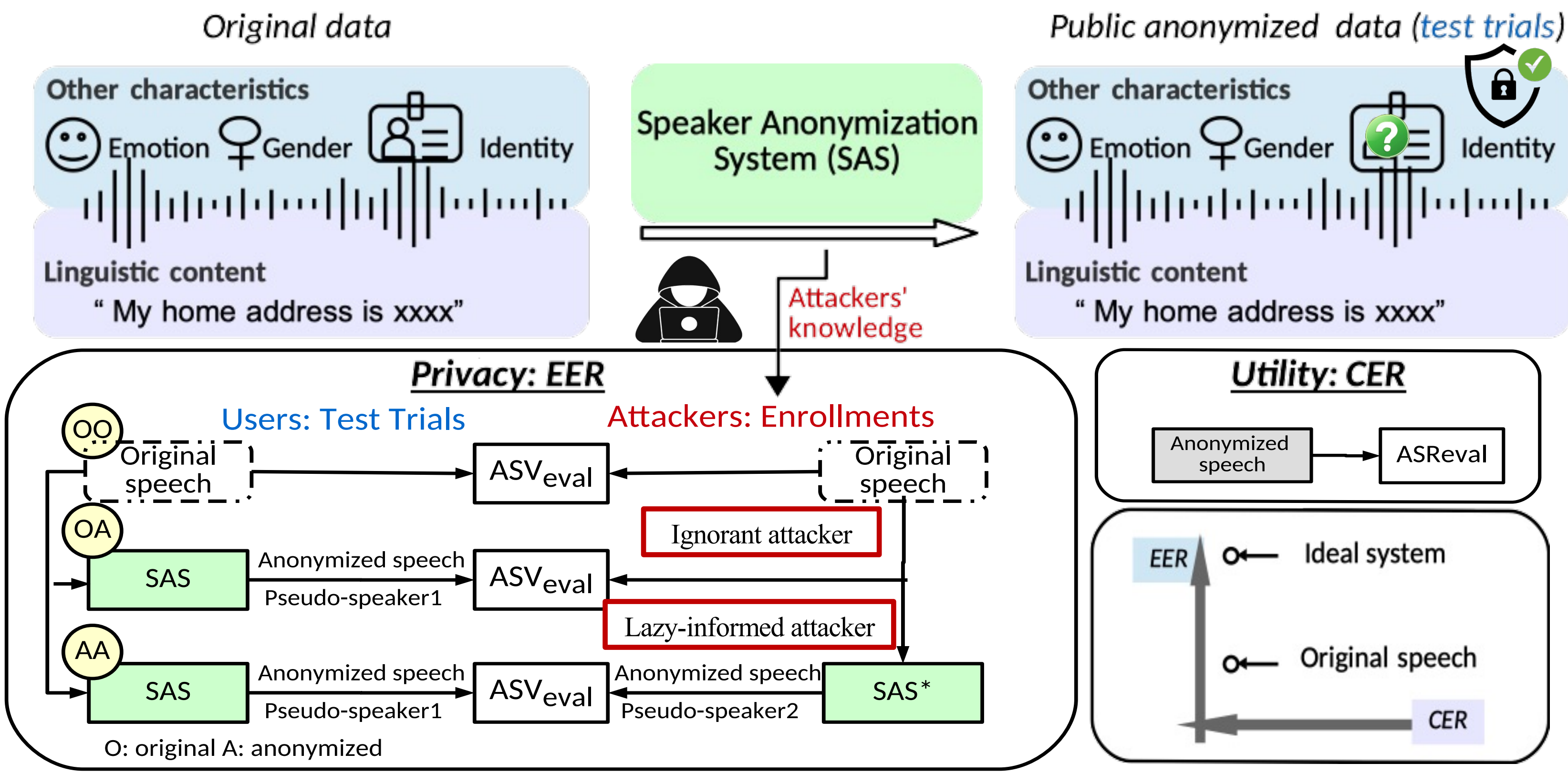
Analyzing Language-Independent Speaker Anonymization Framework under Unseen Conditions

Xiaoxiao Miao[1], Xin Wang[1], Erica Cooper[1], Junichi Yamagishi[1], Natalia Tomashenko[2]

[1] National Institute of Informatics, Japan
[2] LIA, University of Avignon, France

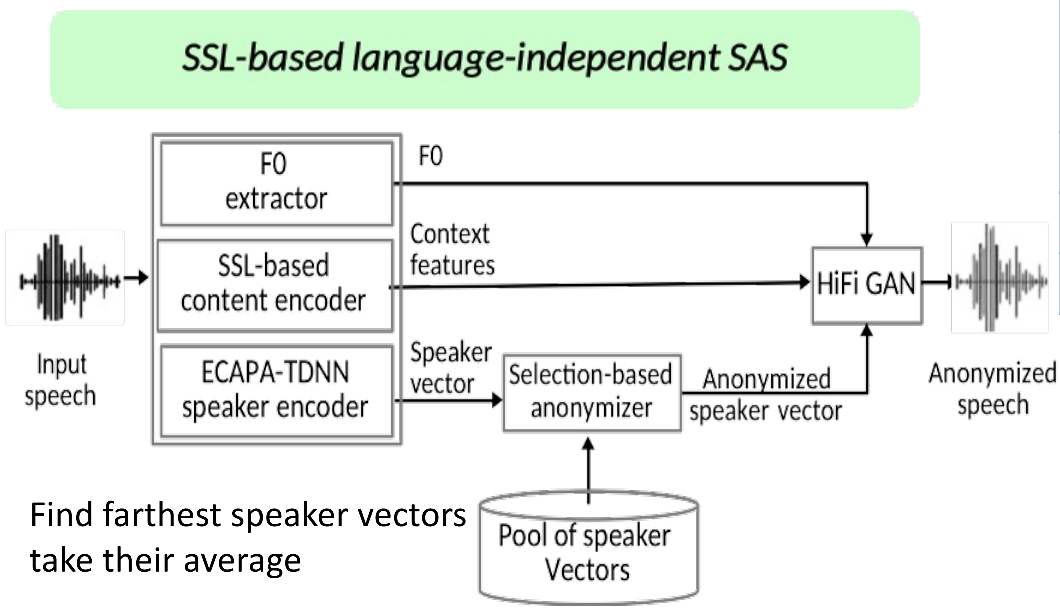
Introduction of speaker anonymization

- Definition[1] from VoicePrivacy challenge (VPC) 2020
  - Suppress the speaker's identity
  - Preserve other information, allow the downstream tasks



SSL-based language independent SAS

- Previously proposed SSL-based SAS[2]:
  - Does not require other language-specific resources, allowing the system to anonymize speech data from any language
  - For English: comparable EER and better WER than VPC baselines
  - For Mandarin: acceptable EER while degraded CER



Find farthest speaker vectors take their average

English	Average EER(%)	
	VPC	SSL
Unpro.	3.73	
Ignorant	51.63	47.06
Lazy-info.	32.44	30.10

English	Average WER(%)	
	VPC	SSL
Original	7.89	
Anony.	10.93	9.25

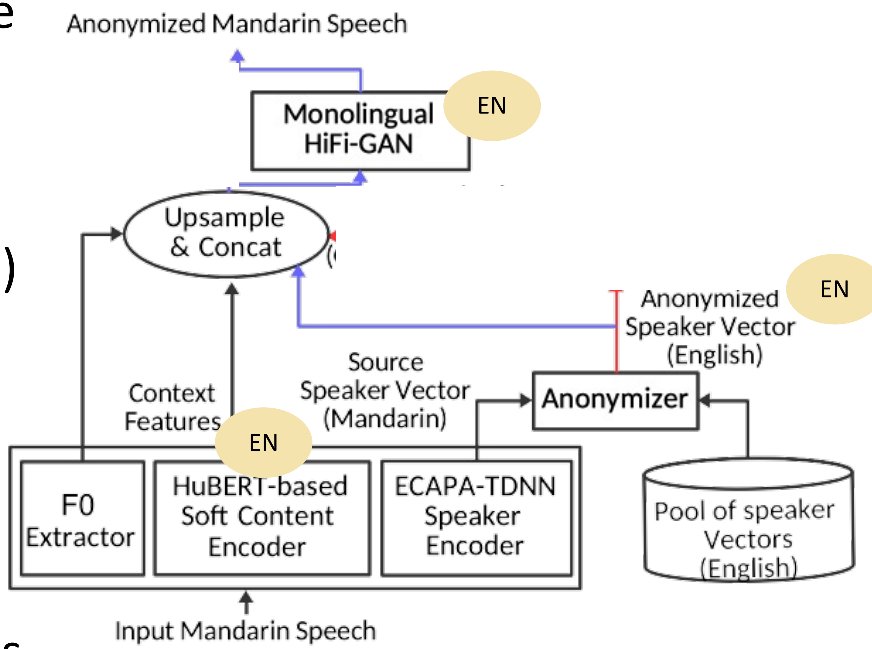
Mandrain	EER(%)-SSL
Unprotect	2.04
Ignorant	37.90
Lazy-info.	22.58

Mandrain	CER(%)-SSL
Original	10.36
Anony.	18.92

[2] X Miao, et al., "Language-independent speaker anonymization approach using self-supervised pre-trained models," Odyssey2022

SSL-based SAS performance bottleneck

- What is the performance bottleneck of SSL-based SAS under unseen conditions?
  - Monolingual content encoder -> multilingual SSL-based soft content ✗
  - Monolingual HiFi-GAN -> multilingual HiFi-GAN ✓
    - To achieve a robust vocoder, the training dataset has to cover diverse speakers and languages[3]
  - Anonymized speaker vector (English) -> map to multilingual or Mandarin space ✓
    - Speaker vectors contain speaker-unrelated information from the source domain, e.g., channel conditions and lexical contents[4,5]



Anonymization stage

[3] J. Lorenzo-Trueba,et al, “Towards Achieving Robust Universal Neural Vocoding,” Interspeech 2019
[4] D. Raj, et al, “Probing the information encoded in x-vectors,” ASRU 2019
[5] J. Williams and S. King, “Disentangling style factors from speaker representations.” Interspeech 2019

Experiment details

- Settings:
  - Test set sampled from AISHELL-3[6]: 10120 enrollment-test trials
  - ASVeval: F-ECAPA trained on CN-Celeb-1&2[7]
  - ASReval: publicly available transformer trained on AISHELL-1[8]
- Vocoder: Monolingual HiFi-GAN vs. Multilingual HiFi-GAN

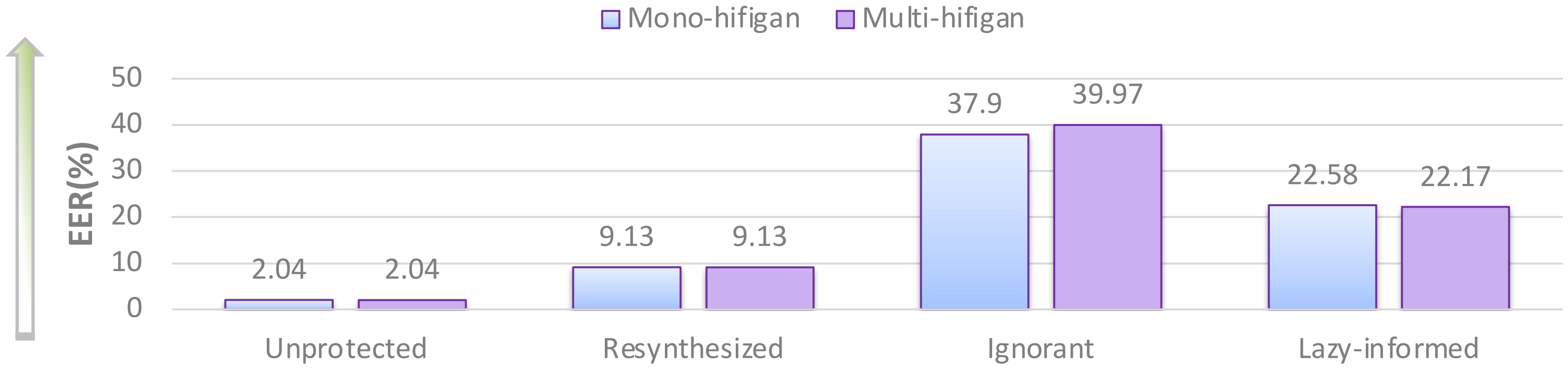
Model	Dataset
Mono-hifigan	LibriTTS train-clean-100[9]
Multi-hifigan	German[10] & Italian[10] & Spanish[10] & LibriTTS train-clean-100

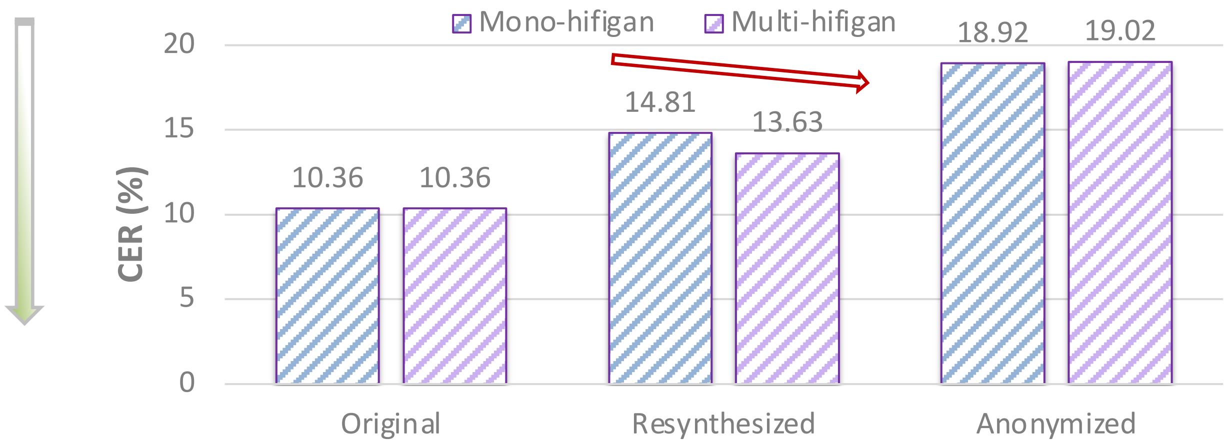
- Anonymized speaker vector: General and Mandarin CORAL

Types	Dataset
General CORAL	German & Italian & Spanish
Mandarin CORAL	AISHELL-3-test-left

[6] Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li, "AISHELL-3: A Multi-Speaker Mandarin TTS Corpus," INTERSPEECH, 2021
[7] Lantian Li, et al., "CN-Celeb: multi-genre speaker recognition," Speech Communication, 2022
[8] Hui Bu, et al.,,"Aishell-1: An open-source Mandarin speech corpus and a speech recognition baseline," O-COCOSDA 2017
[9] H. Zen, et al, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," arXiv preprint arXiv:1904.02882, 2019
[10] V. Pratap, et al, "MLS: A Large-Scale Multilingual Dataset for Speech Research," Interspeech 2020

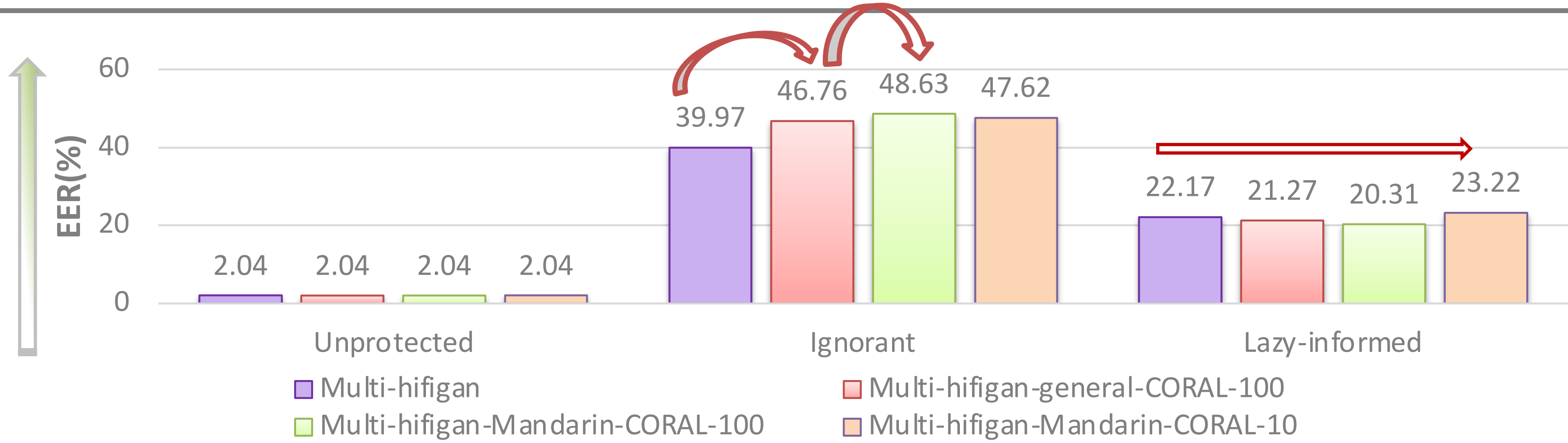
Mono-HiFiGAN vs. Multi-HiFiGAN

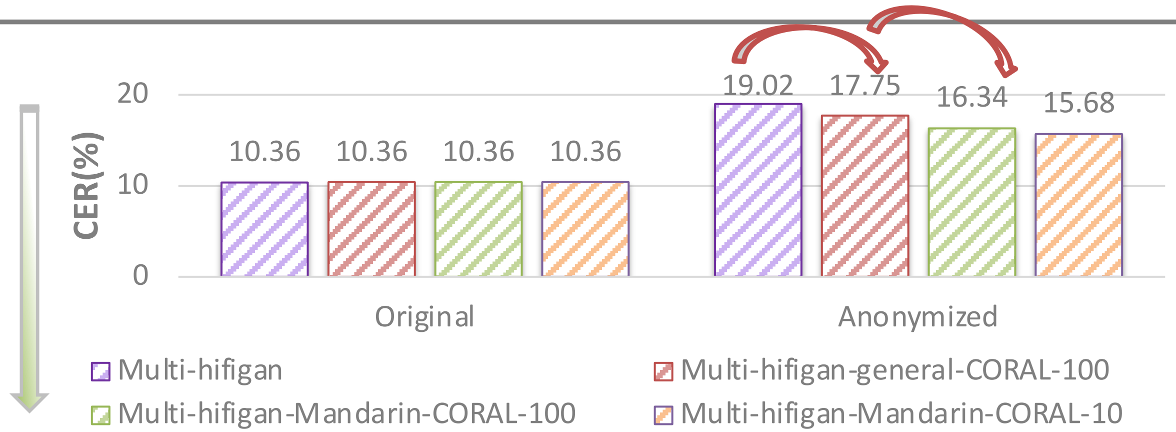




- The multilingual HiFi-GAN:
  - Keep the similar protection ability of the speaker identity
  - Better preservation of the speech contents

Coral trasformation

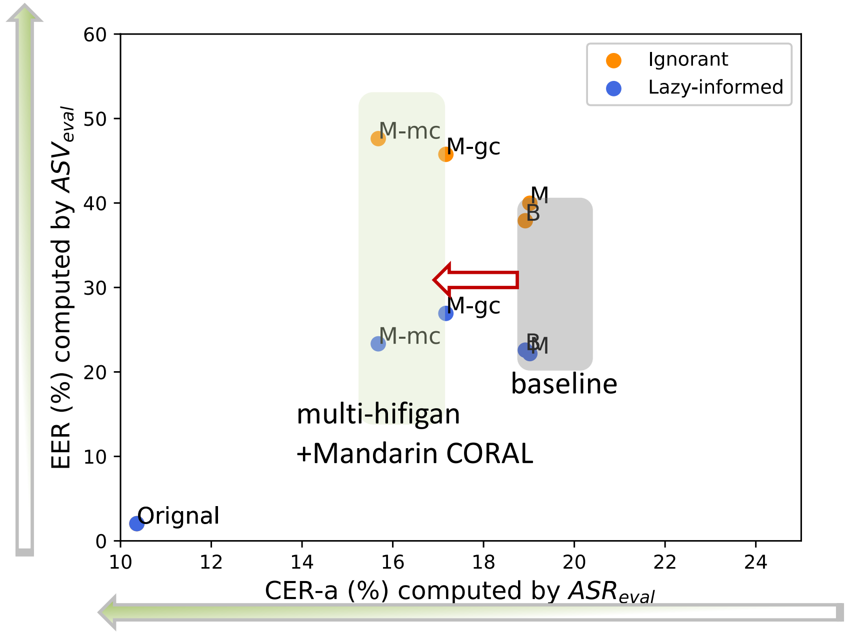




- CORAL achieves higher EER on Ignorant condition and lower CER
- Mandarin CORAL performed better on CERs than the general CORAL
- The mismatch on the anonymized speaker vectors severely affect the SAS

Conclusions

- The performance bottleneck of SSL-based SAS
  - *HiFi-GAN*: increasing the language diversity for the HiFi-GAN benefits the preservation of speech contents
  - *Anonymized speaker vector*: the mismatch on the anonymized speaker vectors severely affect the SAS.
  - The SAS using multilingual HiFi-GAN and CORAL strategy improve both privacy and utility



B: mono-hifigan (SSL-based baseline)
M: multi-hifigan
M-gc: multi-hifigan + general CORAL
M-mc: multi-hifigan + Mandarin CORAL

Audio samples and source code are available at
https://github.com/nii-yamagishilab/SSL-SAS

Thanks for listening
Q&A