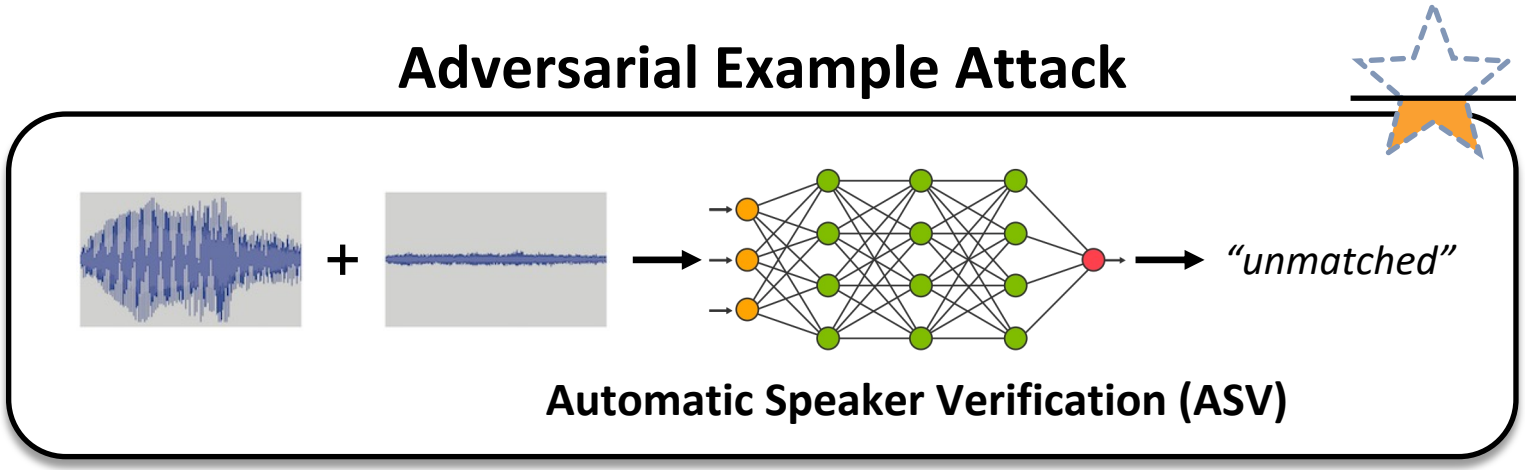
Signal Processing

Anonymity && Timbre-preservation



Preserve the original timbre and achieve anonymity at the same time.

Anonymity && Timbre-preservation



Adapt adversarial examples to achieve the three requirements.

GOAL 1: Real-time Anonymization

Original audio

One-shot generation
(***real-time***)

Anonymized audio

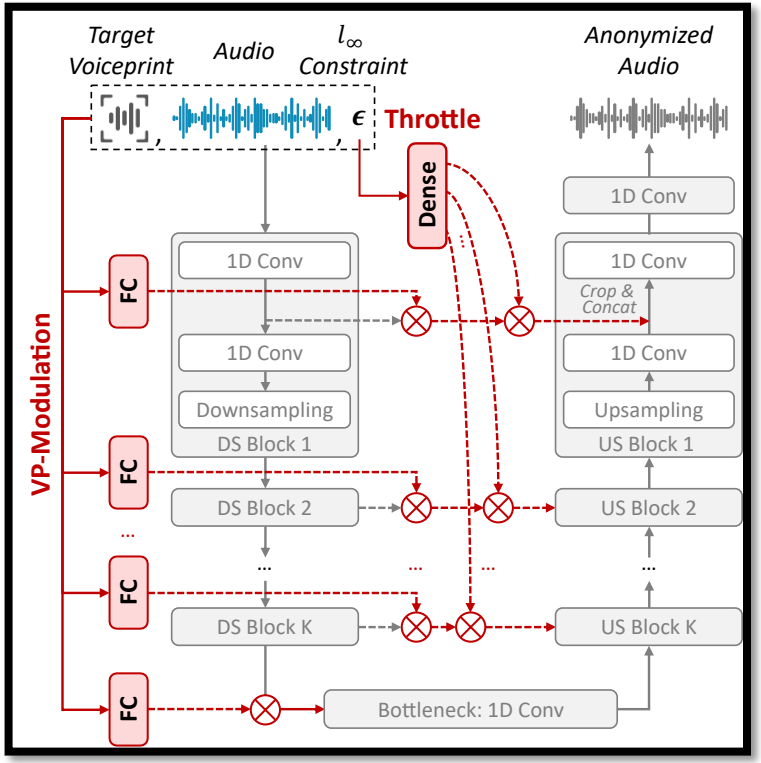
Iterative optimizaiton ✗

GOAL 1: Real-time Anonymization

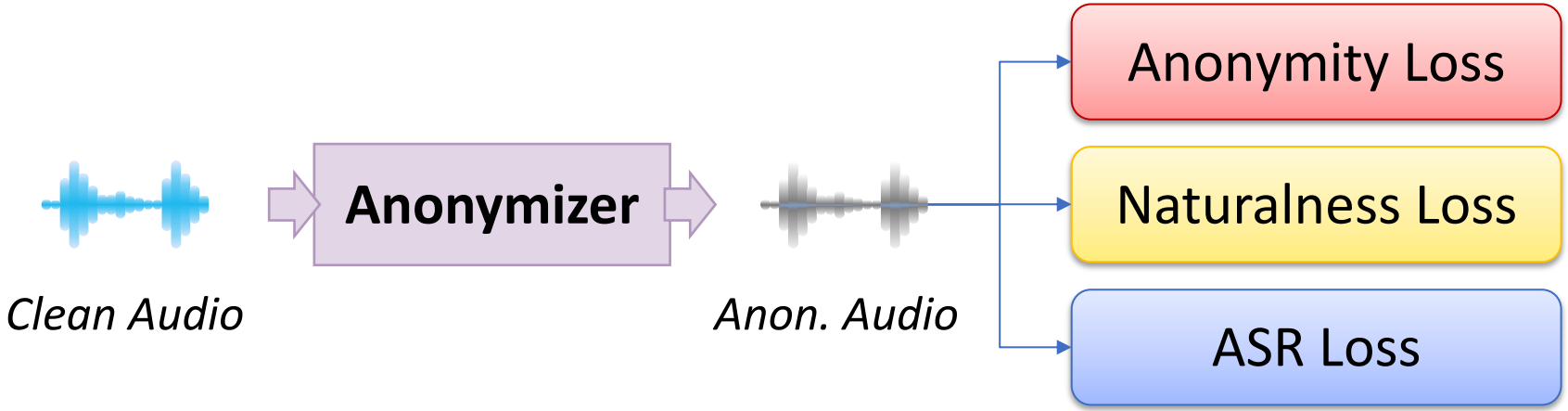
Conv-based audio-to-audio

Fast & effective

- **RNN ✗**
- **Vocoder ✗**



Training Anonymizer



Three loss terms to achieve three requirements

GOAL 1: Real-time Anonymization

Anonymity

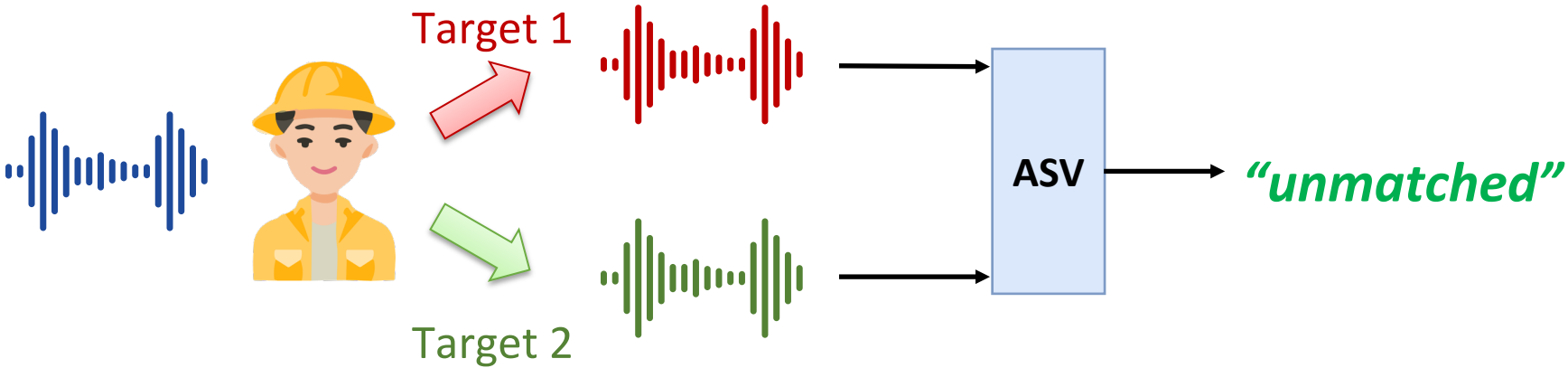
Clean audio ⇔ Anon. audio

ASV ⇨ *“unmatched”*

Anon. audio 1 ⇔ Anon. audio 2

ASV ⇨ *“unmatched”*

GOAL 1: Real-time Anonymization



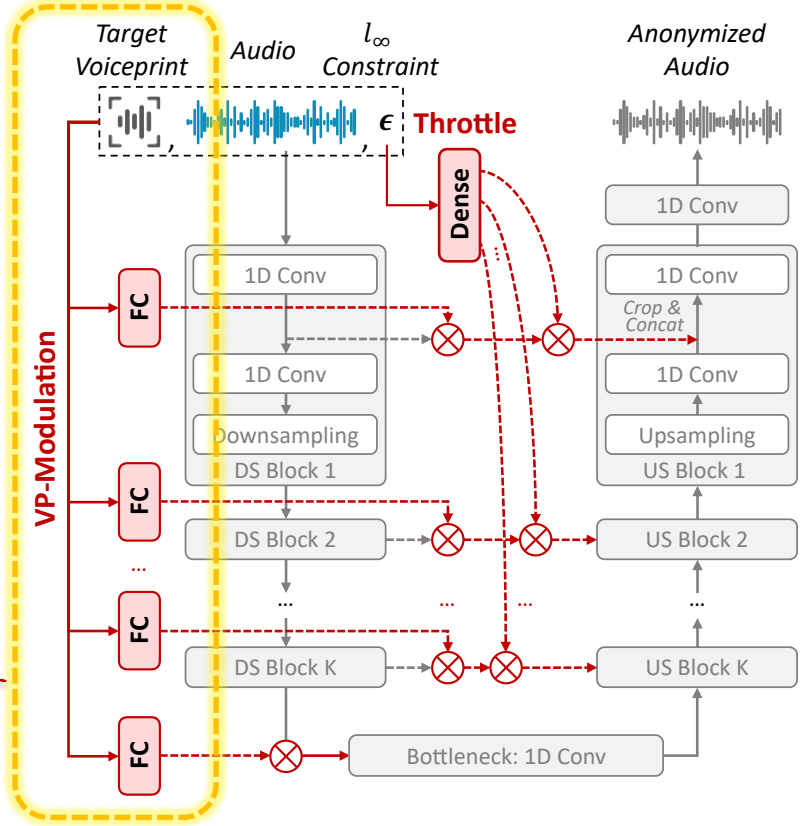
*Perform **targeted** anonymization to achieve anonymity in both cases.*

GOAL 1: Real-time Anonymization

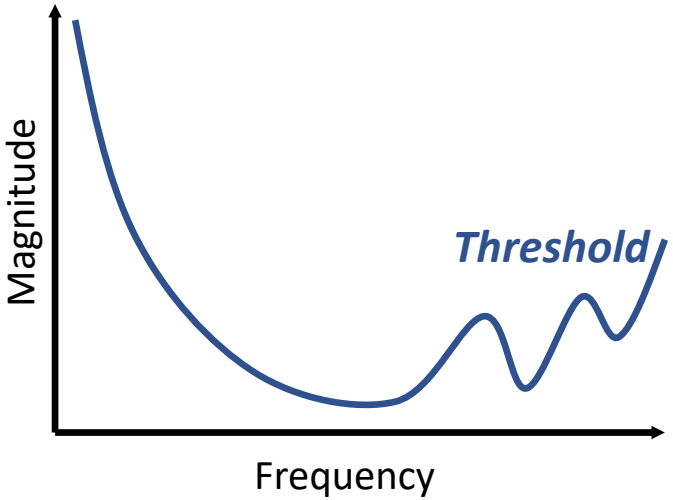
VP-Modulation

To modulate the feature of the original audio at each frequence level **according to the target voiceprint.**

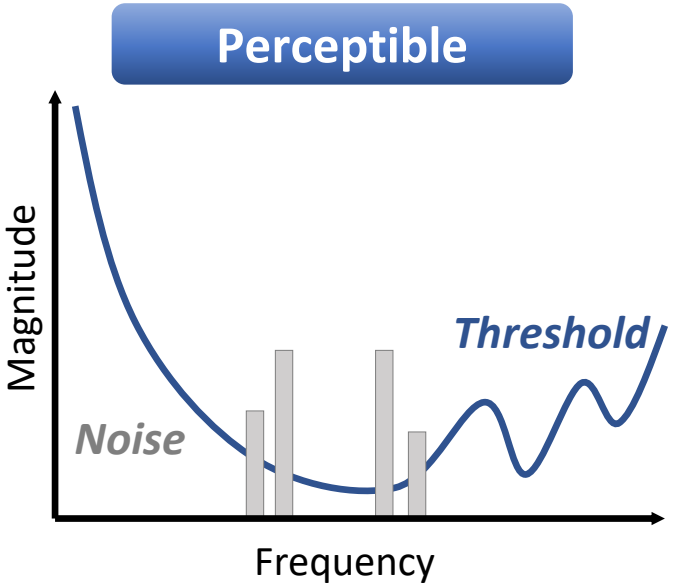
Injecting the target voiceprint



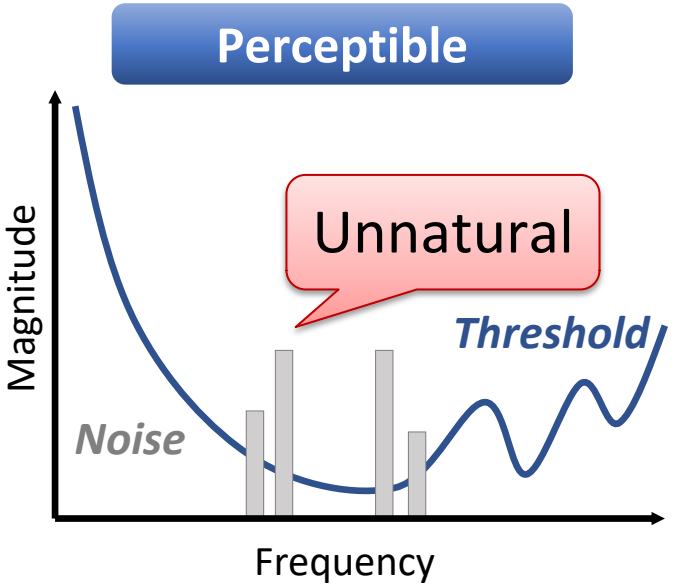
GOAL 2: Naturalness



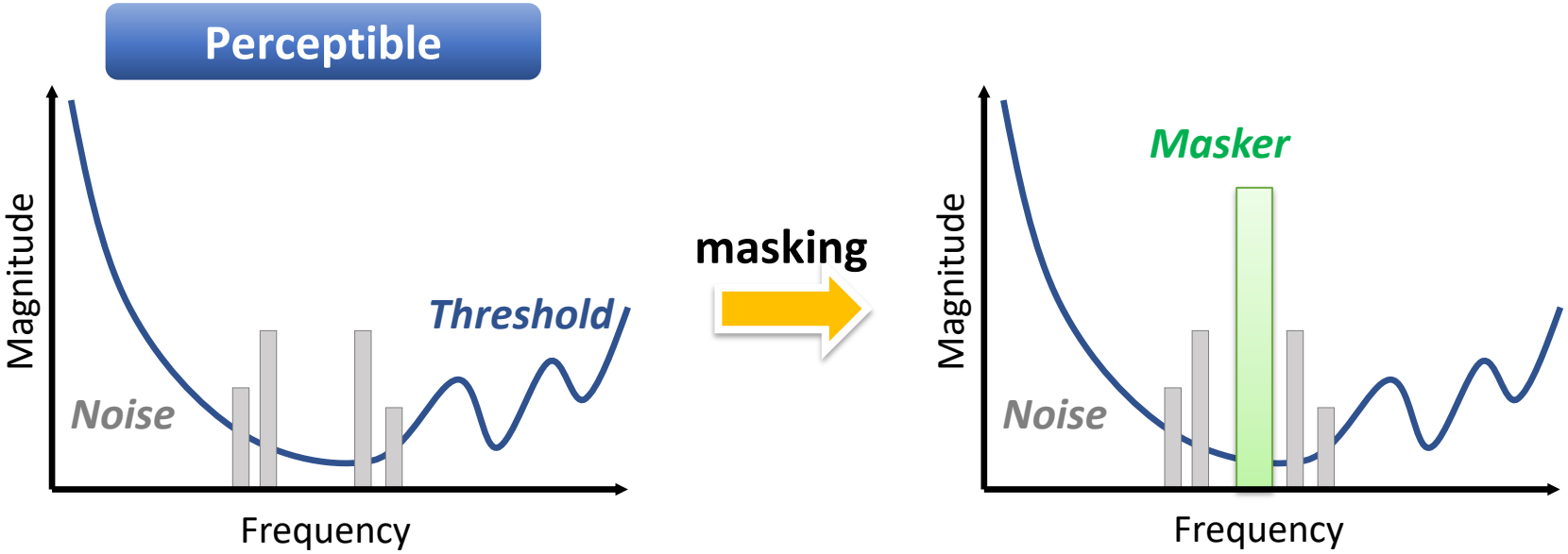
GOAL 2: Naturalness



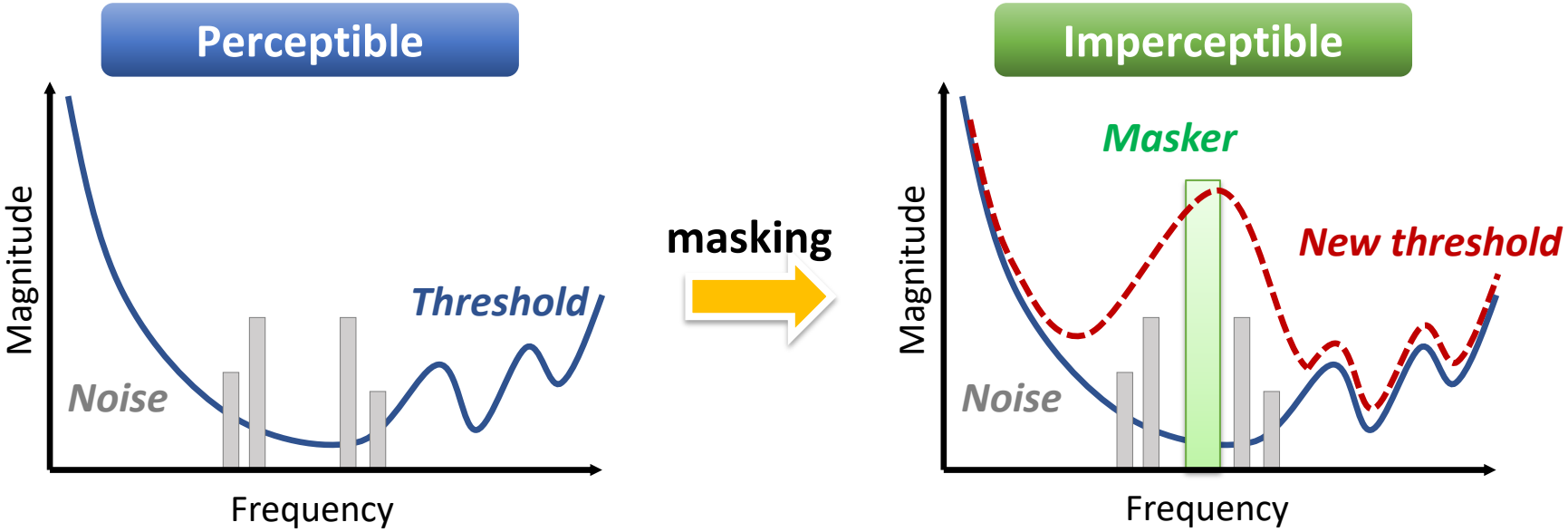
GOAL 2: Naturalness



GOAL 2: Naturalness

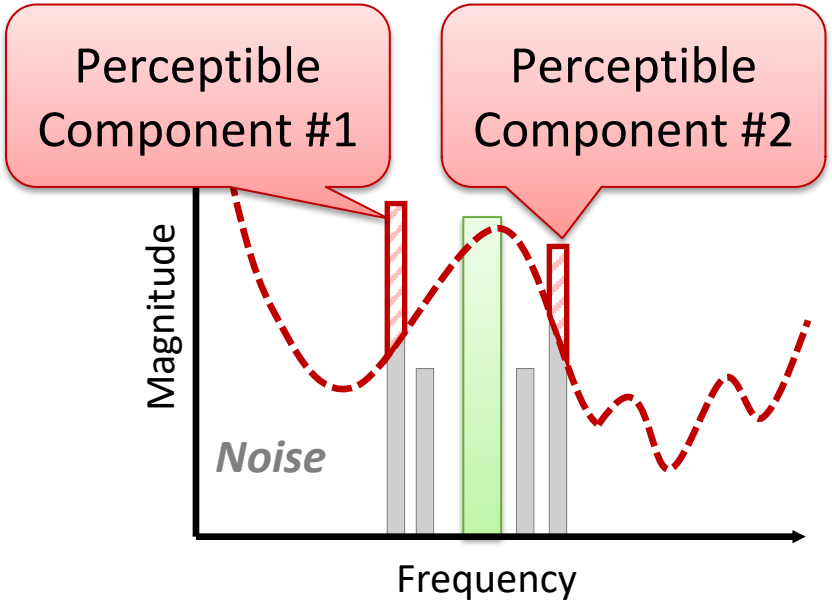


GOAL 2: Naturalness



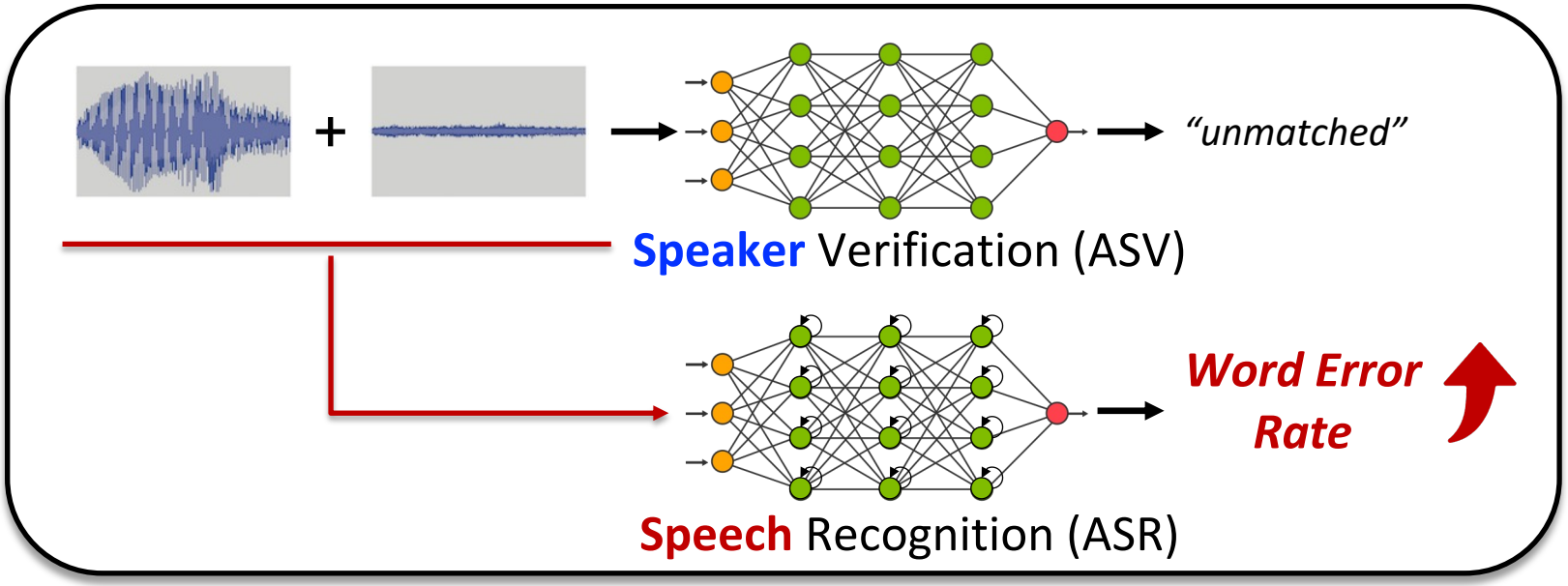
Psychoacoustic Model (Frequency Masking Effect)

GOAL 2: Naturalness



$$L_{\mathrm{PSY}}(\tilde{x}, x) = \frac{1}{F}\sum \max\{0, \mathrm{PSD}(\tilde{x} - x) - M(x)\}$$

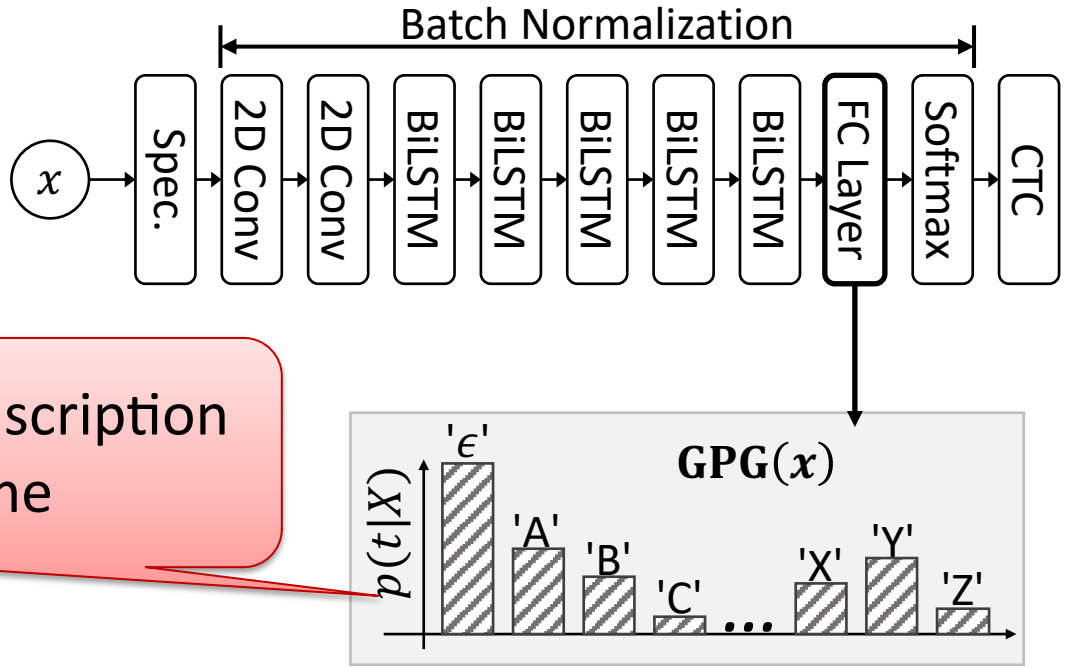
GOAL 3: Transcription Performance



Perturbation that misleads ASV also misleads ASR.

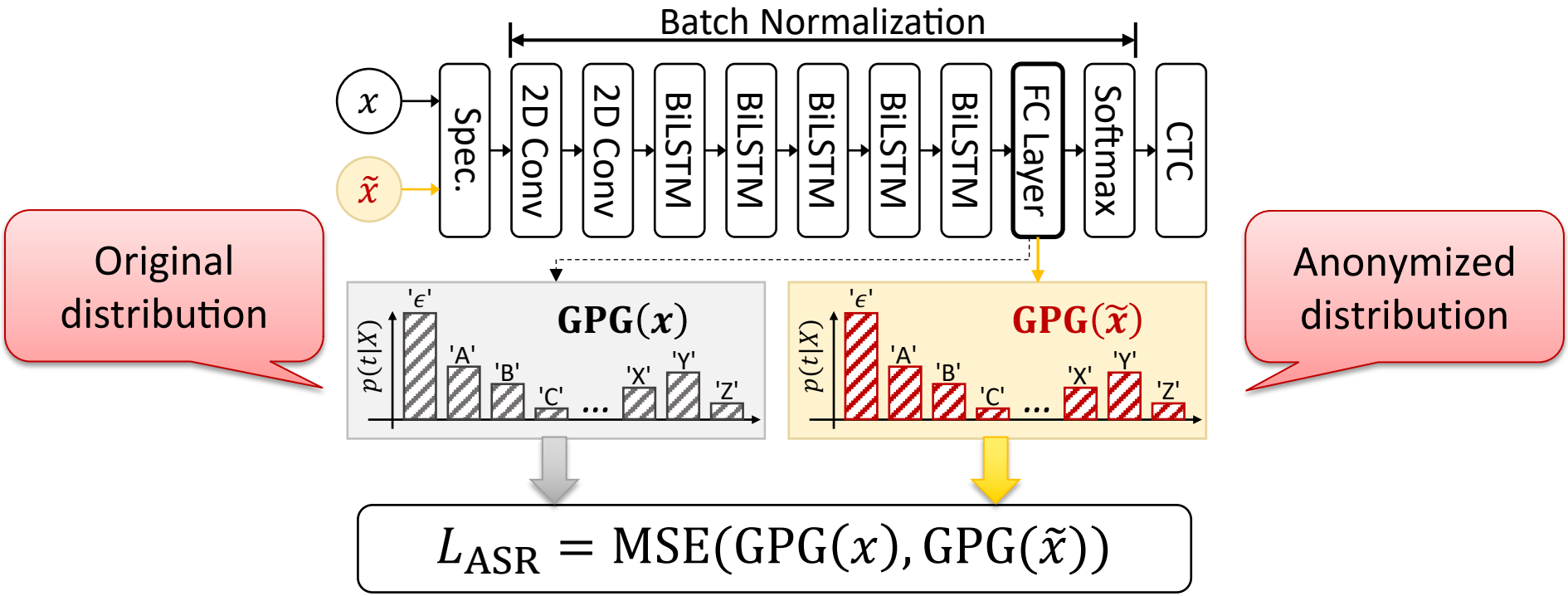
GOAL 3: Transcription Performance

DeepSpeech2



Represent the transcription of each frame

GOAL 3: Transcription Performance



Compared with Existing Work

- **Voice conversion-based**
  - VoiceMask
- **Voice synthesis-based**
  - NSF
  - HiFi-GAN
- **Signal processing-based**
  - McAdams

Compared with Existing Work

Miss-Match Rate ⬆️ *Equal Error Rate* ⬆️ *Word Error Rate* ⬇️

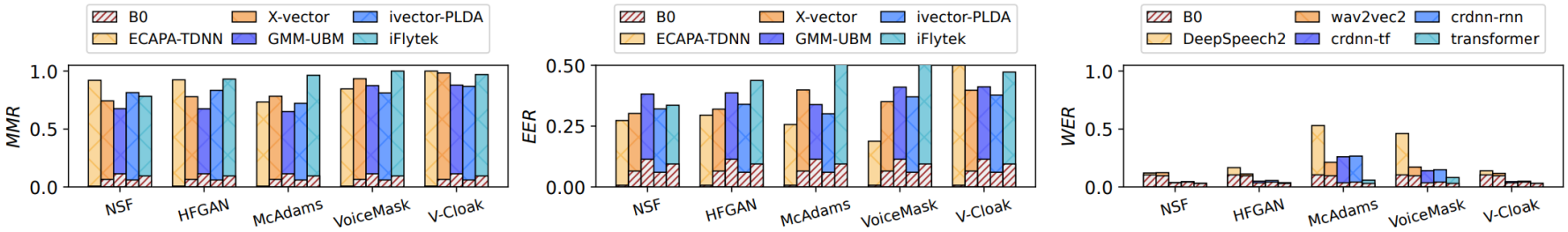


Figure 6: Comparison with existing works. (a) MMR. (b) EER. (c) WER. V-CLOAK yields the highest average MMR of 94.02% and the highest average EER of 46.10%. V-CLOAK obtains a low average WER of 7.65% second only to the NSF (7.19%).

Compared with Existing Work

Miss-Match Rate ⬆ *Equal Error Rate* ⬆ Word Error Rate ⬇

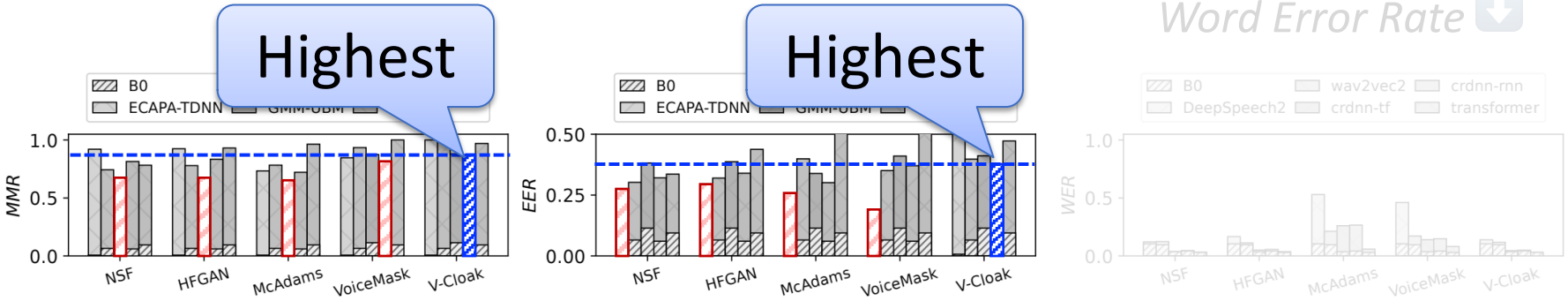


Figure 6: Comparison with existing works. (a) MMR. (b) EER. (c) WER. V-CLOAK yields the highest average MMR of 94.02% and the highest average EER of 46.10%. V-CLOAK obtains a low average WER of 7.65% second only to the NSF (7.19%).

Considering the worst case, V-CLOAK ranks ***the first*** *in terms of* ***anonymity***.

Compared with Existing Work

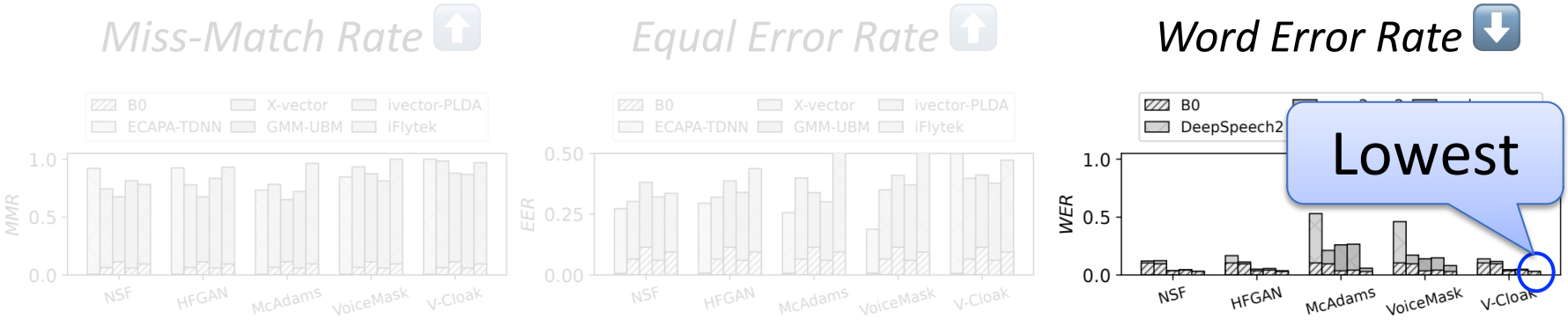


Figure 6: Comparison with existing works. (a) MMR. (b) EER. (c) WER. V-CLOAK yields the highest average MMR of 94.02% and the highest average EER of 46.10%. V-CLOAK obtains a low average WER of 7.65% second only to the NSF (7.19%).

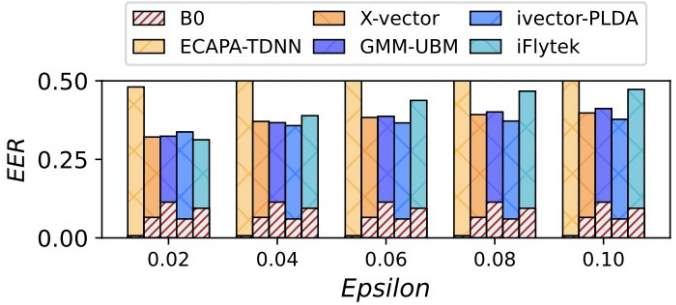
Considering the best case, V-CLOAK ranks ***the first*** *in terms of* ***transcription accuracy***.

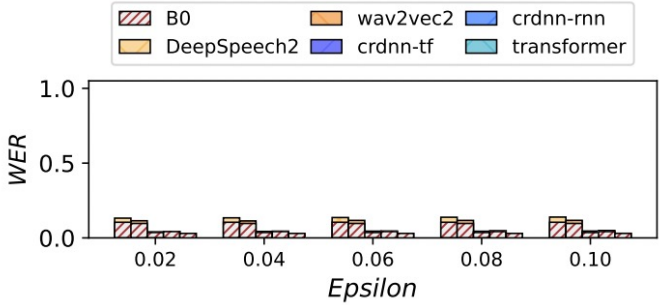
Cross-Language Performance

Equal Error Rate ⬆️

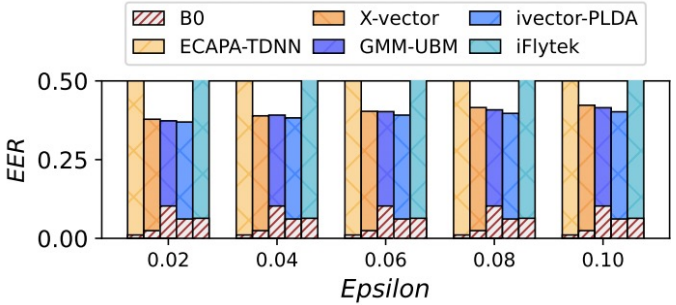
Word Error Rate ⬇️

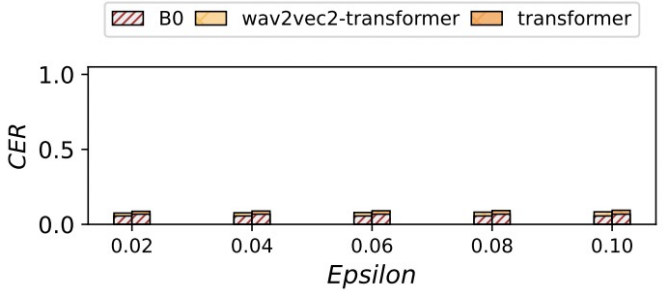
Train on English
*Test on **English***





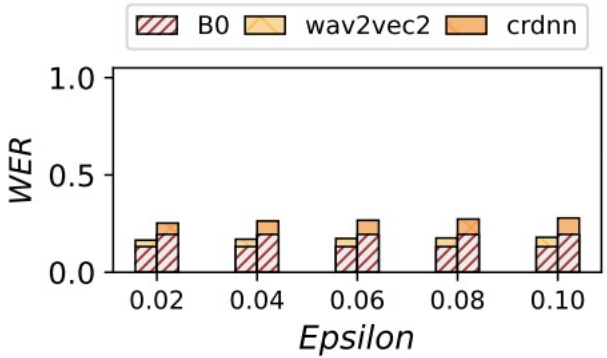
Train on English
*Test on **Chinese***



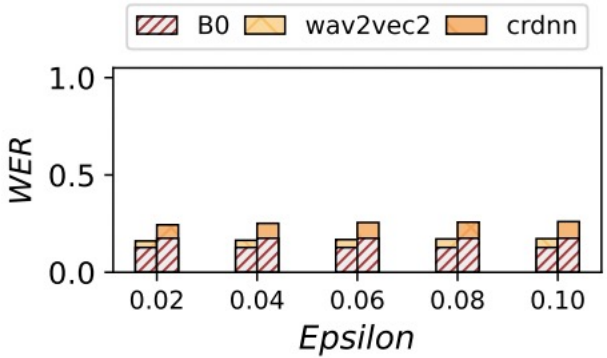


Cross-Language Performance

Word Error Rate ⬇️



Train on English
*Test on **French***



Train on English
*Test on **Italian***

Robustness

Miss-Match Rate ⬆️

Equal Error Rate ⬆️

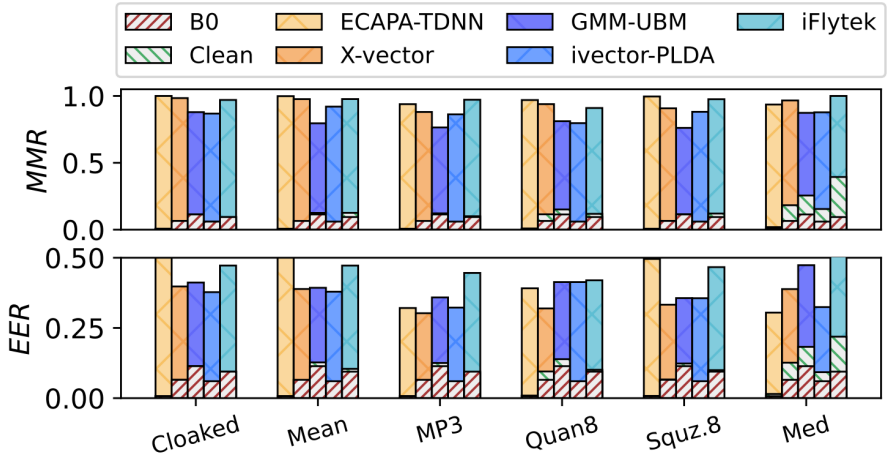


Figure 7: Unidentifiability under adversary A2. The most effective de-noising method, MP3 compression, only causes a decrease in the MMR of 3.27% and the EER of 9.26%.

*Considering the worst case, V-CLOAK ranks **the first** in terms of anonymity under all **de-noising techniques**.*

Anon. Audio 1 ⇔ Anon. Audio 2

Table 4: The performance under adaptive attacker A3.

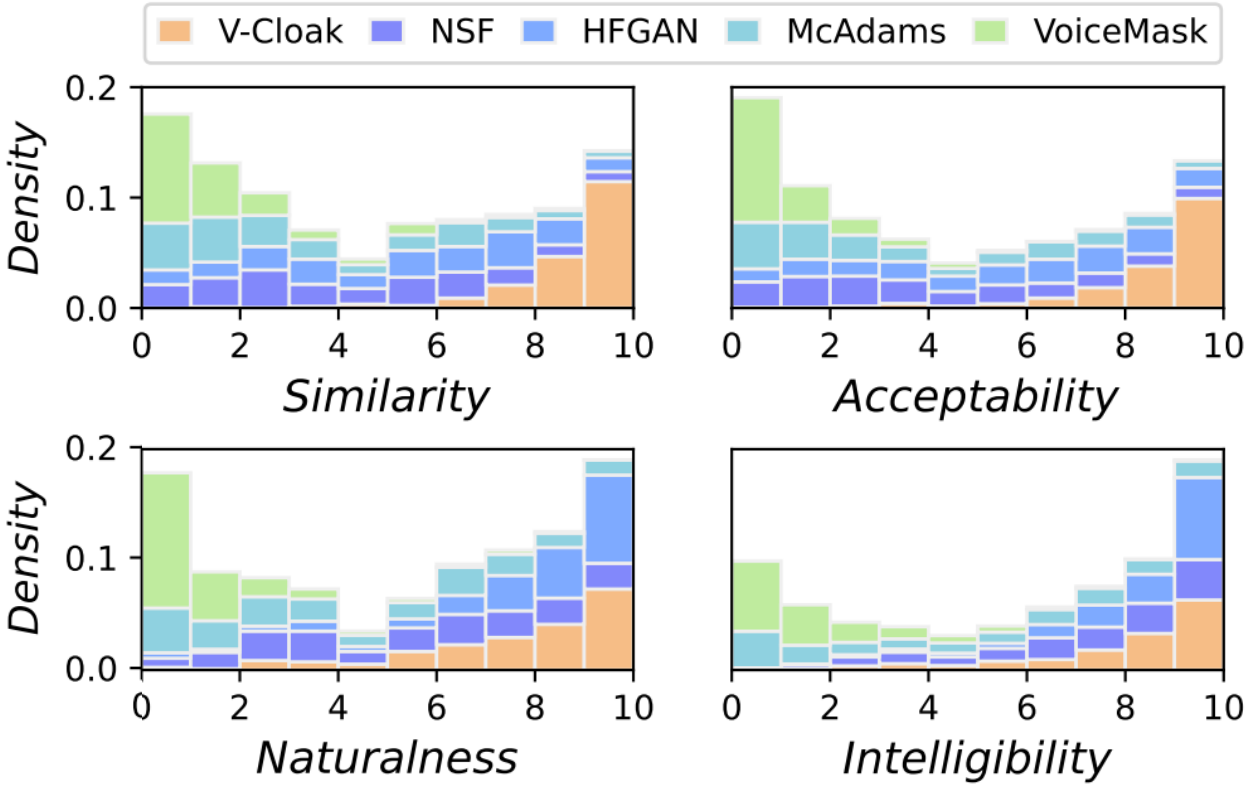
Model		B0 (%)	NSF (%)	HFGAN (%)	McAdams (%)	VoiceMask (%)	V-CLOAK (%)		
							Untargeted	Targeted w/o key†	Targeted w key†
ASV	EP	0.70	24.50	24.79	9.01	8.80	22.03	46.93	32.98
	XV	6.53	27.56	27.11	9.13	14.75	18.20	44.50	29.67
	GMM	11.39	30.36	31.40	23.33	32.58	39.50	43.44	51.15
	IV	6.03	11.17	26.51	8.76	18.25	32.90	40.42	37.22
	IF	9.44	28.24	27.85	13.31	18.95	19.93	37.56	31.34
	AVG	6.82	24.37	27.53	12.71	18.67	26.51	42.57	36.47
	WCS	-	11.17	24.79	8.76	8.80	18.20	37.56	29.67

(i) †: w/ or w/o key means that the voiceprint of the target speaker is known or unknown to the adversary. (ii) **AV**: the average-case scenario, **WCS**: the worst-case scenario. **EP**: ECAPA-TDNN, **XV**: X-vector, **GMM**: GMM-UBM, **IV**: ivector-PLD, iFlytek.

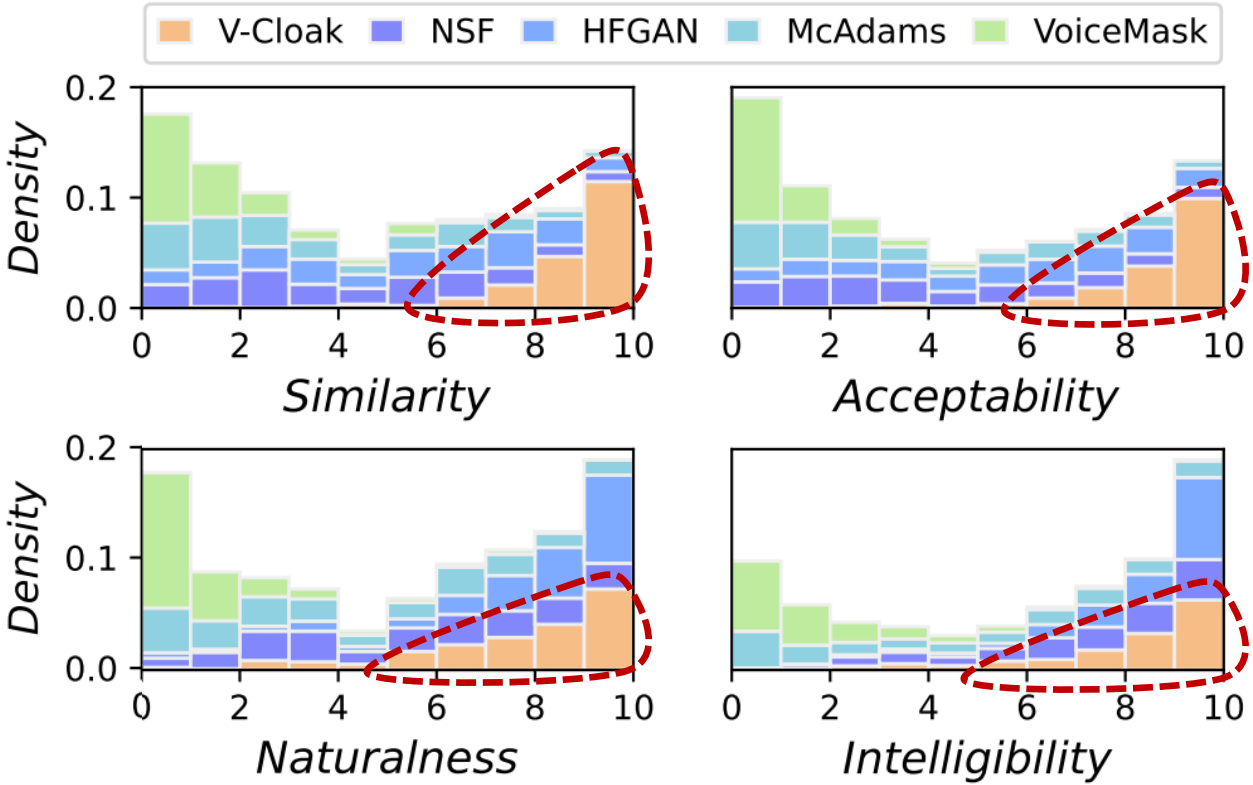
Highest

*V-CLOAK ranks **the first** in terms of unmatching rate between anonymized audios.*

Subjective Results (>100 participants)



Subjective Results (>100 participants)



Demo

Original

Anonymized

For more demo and source code: https://www.v-cloak.com

Conclusion

1. V-CLOAK: an intelligibility-, naturalness and timbre-preserving voice anonymization system.

2. Validated on 5 ASVs, 11 ASRs, 4 languages, 6 defenses, as well as a user study compared with 4 SOTA baselines.

V-CLOAK: Intelligibility-, Naturalness- & Timbre-Preserving Real-Time Voice Anonymization

Paper and demo website: https://www.v-cloak.com

Contact us:
chenyanjiao@zju.edu.cn

USSLAB website:
www.usslab.org