Laboratoire Informatique d'Avignon

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

INTERSPEECH 2020
OCTOBER 25-29/ SHANGHAI , CHINA
SHANGHAI INTERNATIONAL CONVENTION CENTER

Design Choices for X-vector Based Speaker Anonymization

Brij Mohan Lal Srivastava, Natalia Tomashenko, Xin Wang, Emmanuel Vincent, Junichi Yamagishi, Mohamed Maouche, Aurélien Bellet, Marc Tommasi

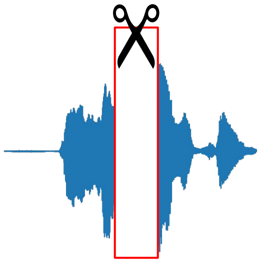
PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR

COMPRISE

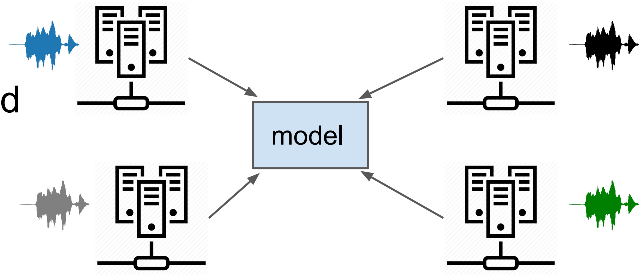
Japan Science and Technology Agency

Methods for privacy protection in speech

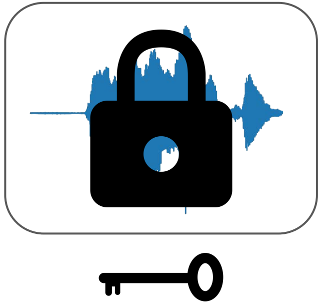
Deletion:



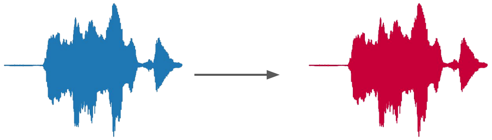
Distributed learning:



Encryption:

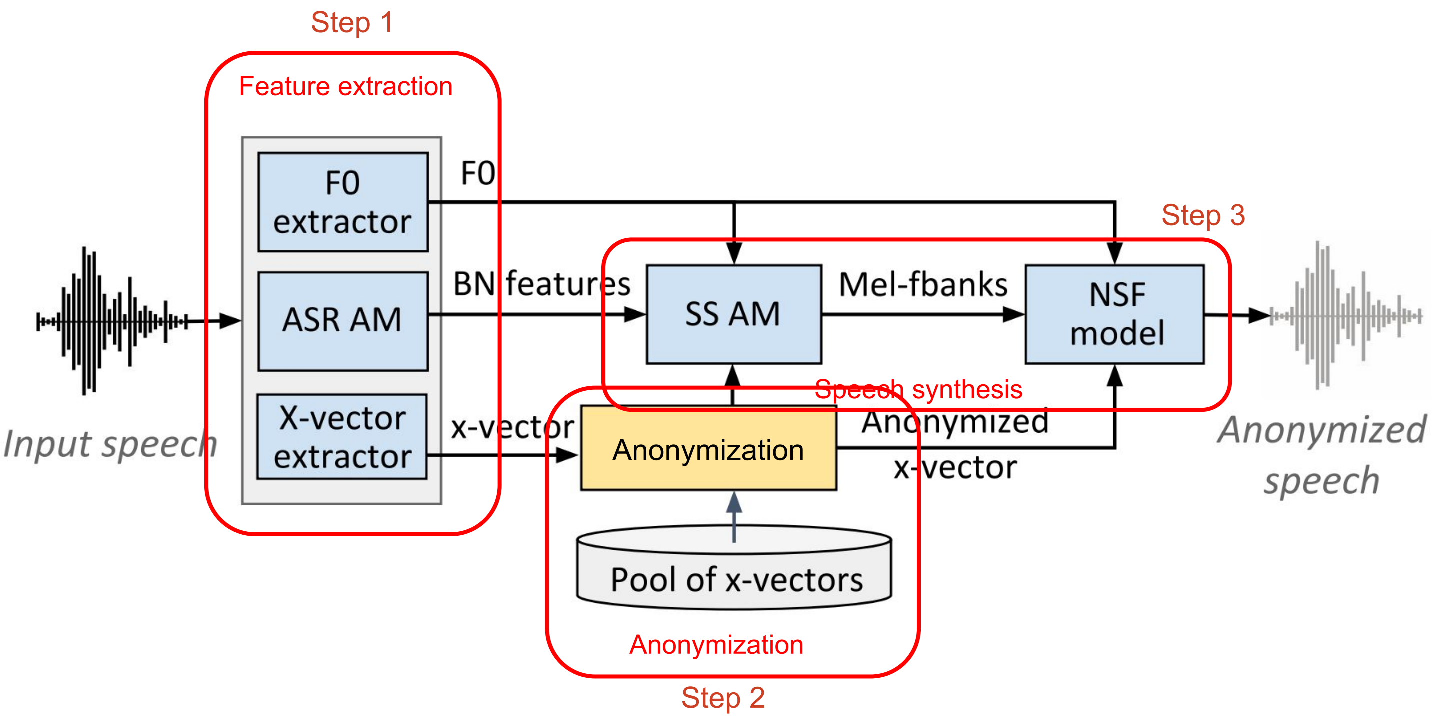


Anonymization:



Suppress speaker’s identity

Anonymization by voice conversion

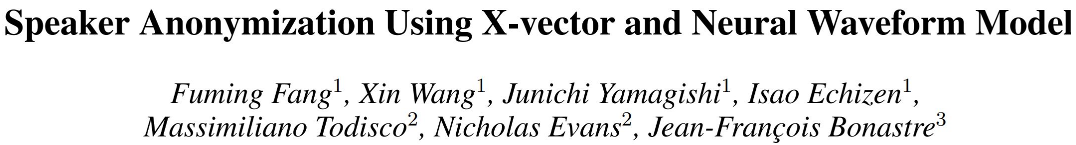


Design choices in speaker anonymization

1. What is the appropriate metric to measure distance between speakers?
2. How to select “**target**” pseudo-speakers from a *small pool of speakers* for robust anonymization?
3. What set of pseudo-speakers will result in high **privacy** protection as well as smaller loss of **utility**?

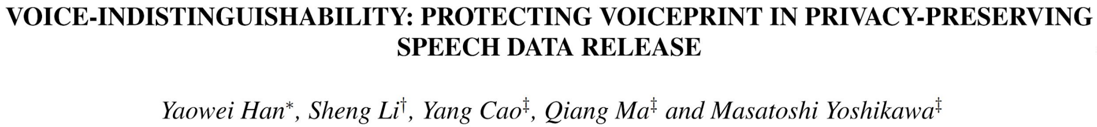
Speaker representation: x-vectors

- Behind the state-of-the-art biometric identification techniques
- Fixed length vector for an utterance regardless of duration (“voiceprint”)
- Intermediate layer of a neural network trained to classify speaker



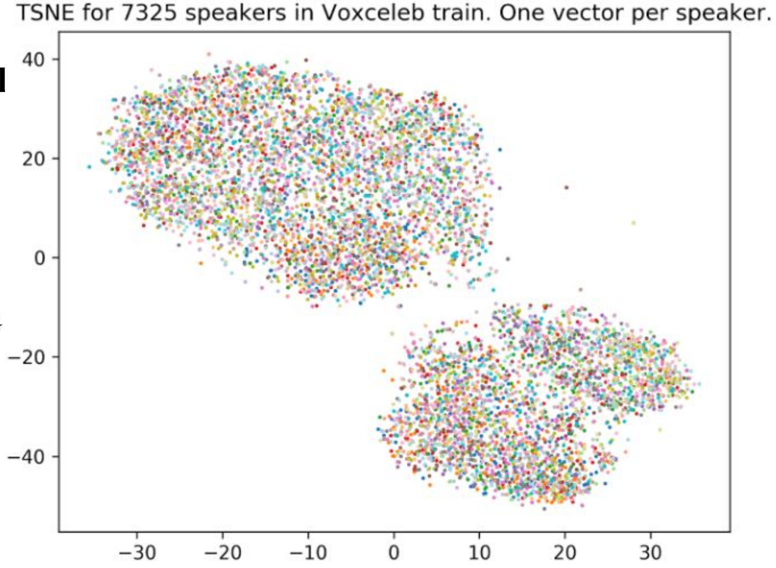
Speaker Anonymization Using X-vector and Neural Waveform Model

Fuming Fang[1], Xin Wang[1], Junichi Yamagishi[1], Isao Echizen[1], Massimiliano Todisco[2], Nicholas Evans[2], Jean-François Bonastre[3]



VOICE-INDISTINGUISHABILITY: PROTECTING VOICEPRINT IN PRIVACY-PRESERVING SPEECH DATA RELEASE

Yaowei Han[*], Sheng Li[†], Yang Cao[‡], Qiang Ma[‡] and Masatoshi Yoshikawa[‡]



TSNE for 7325 speakers in Voxceleb train. One vector per speaker.

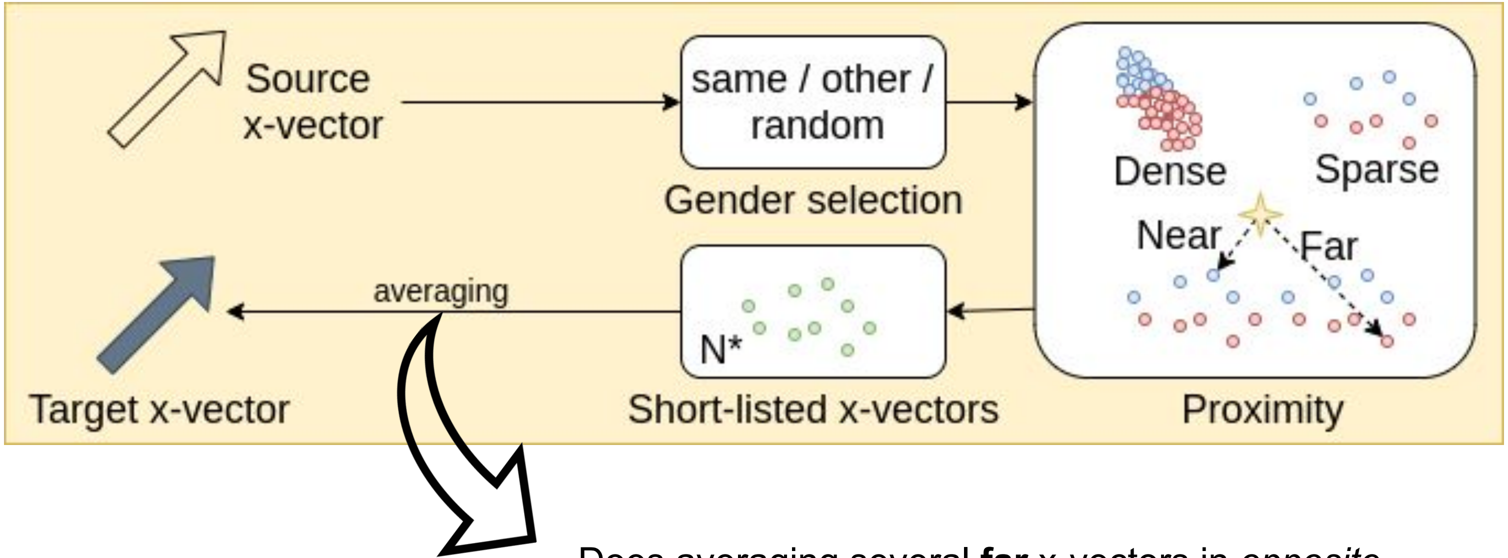
X-vector distance metric

$$\mathrm{cosine}(u, v) = 1 - \frac{u \cdot v}{||u||_2||v||_2}$$

$$\mathrm{PLDA}(u, v) = \log \frac{p(u, v|\mathcal{H}_{\mathrm{same}})}{p(u, v|\mathcal{H}_{\mathrm{different}})}$$

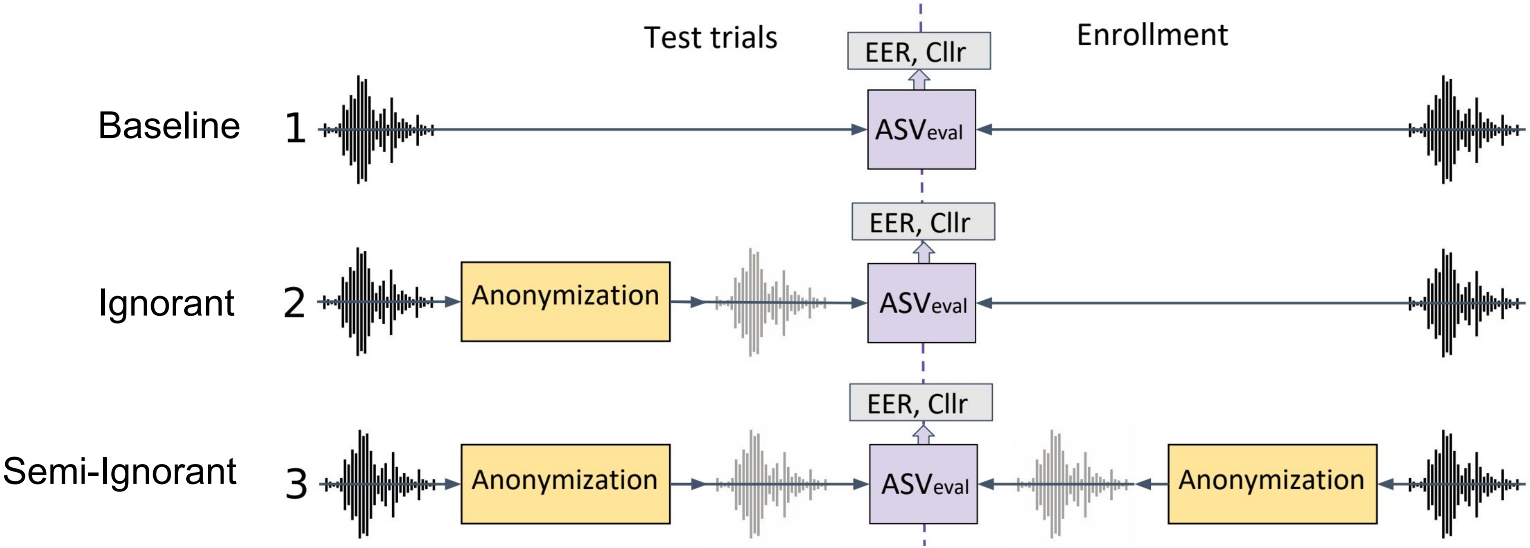
u and v are x-vectors. $\mathcal{H}_{\mathrm{same}}$ and $\mathcal{H}_{\mathrm{different}}$ are the *same-speaker* and *different-speaker* hypotheses respectively.

Target pseudo-speaker selection

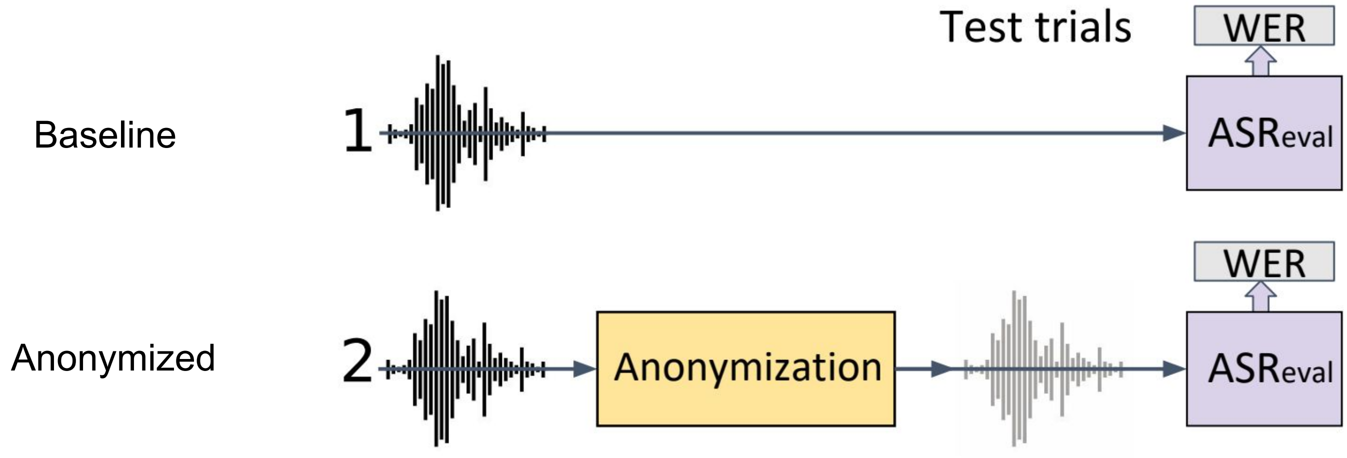


Does averaging several **far** x-vectors in *opposite* directions produce a x-vector close to the source?

Privacy evaluation: Attackers simulated using Automatic Speaker Verification



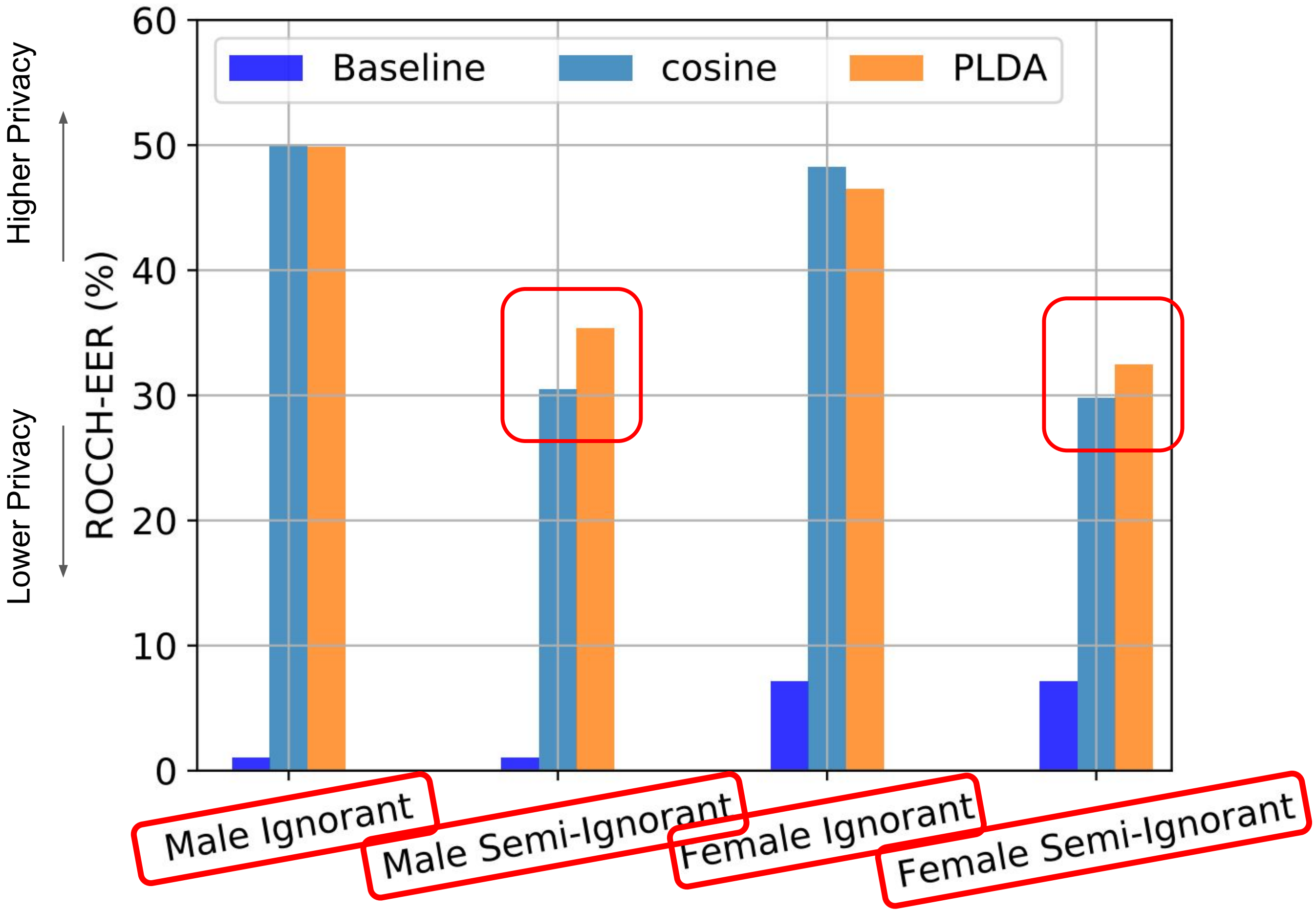
Utility evaluation: Automatic Speech Recognition



Distance

PLDA outperforms cosine distance in x-vector space marginally.

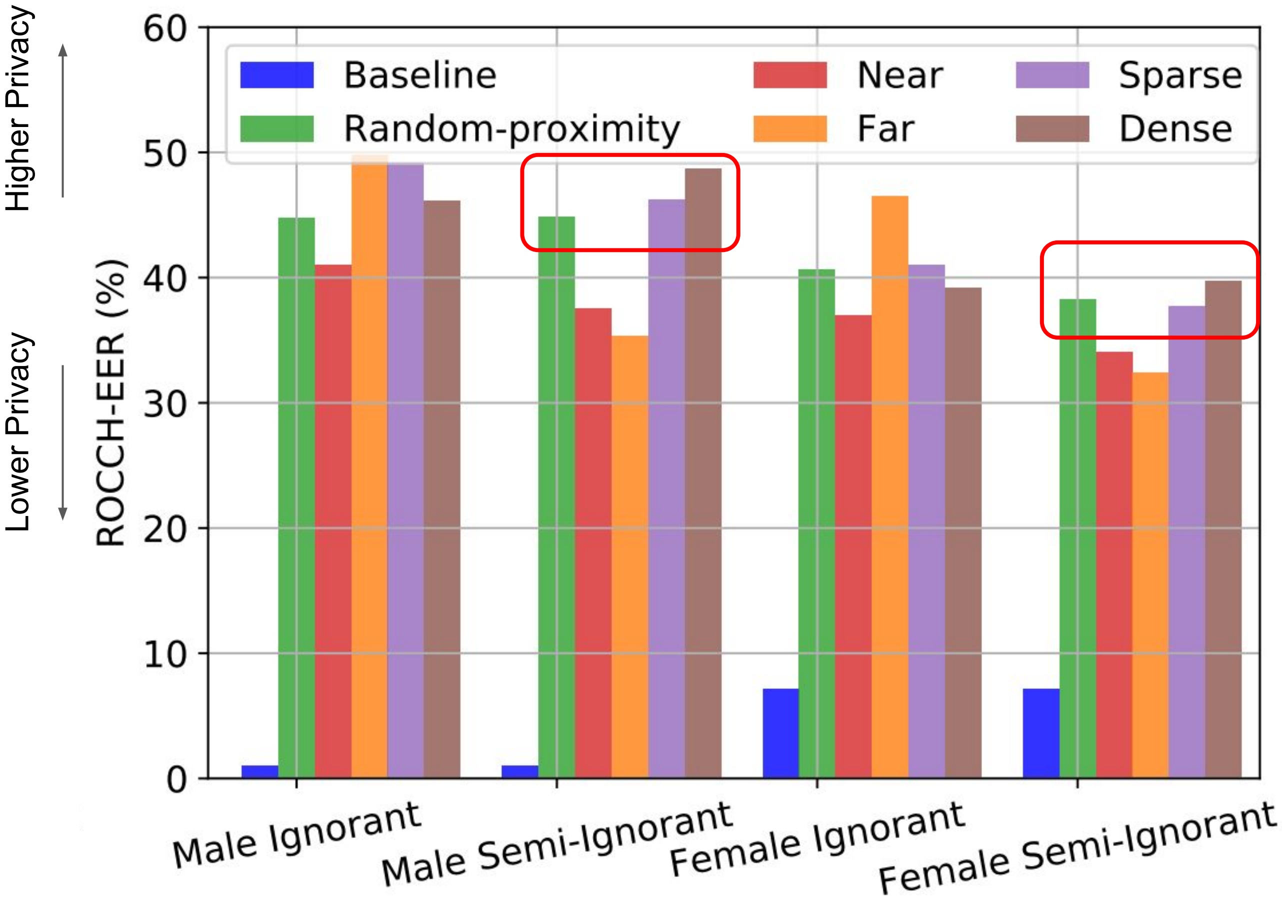
The proximity is fixed to **far** and target gender is **same**.



Proximity

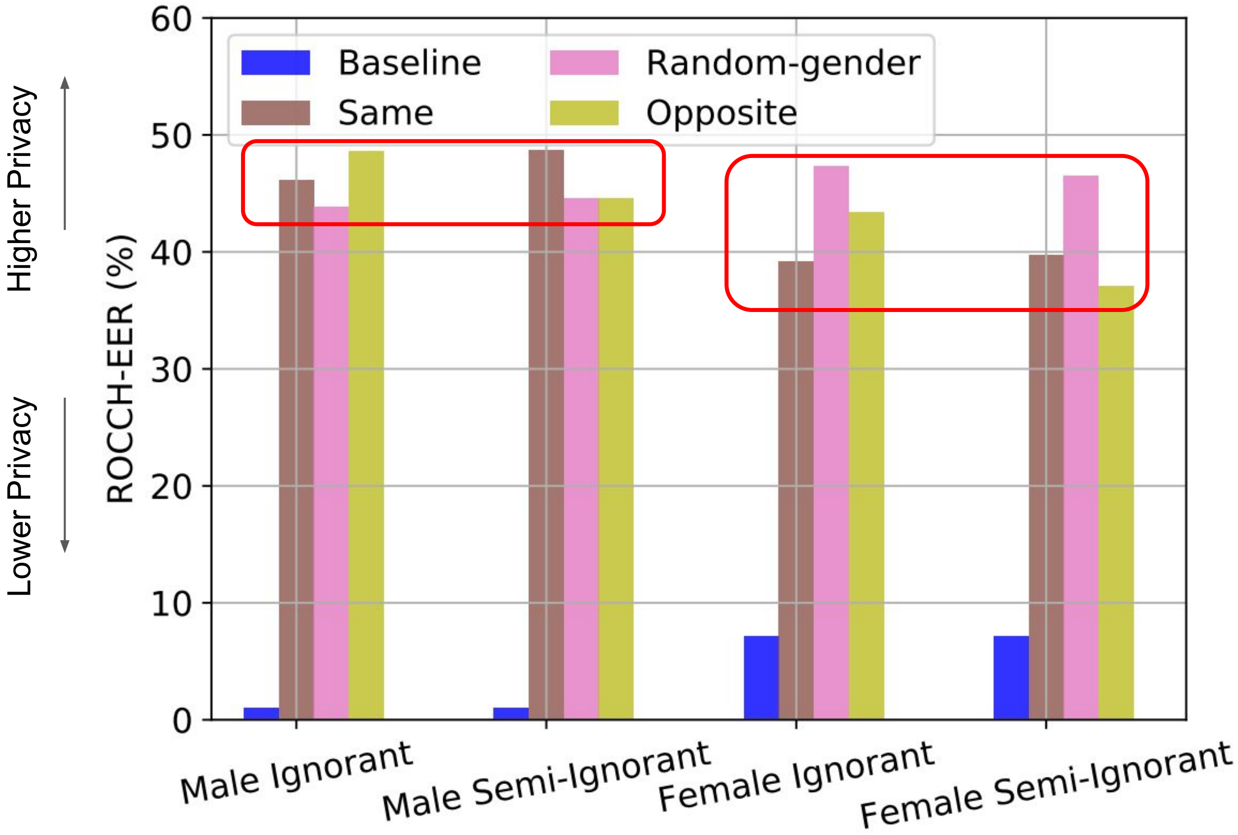
Dense and **Sparse** proximity perform better in semi-ignorant attack resulting in robust anonymization.

Distance is fixed to **PLDA** and target gender is **same**.



Gender selection

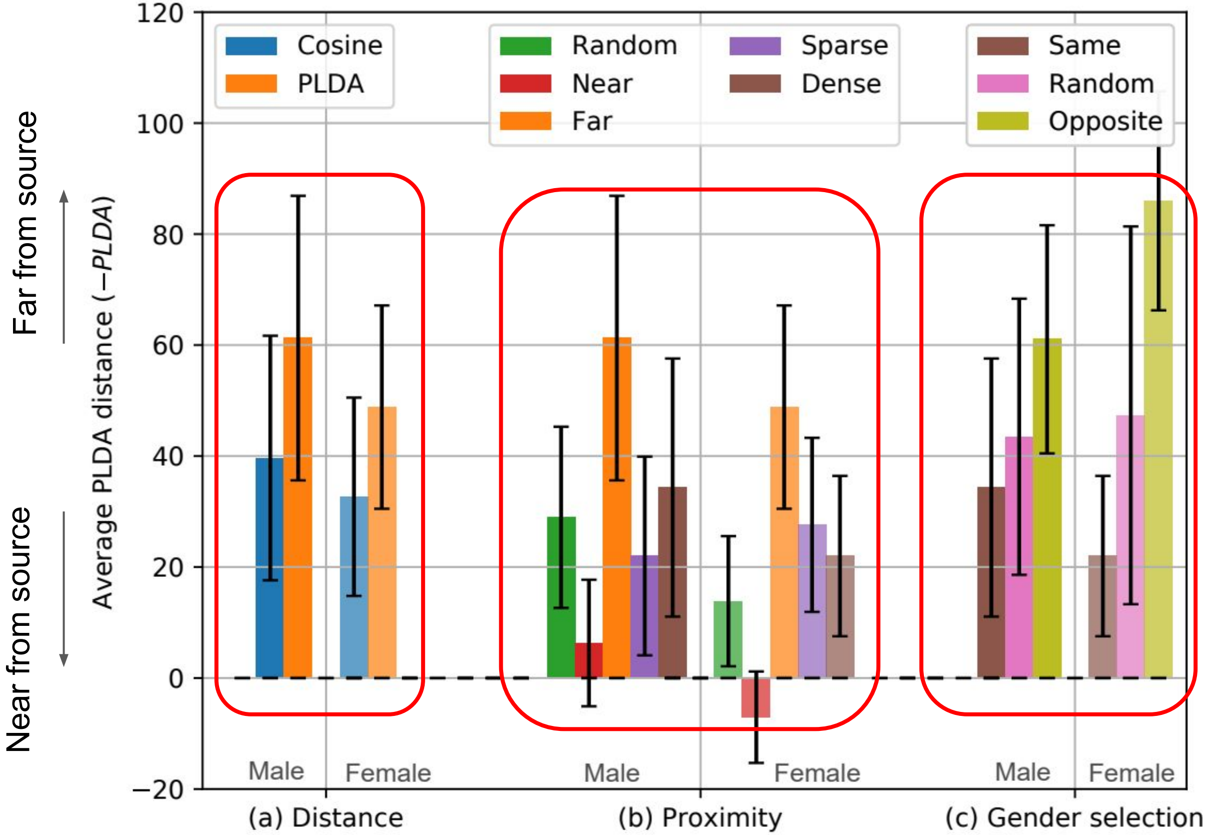
Random target gender produces much stable anonymization across both the gender and both the attackers than using **same** or **opposite**.



Mean PLDA distance

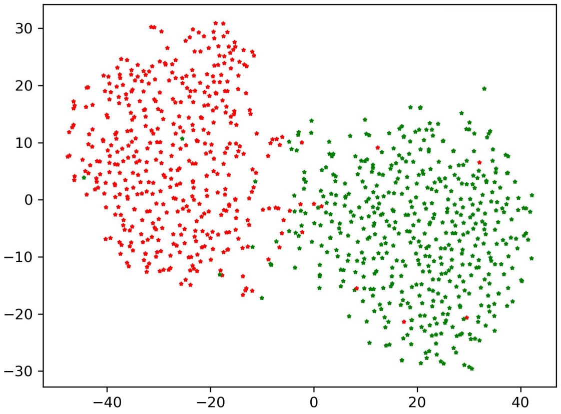
Indeed **Far** proximity exhibits large distance as opposed to **Near**.

Random gender is between Same and Opposite gender.



X-vector space before and after anonymization

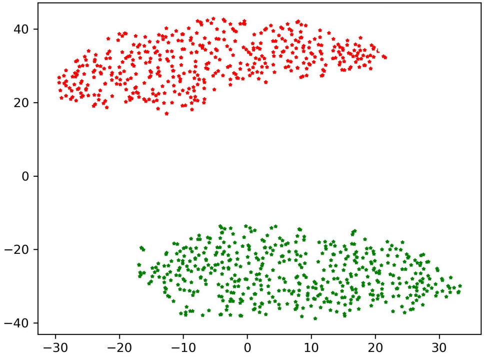
Original



High intra-gender variance
Gender is separable

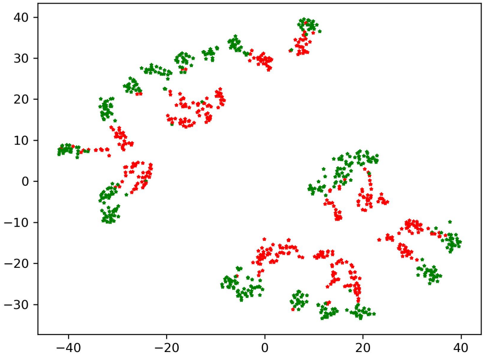
● Male

● Female



Random proximity,
Same gender

More tightly
clustered

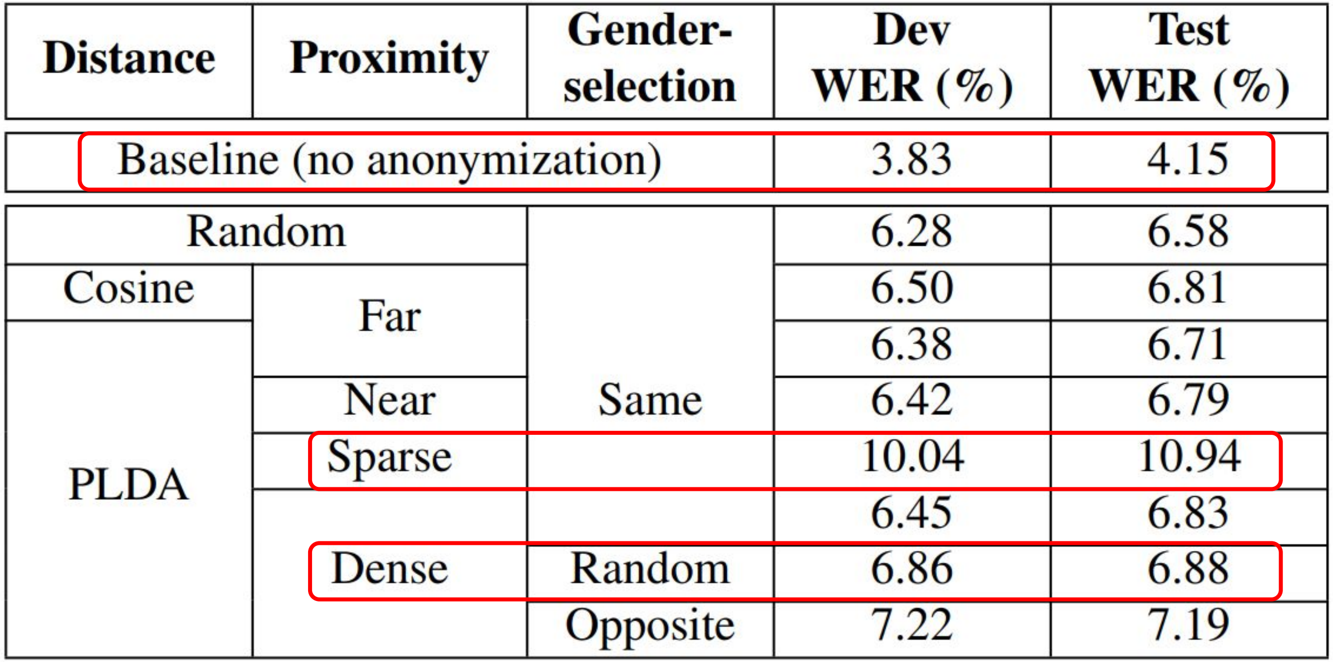


Dense proximity,
Random gender

Gender becomes
inseparable

Word Error Rate

Dense proximity with **Random** gender selection produces reasonable loss of utility as compared to other combinations.



Distance	Proximity	Gender-selection	Dev WER (%)	Test WER (%)
Baseline (no anonymization)			3.83	4.15
Random		Same	6.28	6.58
Cosine	Far		6.50	6.81
PLDA			6.38	6.71
	Near		6.42	6.79
	Sparse		10.04	10.94
	Dense		6.45	6.83
		Random	6.86	6.88
		Opposite	7.22	7.19

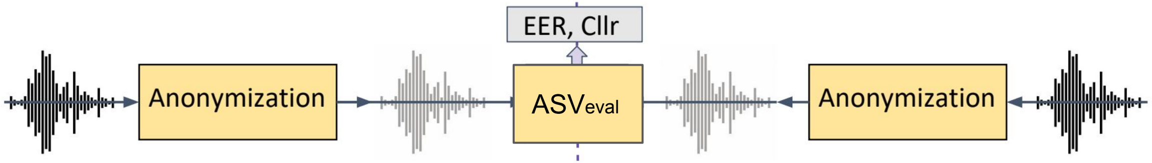
Conclusion

- PLDA distance marginally better than cosine distance in x-vector space.
- Among the different proximity choices, **Dense region** in combination with **Random gender selection** produce reasonable privacy as well as utility.

Future directions

Stronger attacker:

Semi-Informed
(Re-trained ASV$_{eval}$ model)3



1. Is this anonymized data **viable** for ASR training?
2. What is the residual speaker information after anonymization (leakage from BN features and F0)?

Thanks for your attention!

More details on :

https://brijmohan.github.io/

Email : brij.srivastava@inria.fr