Protecting gender and identity with disentangled speech representations

Dimitrios Stoidis and Andrea Cavallaro

dimitrios.stoidis@qmul.ac.uk, a.cavallaro@qmul.ac.uk

Proc. Interspeech 2021, 1699-1703, doi: 10.21437/Interspeech.2021-2163

2021 Intelligent Sensing Winter School (December, 7-10)

CIS centre for intelligent sensing

Queen Mary University of London

Introduction

- Motivation
  - Privacy risks related to sharing our voice (speaker profiling, biased decision-making)
  - Biometric information in speech not necessary for Automatic Speech Recognition (ASR)
  - Use gender to protect identity information as well

- Contributions
  - Learnable embeddings to independently encode gender and identity attributes
  - Improvement on the privacy-utility trade-off by with existing state-of-the-art methods
  - Gender information used to protect identity of the speaker

Privacy Settings

Two biometric attributes, identity and gender, defining 5 privacy settings.

	Identity	Gender
Same	SI	SG
Random	RI	RG
	RISG	SIRG

Table of privacy settings considered in this work. 5 privacy settings are defined combining identity and gender information, by keeping one attribute fixed and the other random. A random identity from a speaker in the set will be chosen in the random identity (RI) setting.
KEY-- Same (S): Biometric attribute is same as source speaker,
Random (R): Biometric attribute is randomly assigned.

Proposed Method

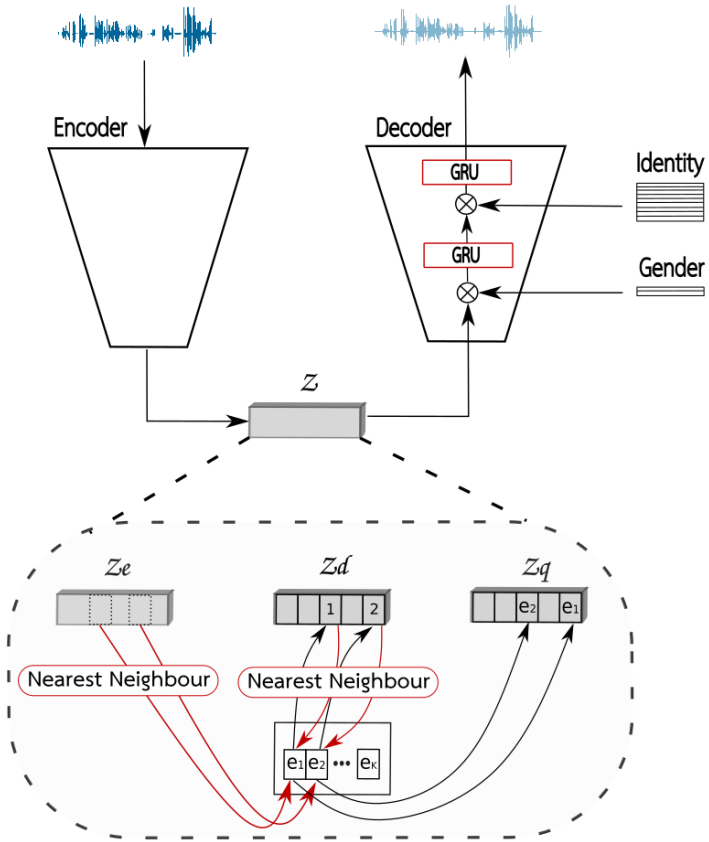
Disentangled Representation Learning to protect gender and identity attributes:

1. Vector Quantisation (VQ) of speech content into discrete latent space
2. Train decoder to disentangle identity and gender from content information
3. Reconstruct voices by decoding according to defined privacy settings
4. Validate voices with respect to performance on privacy and utility

Model Architecture

1. Encoder compresses speech information:
    - Discretisation of content with Vector Quantisation (VQ).
    - Nearest neighbour lookup between embedding vectors (codewords) from learned embedding space (codebook) and quantised content vectors.
2. Decoder reconstructs speech with respect to desired privacy setting:
    - Additional Gated Recurrent Unit (GRU) layer to combine identity and gender embeddings.
    - Disentangled identity and gender embeddings are concatenated with quantized content

Vector-Quantised VAE with privacy settings



Results (1)

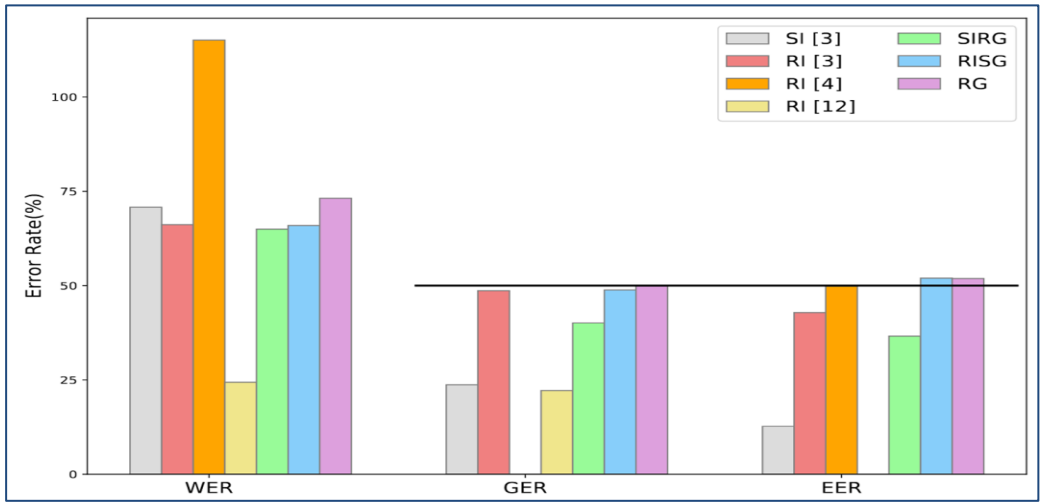
Ref.	Settings	Utility	Privacy	
		WER	Acc	EER
EDGY [3]	SI	70.78	76.27	12.70
EDGY [3]	RI	66.17	51.36	42.85
Disentangl.VC [4]	RI	115.10	–	49.85
Client-VAE [12]	RI	24.38	77.80	–
ours	SIRG	64.99	59.89	36.63
	RISG	65.92	51.15	52.00
	RG	73.16	50.01	51.88

Table of results on utility and privacy on LibriSpeech clean test set. Improved privacy-utility trade-off with respect to current methods using identity information. Random Gender (RG) setting shows that gender can be used to protect identity information against attribute inference attacks.

Key -- WER: word error rate (%),
Acc: gender (binary) classification accuracy (%),
EER: equal error rate (%).

[3] R. Aloufi et al., “Privacy-preserving Voice Analysis via Disentangled Representations”, in Proc. of the ACM SIGSAC Conference on Cloud Computing Security Workshop, Nov, 2020.
[4] B. M. lal Srivastava et al., “Evaluating Voice Conversion-Based Privacy Protection against Informed Attackers”, in IEEE ICASSP, 2020, pp. 2802-2806.
[12] R. Wu et al., “Understanding the Tradeoffs in Client-Side Privacy for Speech Recognition”, preprint arXiv:2101.08919.

Results (2)



Privacy and utility error rate values across models considering gender and/or identity information on the LibriSpeech clean test set. The horizontal line at 50% error rate denotes privacy target.

Key -- WER: word error rate (%),
GER: gender error rate(%),
EER: equal error rate (%).

[3] R. Aloufi et al., “Privacy-preserving Voice Analysis via Disentangled Representations”, in Proc. of the ACM SIGSAC Conference on Cloud Computing Security Workshop, Nov, 2020
[4] B. M. lal Srivastava et al., “Evaluating Voice Conversion-Based Privacy Protection against Informed Attackers”, in IEEE ICASSP, 2020, pp. 2802-2806
[12] R. Wu et al., “Understanding the Tradeoffs in Client-Side Privacy for Speech Recognition”, preprint arXiv:2101.08919

Conclusion

- Disentangled representations of speech can protect against gender and identity attribute inference attacks.
- Gender information can be used to protect identity of the speaker.
- Improvement on the privacy-utility trade-off by including gender information.