SecureSpeech: prompt-based speaker and content protection

SINGAPORE INSTITUTE OF TECHNOLOGY · 昆山杜克大学 DUKE KUNSHAN UNIVERSITY · NII 大学共同利用機関法人 情報・システム研究機構 国立情報学研究所 National Institute of Informatics

Belinda Soh Hui Hui, Xiaoxiao Miao, Xin Wang (presenter, wangxin@nii.ac.jp)

Poster session: #2, 18
PaperID 49

Background: protection of private speaker information

- Information in speech utterances

***Linguistic**:* contents, prosody of timing, pitch, …

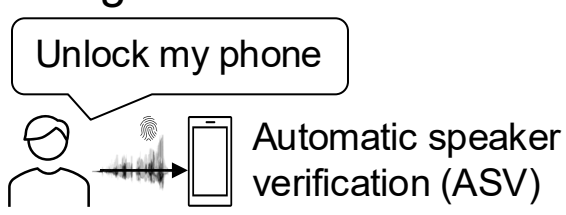
***Paralinguistic**:* emotion, …

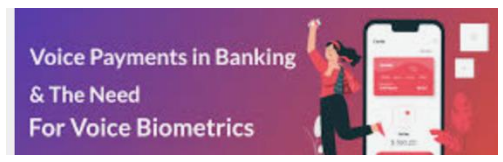
Biometric voiceprint, speaker ID

Your voice is your identity as a speaker

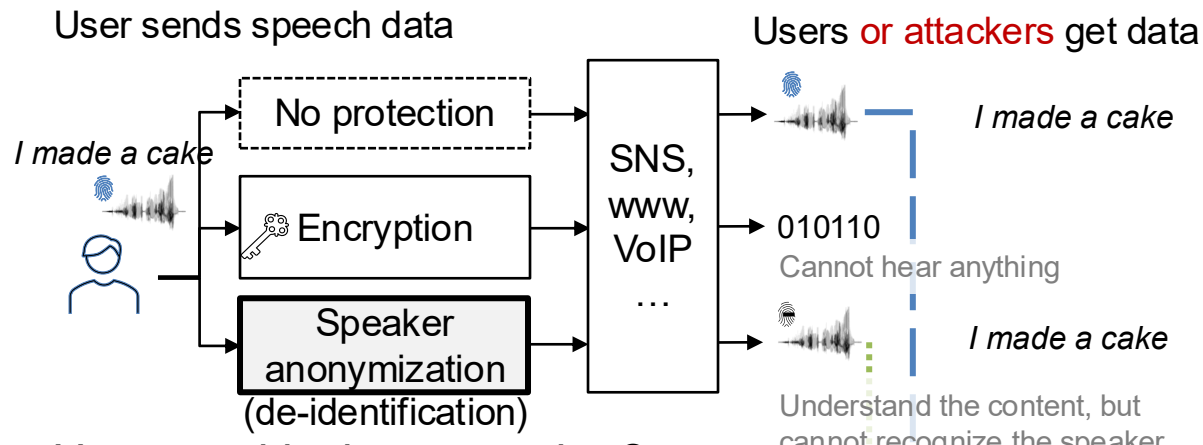
This poster is only on protection of speaker identity

- Usage of biometric information in speech

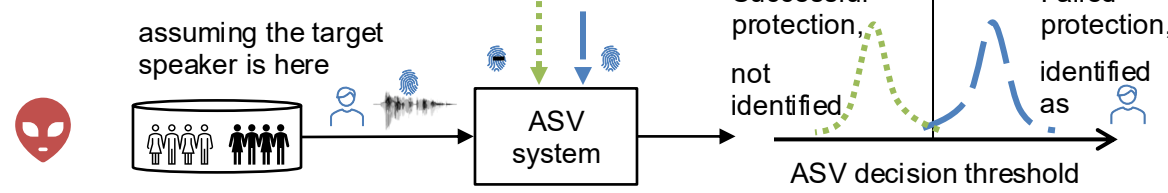




- How to protect biometric information in speech?
  - Similar idea to face de-identification



- How good is the protection?
  - Attacker:

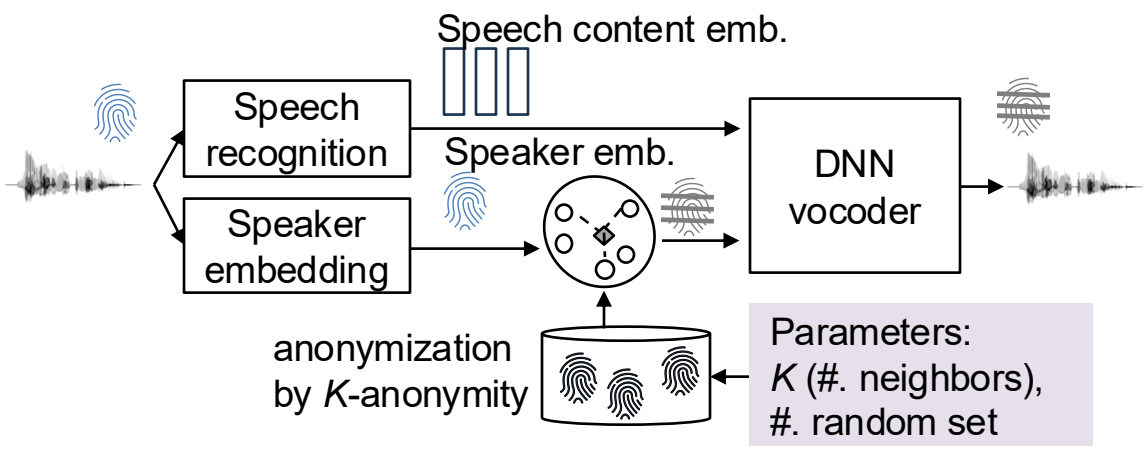


better protection = attacker gets a lower recognition rate

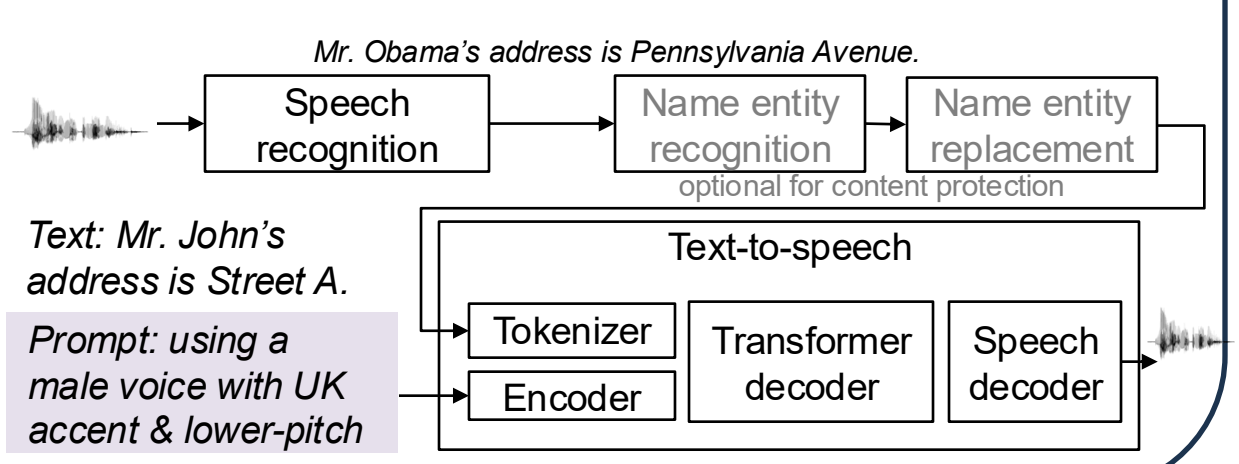
  - User: the protected speech is intelligible & natural

Protecting speaker biometric information by anonymization

- Conventional approach [1]
  - Similar to deep neural network (DNN) voice conversion
  - Users set parameters via trial-and-errors
  - Attacker can still link anonymized & original speakers



- Proposed **SecureSpeech**
  - Automatic speech recognition (ASR) + text-to-speech (TTS)
  - User describes the voice using text prompt – easier to use
  - Not linked to original speaker identity – better protection



Experiment configuration

- Content protection is off
- Evaluation dataset: SLUE-VoxPopuli [2]
  - #. English speakers: 161
  - #. utterances (in total): 3,729
- Attacker's ASV system (pre-trained)
  - Popular ECAPA-TDNN, on VoxCeleb2 [3]
  - (ignorant attacker in Voice Privacy Challenge [1])
- Proposed system (pre-trained modules)
  - ASR: wav2vec 2.0-large ft. on Librispeech 960 [4]
  - TTS: Parler-TTS [5]
    - Transformer decoder: 24 blocks
    - Speech decoder: neural codec DAC [6]
  - Speaker prompt: randomly combined from templates of gender, English accent, speaking rate…

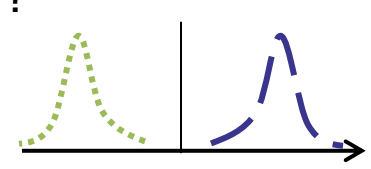
Conclusions:
- The proposed system is effective against attackers using pre-trained ASV; easy to use (text prompt)
- Future work: stronger attacking model

Experiment result

See other results in the paper

- Is speaker ID protected from attacker?

	No protect	Proposed
ASV rec. rate by attacker	100%	0%
	the lower the better ↓	



  - Not surprising: Parler-TTS's training speakers are different from test speakers

Yes, no link to original speakers

- Does the protected speech sound good?

	No protect	Proposed
ASR error ↓	23%	**16%**
MOS (squim) ↑	**4.48**	4.01

Yes, quality is not degraded severely

- Impact of text prompts?
  - Fix one attribute, randomize other attributes
  - Protection is **equally** good: 0% ASV rec. rate
  - Speaking "quickly" degrades quality

Attributes	Subcategories	ASR err.↓	MOS↑
Gender	Female	14.85	4.10
	Male	16.88	4.07
Pitch	Low-pitched	13.25	4.12
	Normal	14.10	4.25
	High-pitched	16.01	4.05
Speaking rate	Slowly	15.25	4.28
	Normally	12.39	4.23
	Quickly	15.55	3.91

[1] M. Panariello et al., "The VoicePrivacy 2022 Challenge: Progress and Perspectives in Voice Anonymisation," IEEE TASLP, pp. 1–14, 2024
[2] S. Shon et al., "SLUE: New Benchmark Tasks for Spoken Language Understanding Evaluation on Natural Speech," Proc. ICASSP, 7927–7931, 2022
[3] B. Desplanques et al, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in Proc. Interspeech, 2020, pp. 3830–3834.
[4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NIPS*, 2020, pp. 12449–12460.
[5] D. Lyth and S. King, "Natural language guidance of high-fidelity text-to-speech with synthetic annotations," Feb. 02, 2024, arXiv: arXiv:2402.01912.
[6] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved RVQgan," *Proc. NIPS*, vol. 36, 2024.