Recent approaches towards speaker anonymization

Brij Mohan Lal Srivastava
PhD student at INRIA

15 Sep, 2020

Inria

COMPRISE

Outline

1. Objectives and Roadmap
2. Anonymization via Adversarial Representation Learning
3. Anonymization via X-vector based Voice Conversion
4. Conclusion

Speaker recognition = Biometric identification

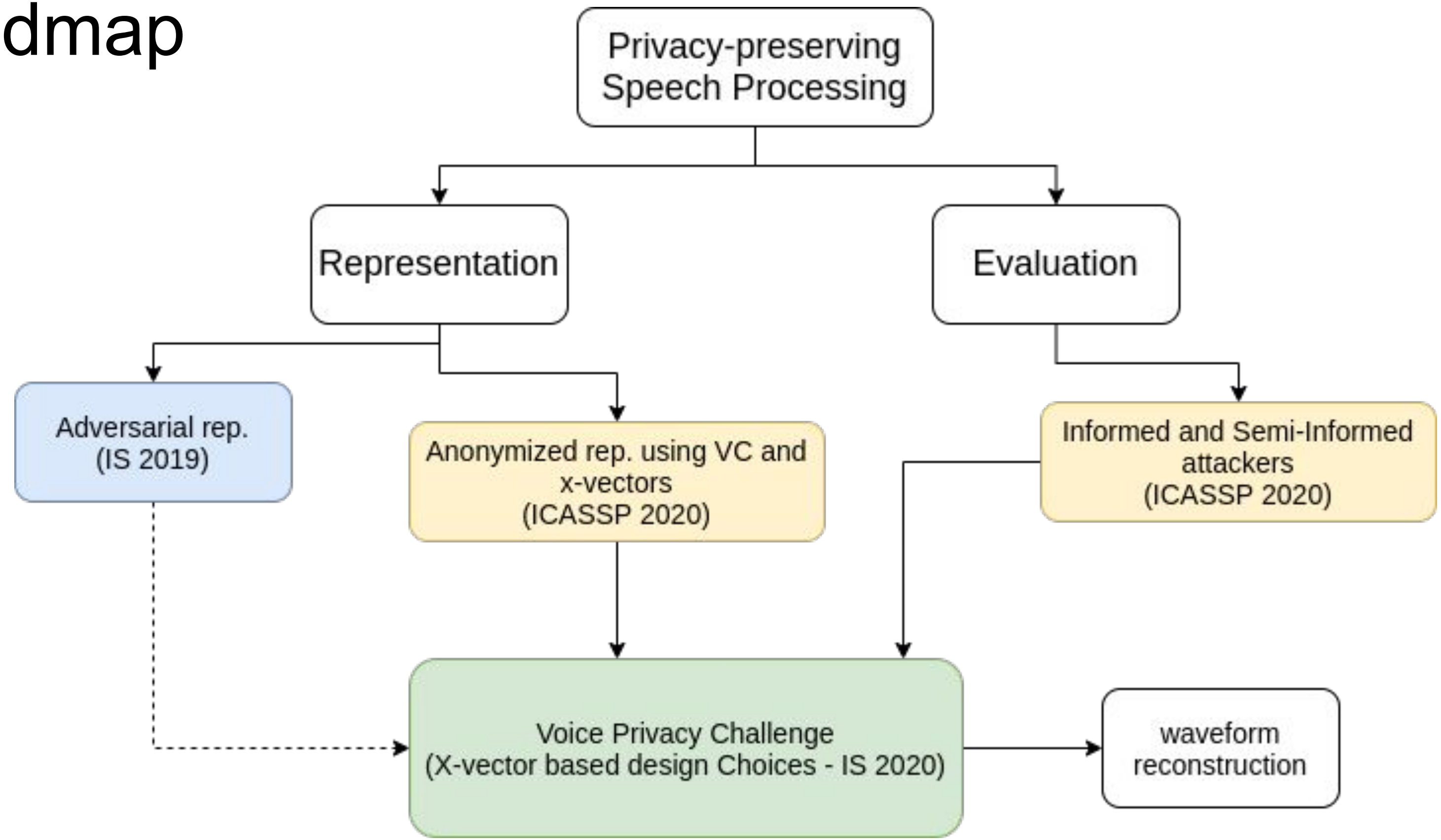
- Non-invasive / without contact
- **Distinctive** and **replicable** templates can be generated (x-vectors).
- Speaker identification and verification/authentication error rates are close to zero : X-vector + PLDA yields 2-3% error rate *(Garcia-Romero et al. 2019)*
- Increasing privacy threats require more research on speaker anonymization.

Two objectives of anonymization

- (Privacy) Data shared by the speaker cannot be linked back to the speaker.
  - Amount of privacy protection must be reported in all possible attack scenarios.
  - All attributes of speaker’s identity such as speaking rate, timbre, emotional traits, health conditions, etc. must be handled.

- (Utility) Anonymization should not affect the utility of speech, e.g. linguistic variability and content.
  - Output must be usable for further processing, e.g. pitch extraction, phonetic analysis, etc.
  - Output must be intelligible and suitable for annotation and training of automatic speech recognition (ASR) systems.

Roadmap

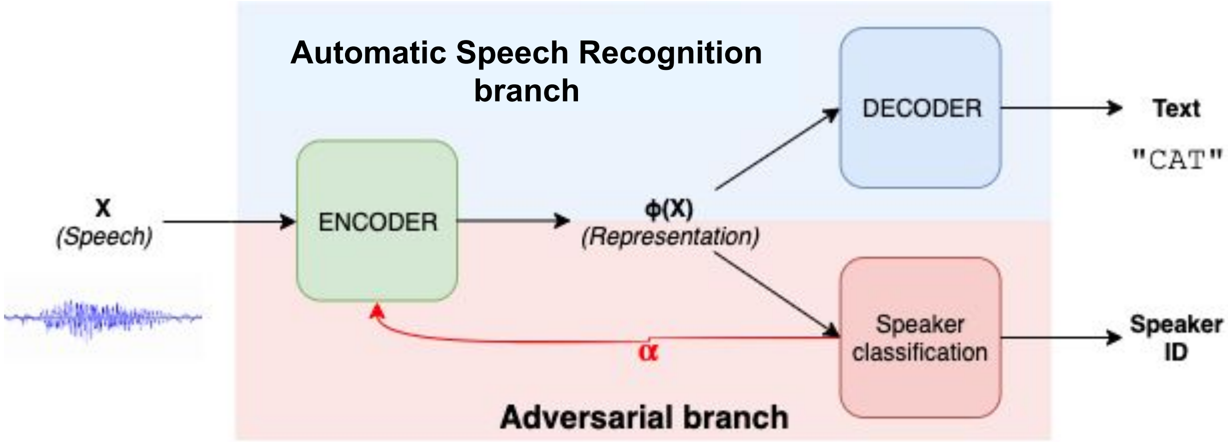

Privacy-preserving Speech Processing
Representation
Evaluation
Adversarial rep. (IS 2019)
Anonymized rep. using VC and x-vectors (ICASSP 2020)
Informed and Semi-Informed attackers (ICASSP 2020)
Voice Privacy Challenge (X-vector based design Choices - IS 2020)
waveform reconstruction

Outline

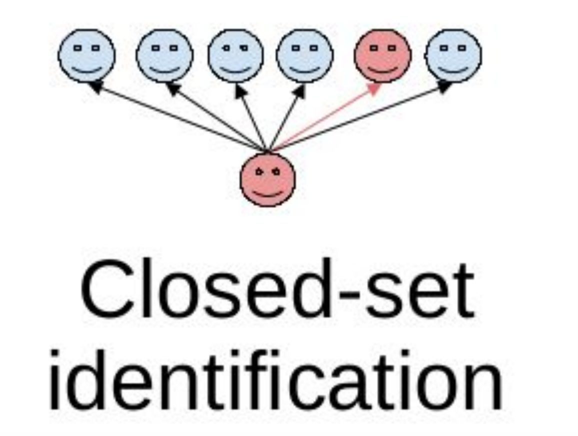
1. Objectives and Roadmap
2. Anonymization via Adversarial Representation Learning
3. Anonymization via X-vector based Voice Conversion
4. Conclusion

Adversarial anonymization

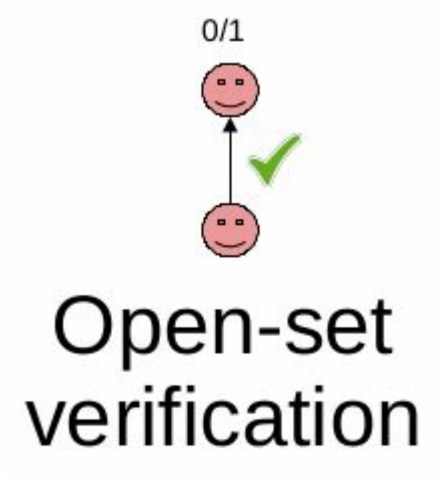
- The Adversary neural network (red) tries to learn relevant speaker-specific features
- Provides feedback to Encoder network scaled by a parameter (α) which decides the strength of anonymization



Attacker scenarios - evaluation schemes



Inside the adversarial ASR



X-Vector based Speaker Verification

Results (open vs closed set)

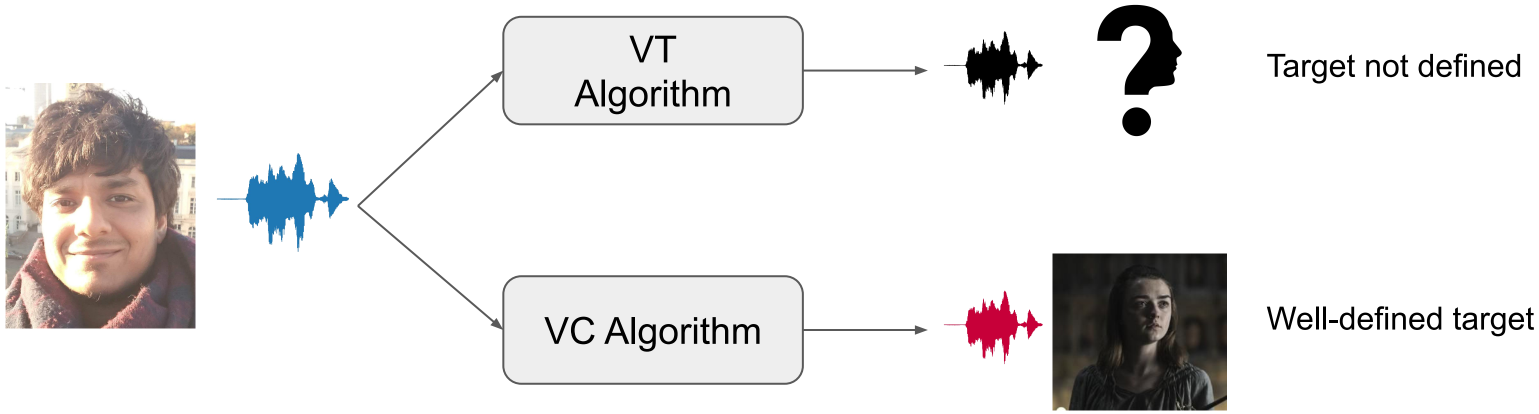
	Raw speech	Blue branch only	Adversarial Learning
Word Error Rate (ASR)		9.40	11.30 ⬆
Classification Error (closed)	2.78	51.37 ⬆	94.40 ⬆
Equal Error Rate (open)	4.31	24.77 ⬆	25.97 ⬆

- WER increases slightly indicating bearable utility loss.
- Speaker classification error (closed-set) increases significantly = significant privacy gain.
- Speaker verification error only increases slightly = insignificant privacy gain

Outline

1. Objectives and Roadmap
2. Anonymization via Adversarial Representation Learning
3. Anonymization via X-vector based Voice Conversion
4. Conclusion

Voice Conversion vs Voice Transformation



Adversarial Learning (VT technique) because we define what we **do not** want.

In VC we define what we **do** want.

Voice Privacy Challenge

The challenge is to develop anonymization solutions which suppress personally identifiable information contained within speech signals.

Using freely available datasets.

https://www.voiceprivacychallenge.org/

Baseline recipe available at:

https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2020

Organized by:

Laboratoire Informatique d'Avignon

Inria

NII Inter-University Research Institute Corporation / Research Organization of Information and Systems National Institute of Informatics

EURECOM

Supported by:

ANR

COMPRISE Cost-effective, Multilingual, Privacy-driven voice-enabled Services

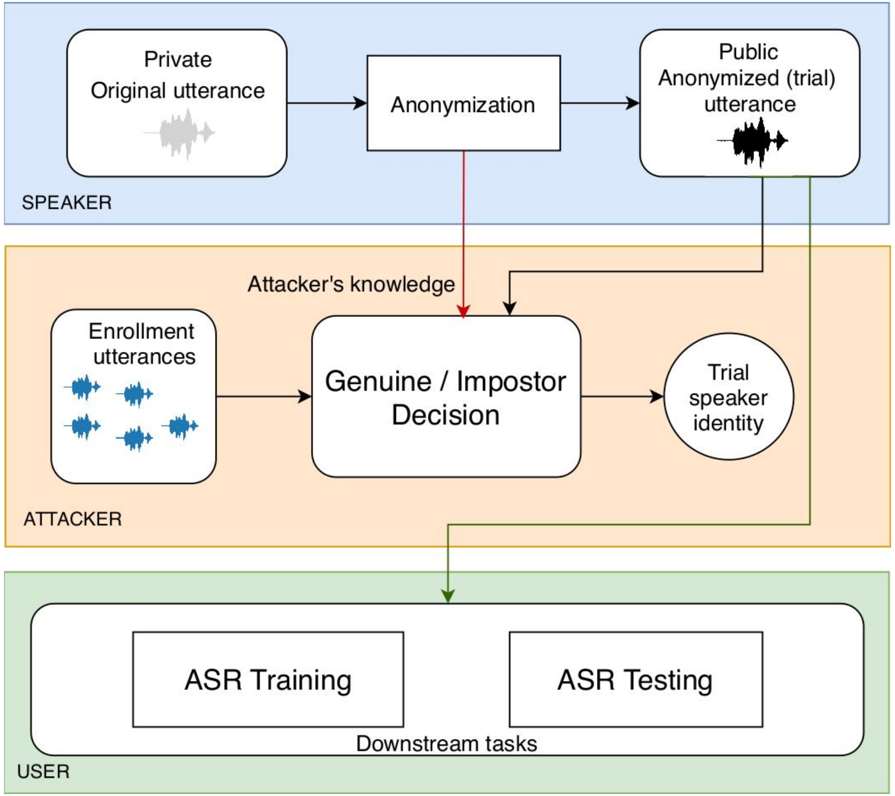
European Commission

JST Japan Science and Technology Agency

Threat model

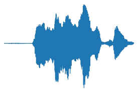
Actors:

1. Speaker
2. Attacker
3. User



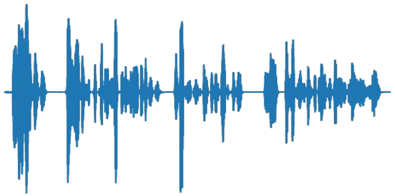
X-vectors

- Behind the state-of-the-art biometric identification techniques
- Fixed length vector to represent an utterance regardless of duration.
- Intermediate layer of a neural network trained to classify speaker



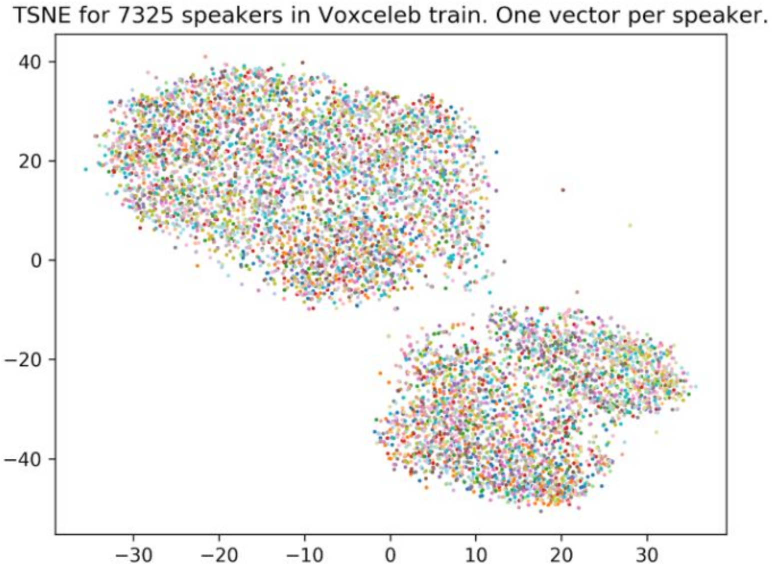
2 seconds

$$\underbrace{\begin{bmatrix} 0.2 \\ -1.6 \\ . \\ . \\ . \\ 2.5 \\ -0.3 \end{bmatrix}}_{\text{512 dimension}}$$

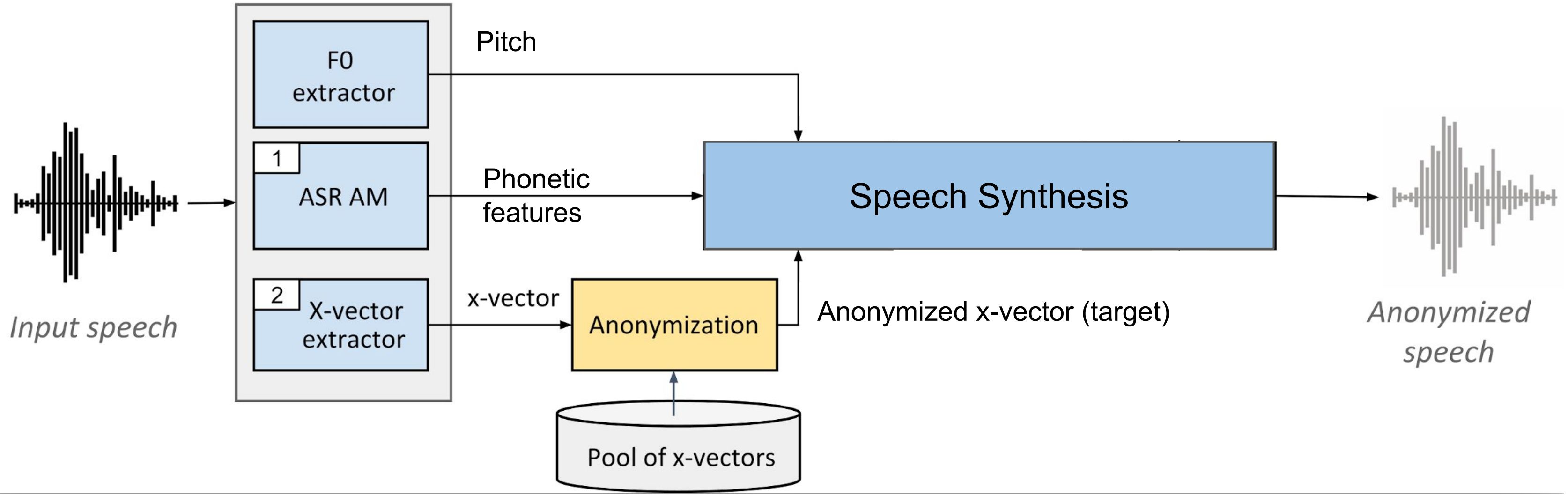


20 seconds

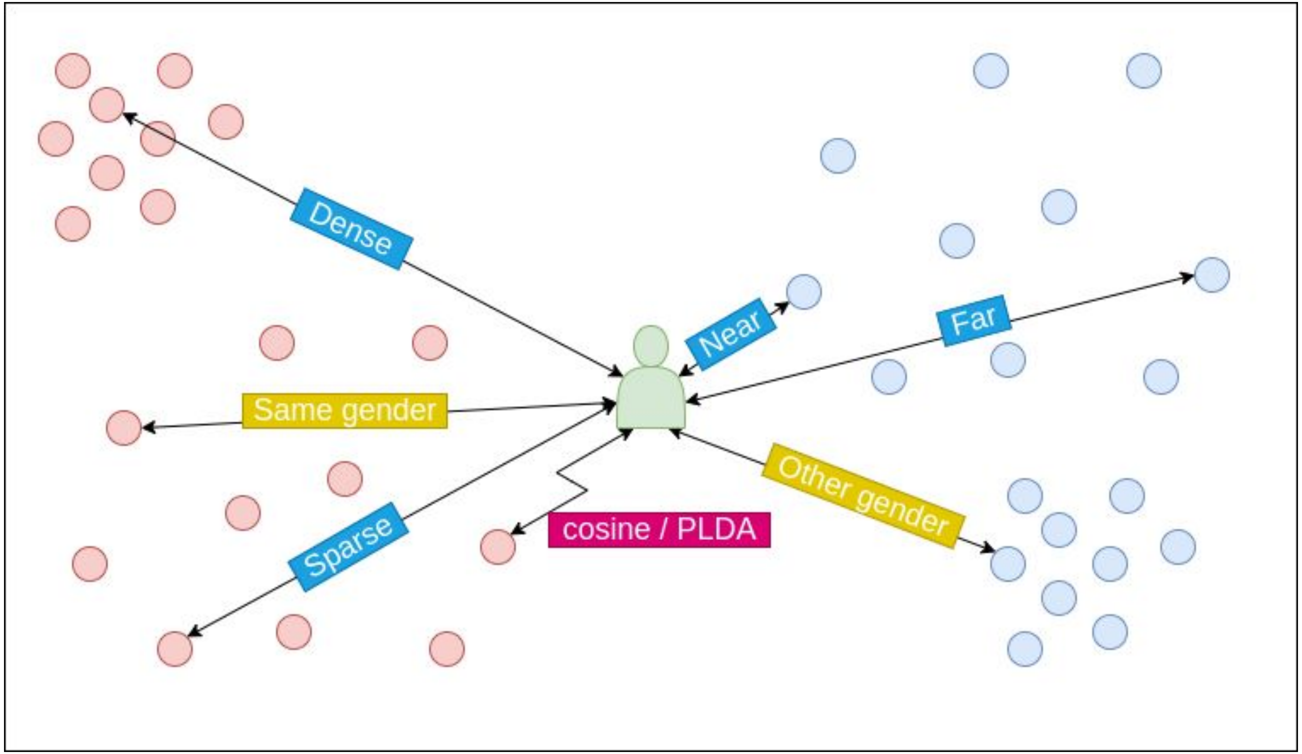
$$\underbrace{\begin{bmatrix} 0.7 \\ -4.4 \\ . \\ . \\ . \\ 0.5 \\ -5 \end{bmatrix}}_{\text{512 dimension}}$$



X-vector based speaker anonymization framework



How to optimally select target speakers from a small pool of speakers? (Speaker's Perspective)



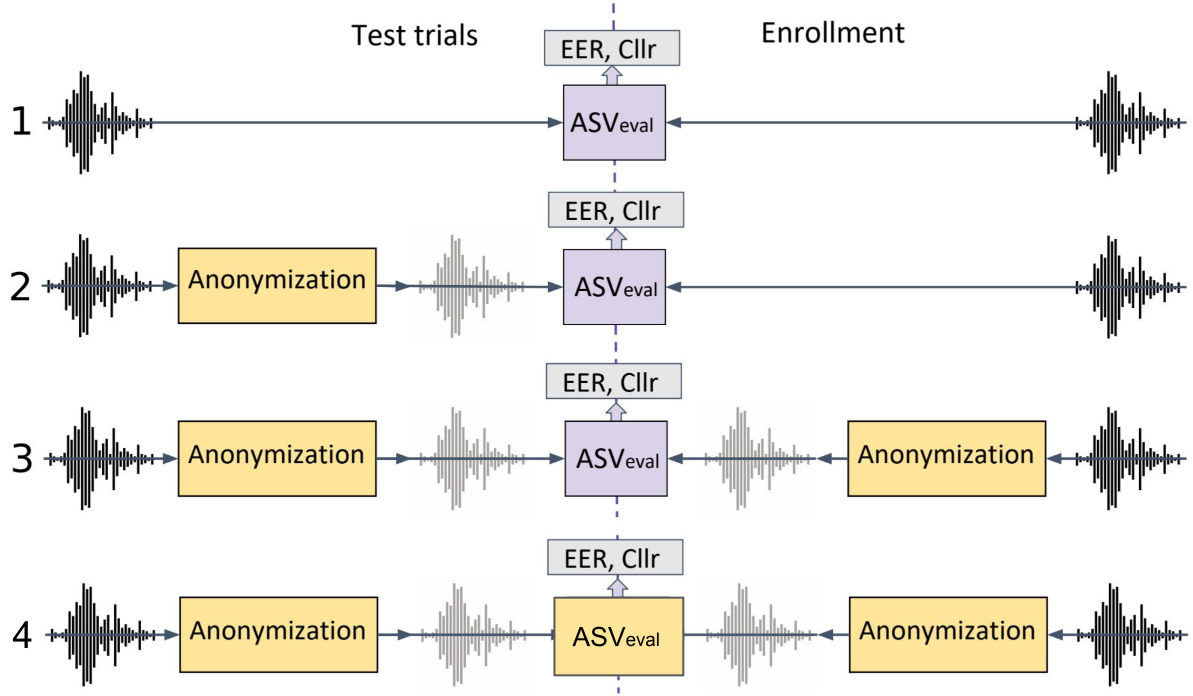
Privacy Evaluation - Attackers

Baseline

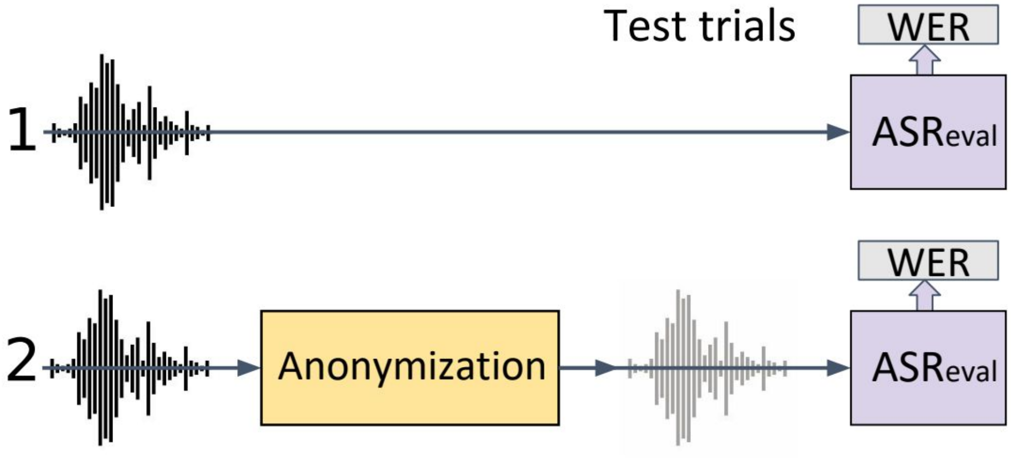
Ignorant

Semi-Ignorant

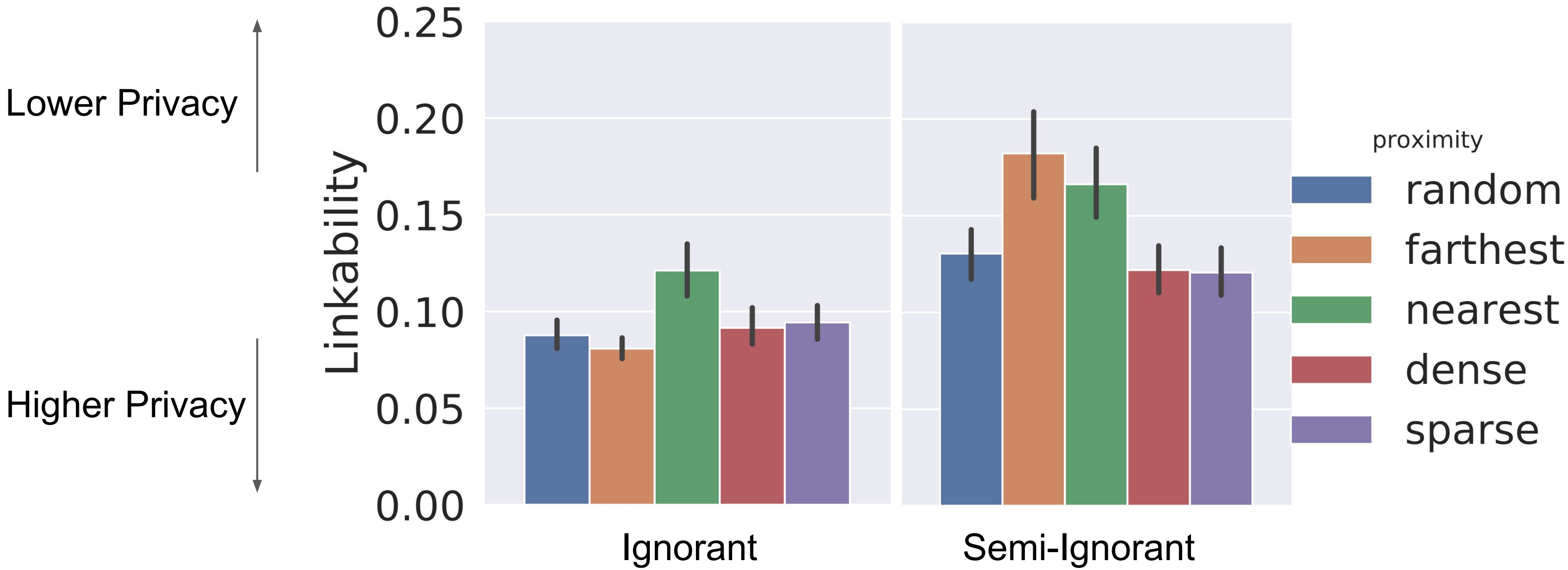
Semi-Informed
(Re-trained ASVeval model)



Utility Evaluation



Proximity (Privacy)



Baseline = **0.86**

Mapping in DENSE region can be considered as "losing your identity in the crowd".

User's perspective

Is the resulting speech corpus suitable for downstream tasks?

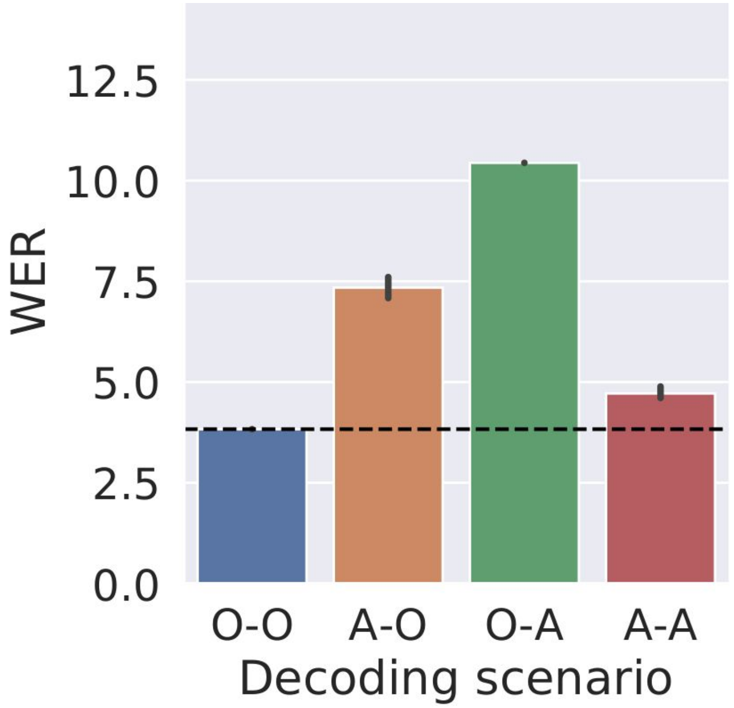
Preserve speech quality in terms of naturalness and intelligibility

- Measured using viability to train ASR models

Informed ASR (Proximity: DENSE)

X-Y = Decoding X using ASR trained on Y

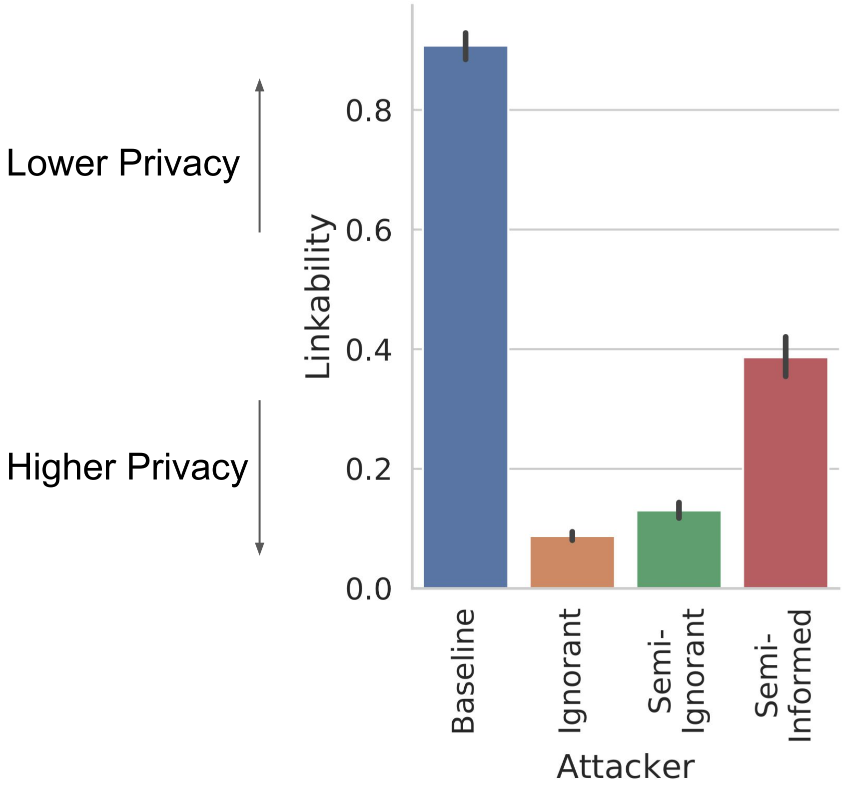
O = Original
A = Anonymized



Attacker’s Questions

1. Does the information about anonymization help discover the speaker’s identity? How to use this information?
2. How to optimize the search space using side-information to efficiently discover the speaker’s identity?

Informed ASV (Proximity: DENSE)



Conclusion

1. Adversarial Training effectively removes speaker's information in a closet-set but does not generalizes to open-set speakers.
2. During Voice Conversion, mapping the "target speaker" in **dense** region with **random** gender selection produces *state-of-the-art* speaker anonymization.
3. The resulting speech corpus can be utilised for tasks such as: training an ASR model.
4. X-vector based target selection proves to be robust against "Semi-Ignorant" and "Semi-Informed" attacks.

Thanks for your attention!

More details on :

https://brijmohan.github.io/

Email : brij.srivastava@inria.fr