Language-independent Speaker Anonymization Approach using Self-supervised Pre-trained Models

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

Odyssey 2022

Outline

- Introduction
- Motivation
- Language-independent SSL-based Speaker Anonymization System
  - SSL-based Content Encoder
  - ECAPA-TDNN Speaker Encoder
  - HiFi-GAN
- Experimental Results and Analysis
- Conclusions and future work

Introduction

- Definition[1] from VoicePrivacy challenge (VPC) 2020
  - Suppress the speaker's identity
  - Preserve the linguistic content, other paralinguistic attributes such as age, gender, emotion, and the diversity of speech
  - Allowing downstream tasks: human communication, automated processing, model training, etc.
- VPC2020 primary baseline[1]: ASR + TTS. - B1

  Step1: disentangle speech into F0, BN, and x-vector

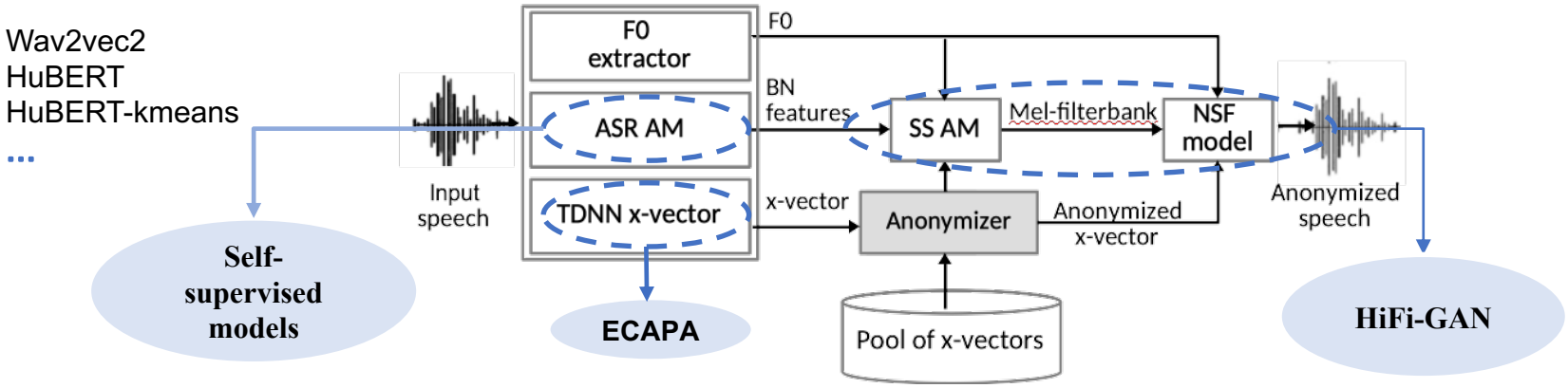
  Step2: anonymize x-vector

  Step3: synthesize anonymized speech using source F0, BN, and anonymized x-vector

[1] N. Tomashenko, et al., "Introducing the VoicePrivacy Initiative," Interspeech2020

Motivation

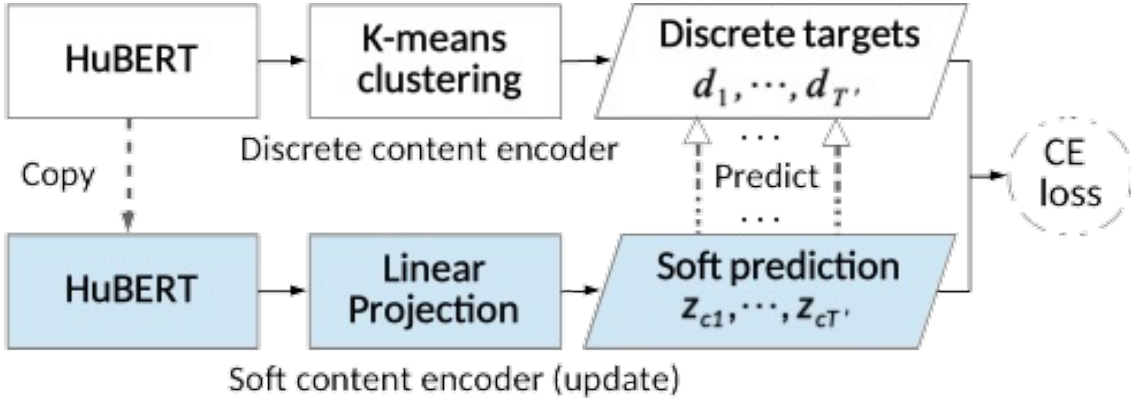
- Can we create one speaker anonymization system (SAS) that can anonymize speech from unseen languages?
  - Directly use B1- content is distorted because of the language-specific ASR AM



- SSL-based content encoder provides general context representations
- Updating the TDNN x-vector to the ECAPA-TDNN speaker encoder
- Using a HiFi-GAN as the speech waveform generation model instead of the traditional TTS pipeline to make the system more efficiently

SSL-based SAS --SSL

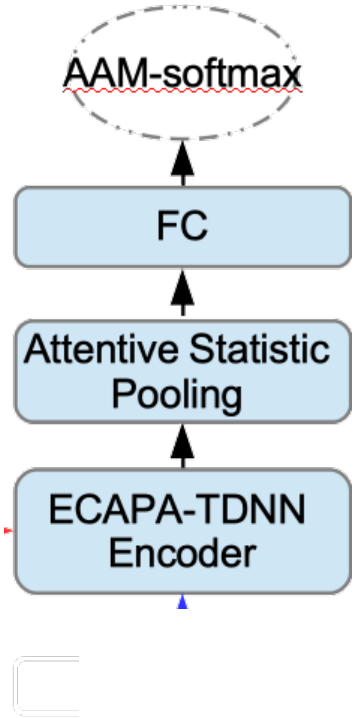
- HuBERT[2]
  - Continuous features contain both context and speaker information.
    Not suitable for speech disentanglement
- HuBERT-km[3]
  - Apply k-means algorithm over HuBERT continuous features
  - Discrete features get rid of speaker identity attributes.
    While inaccurate discrete features lead to incorrect pronunciations
- HuBERT-based Soft Content Encoder[4]
  - Trade-off between continuous and discrete features



[2] Wei-Ning Hsu, et al., "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," TASLP 2021
[3] Kushal Lakhotia, et al., "Generative spoken language modeling from raw audio," TACL 2021
[4] Benjamin van Niekerk, et al, "A comparison of discrete and soft speech units for improved voice conversion," ICASSP 2022

SSL-based SAS --ECAPA

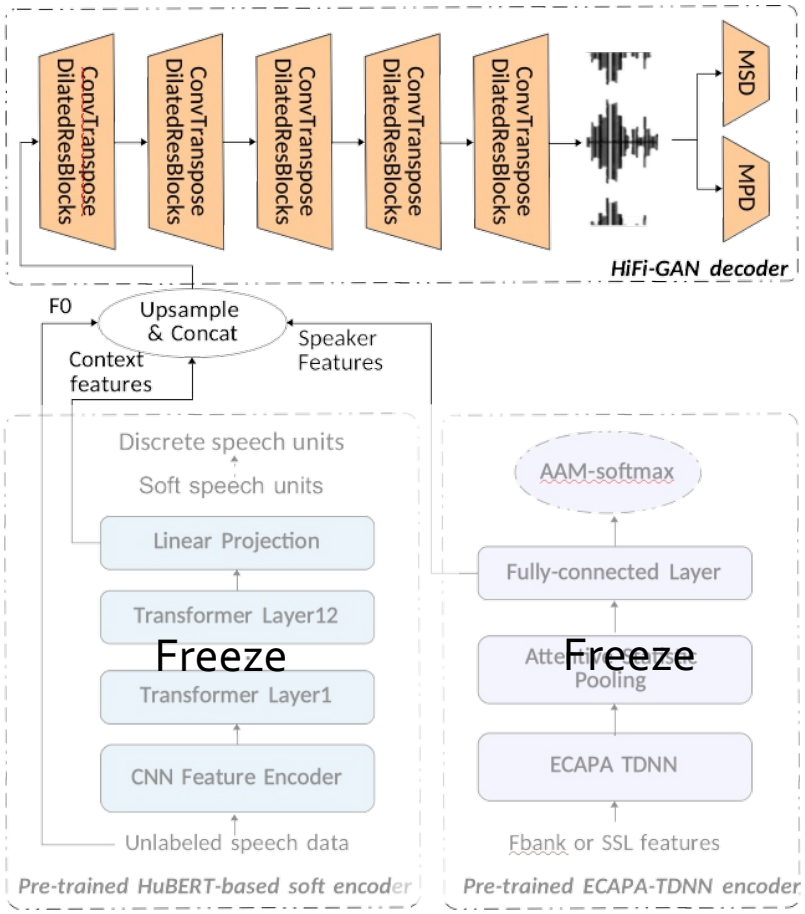
- ECAPA-TDNN[5]



[5] Brecht Desplanques, et al., "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," INTERSPEECH 2020

SSL-based SAS --HiFiGAN

- Frame-wise F0, context, and segment al-level speaker embedding are up-sampled and concatenated
- Then passed to HiFi-GAN[6], which contains one generator and two discriminators:
  - Multi-period discriminator (MPD) captures the periodic patterns
  - Multi-scale discriminator (MSD) exploring long-range and consecutive interactions



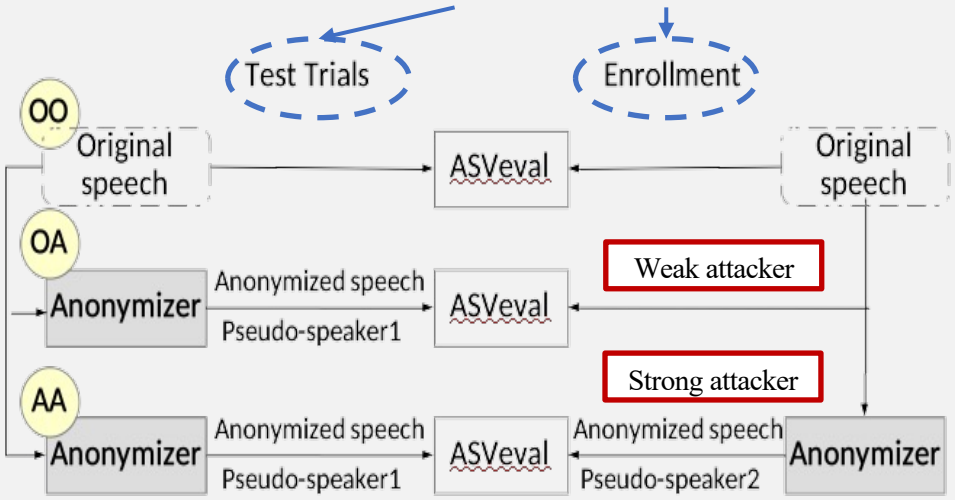
[6] Jungil Kong, et al., "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," NIPS 2020

Experimental Results and Analysis

- Evaluation Plan

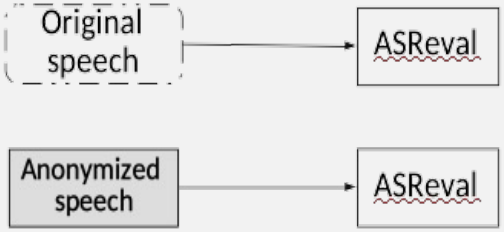
Privacy: EER

Game between users and attackers



- Unprotected (OO): no anonymization
- Ignorant attacker (OA): users anonymize
- Lazy-informed (AA): users & attackers anonymize, attackers know the SAS, but don't know the specific parameters, would get the different pseudo speaker.

Utility: WER



Experiment details

Training set

HuBERT soft : *LibriSpeech-100*[9]

HiFi-GAN: *LibriTTS-100*[10]

Test set :

LibriSpeech[10], *VCTK*[11]

- ASV_{eval}^{Eng}: TDNN
- ASR_{eval}^{Eng}: TDNN-F

} *LibriSpeech-360*[9]

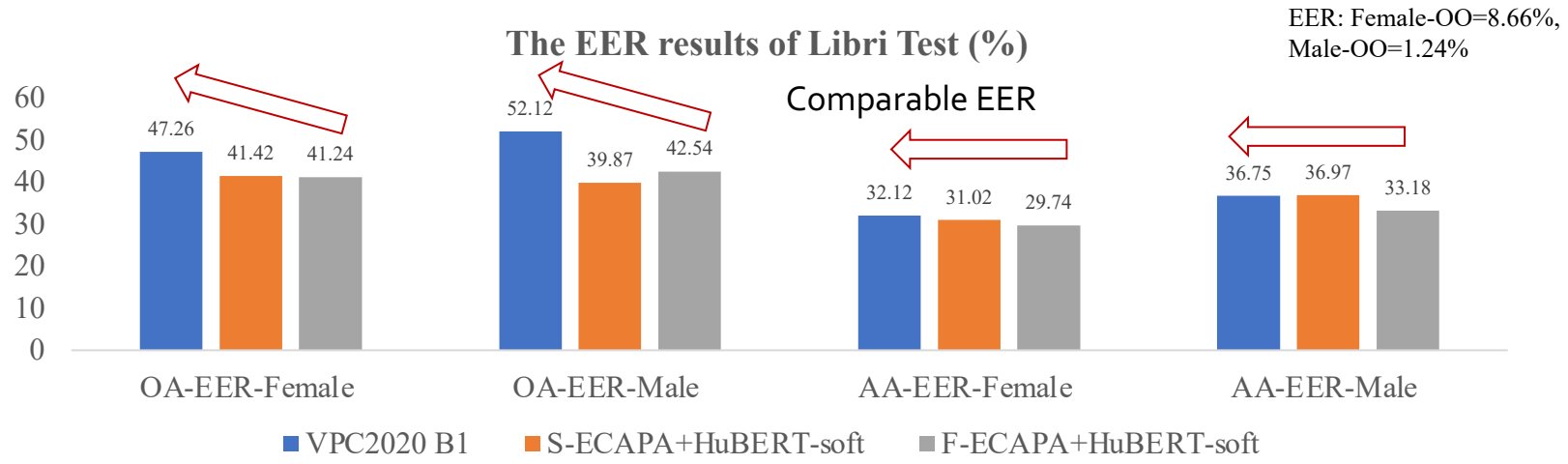
[9] Vassil Panayotov, et al., "Librispeech: an ASR corpus based on public domain audio books," ICASSP 2015

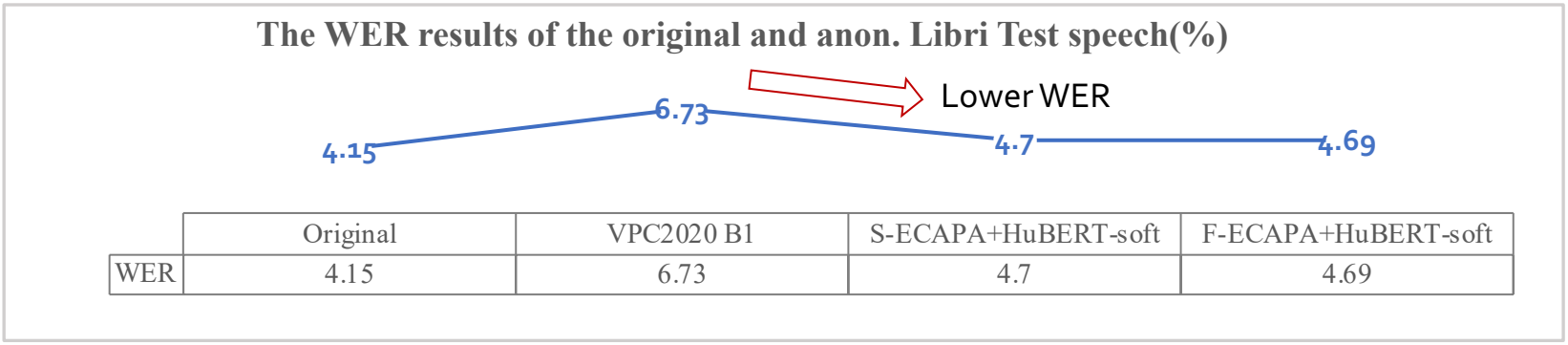
[10] Heiga Zen, et al., "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," arXiv preprint arXiv:1904.02882, 2019

[11] Christophe Veaux, et al., "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019.

Experimental Results and Analysis

- English Speaker Anonymization





	Original	VPC2020 B1	S-ECAPA+HuBERT-soft	F-ECAPA+HuBERT-soft
WER	4.15	6.73	4.7	4.69

- The proposed SAS evaluated on English data:
  - Protect the speaker information almost as well as VPC B1
  - Provide more reliable linguistic information.

Experimental Results and Analysis

- Mandarin Speaker Anonymization
  - Experiment details
    - Test set sampled from AISHELL-3[12]: 10120 enrollment-test trials
    - ASV_{eval}^{mand}: F-ECAPA trained on CN-Celeb-1&2[13]
    - ASR_{eval}^{mand}: publicly available Transformer* trained on AISHELL-1[14]

(%)		S-ECAPA + HuBERT-soft	F-ECAPA + HuBERT-soft
EER	OO	2.04	
	OA	40.81	37.58
	AA	23.26	22.98
CER	Original	10.36	10.36
	Anony.	21.18	18.86

    - Proposed SAS is simpler without any language-specific models can be adopted to other languages successfully
    - CERs on the anonymized trials were increased to about 20%, suggesting degradation on the speech content

*SpeechBrain: https://github.com/speechbrain/speechbrain/tree/develop/recipes/AISHELL-1/ASR/transformer

[12] Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li, "AISHELL-3: A Multi-Speaker Mandarin TTS Corpus," INTERSPEECH, 2021
[13] Lantian Li, et al., "CN-Celeb: multi-genre speaker recognition," Speech Communication, 2022
[14] Hui Bu, et al.,"Aishell-1: An open-source Mandarin speech corpus and a speech recognition baseline," O-COCOSDA 2017

Conclusions and Future Work

- Can we create one speaker anonymization system (SAS) that can anonymize speech from unseen languages?
  - Yes, we can use Language-independent SSL-based SAS.

- The limitation of the proposed SAS:
  - There is still room to improve the performance of the SAS under unseen condition.
  - See our Interspeech2022 paper: https://arxiv.org/abs/2203.14834

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

Thanks for listening

Q&A