INTERSPEECH-2022

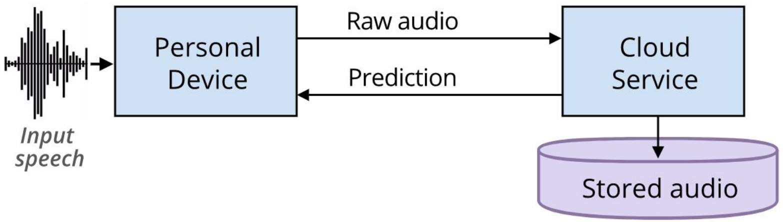
Are disentangled representations all you need to build speaker anonymization systems?

Pierre Champion[1,2], Denis Jouvet[1], Anthony Larcher[2]

[1]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

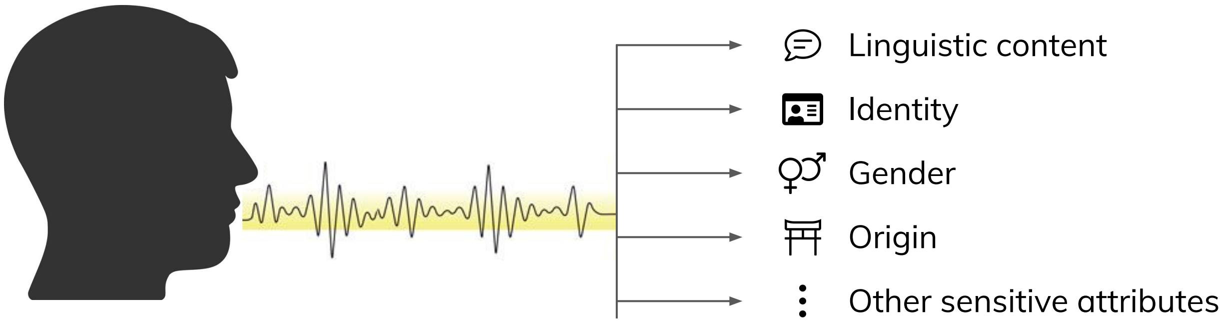
[2]Le Mans Université, LIUM, France

Context: Speech recognition in the cloud

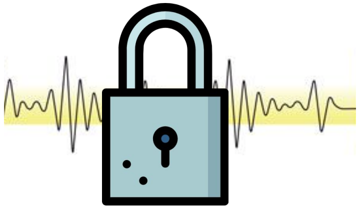


The cloud runs a classifier on the raw audio feature and sends the result back to the user

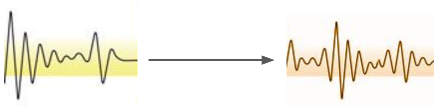
Issue: Privacy



Two general solution:



cryptography



anonymization

Voice Privacy Challenge:

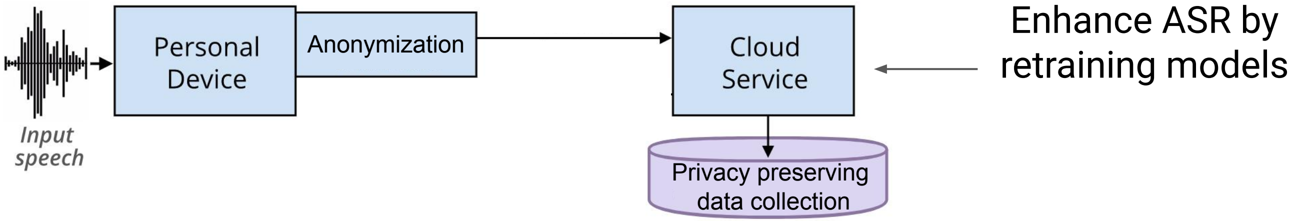
Promote research in privacy related to speech

https://www.voiceprivacychallenge.org/

Outline:

- Introduction
  - Application case
  - Threat model
- Anonymization pipeline
  - Speaker representation
  - Content representation
  - Prosodic representation
- Results
- Conclusions

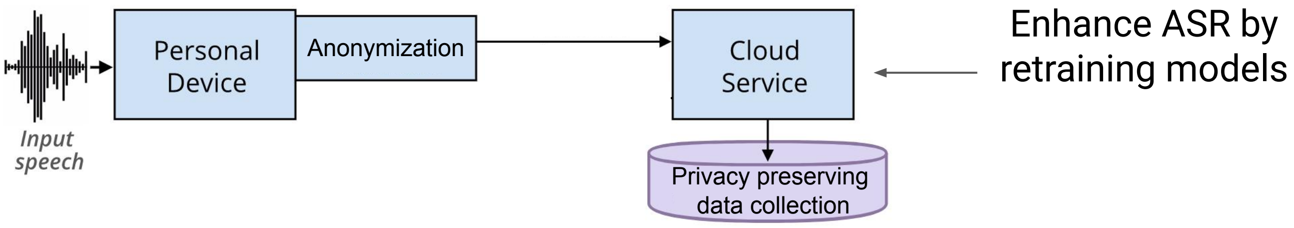
Application context: Share speech data for training new ASR models



Collecting large speech dataset representative of **real users and various usage conditions** is important to improve ASR systems
Must be done while preserving user's privacy => keep the speaker's identity private

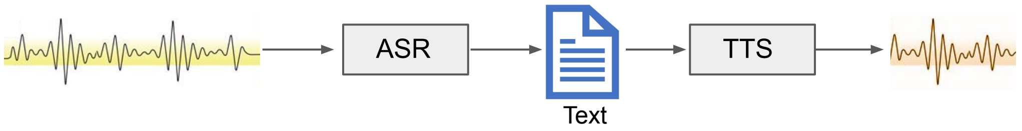
1. Osia, Seyed Ali et al., « A Hybrid Deep Learning Architecture for Privacy-Preserving Mobile Analytics », in : *IEEE Internet of Things Journal* (2020).

Application context: Share speech data for training new ASR models



Collecting large speech dataset representative of **real users and various usage conditions** is important to improve ASR systems
Must be done while preserving user's privacy => keep the speaker's identity private

ASR transcription + TTS is **not** a solution

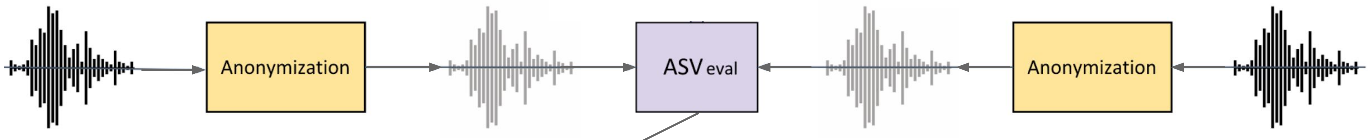


1. Osia, Seyed Ali et al., « A Hybrid Deep Learning Architecture for Privacy-Preserving Mobile Analytics », in : *IEEE Internet of Things Journal* (2020).

Threat model: Linkability of the speaker's speech

ISO/IEC international Standard 24745 on biometric data protection

Speaker Verification Attack



Impostor scores: original speakers are different

Genuine scores: original speakers are the same

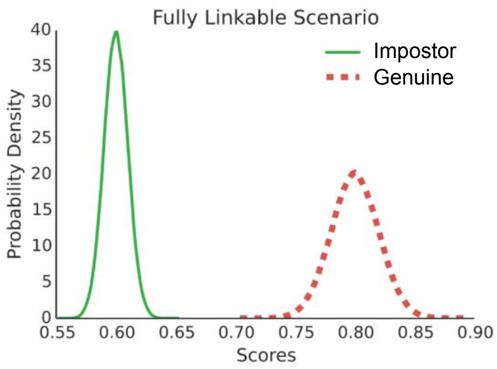
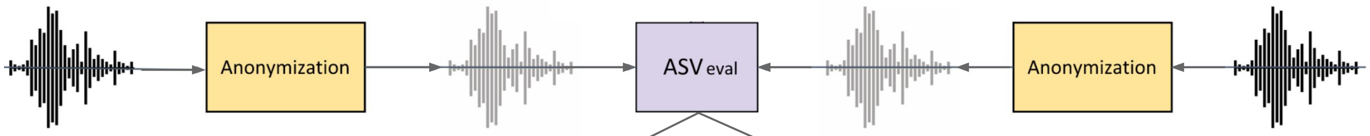


Image from: General Framework to Evaluate Unlinkability in Biometric Template Protection Systems, Gomez-Barrero et al

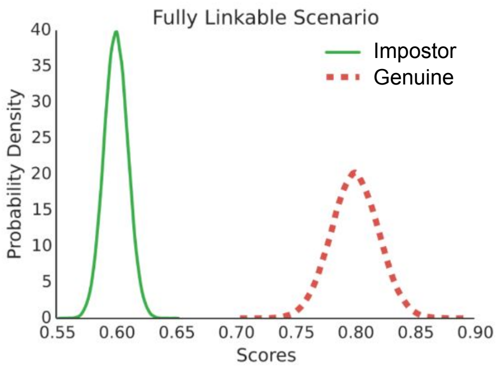
Threat model: Linkability of the speaker's speech
ISO/IEC international Standard 24745 on biometric data protection

Speaker Verification Attack



Impostor scores: original speakers are different

Genuine scores: original speakers are the same



Anonymization

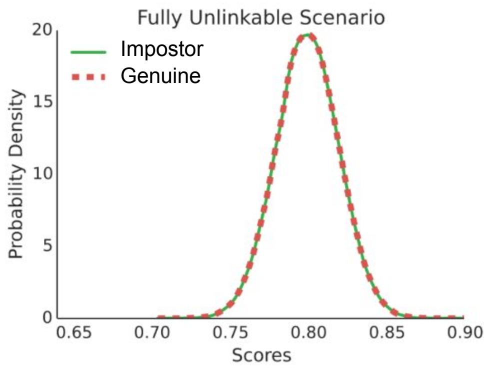


Image from: General Framework to Evaluate Unlinkability in Biometric Template Protection Systems, Gomez-Barrero et al

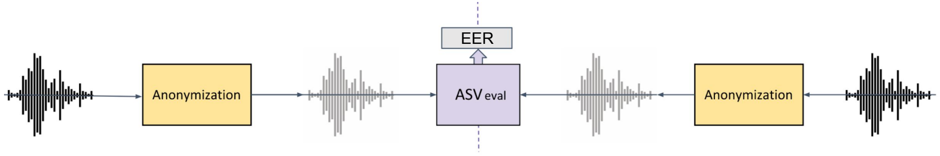
Voice Privacy: Evaluation protocol

Voice privacy challenge 2022 **informed attacker evaluation**

Goal:

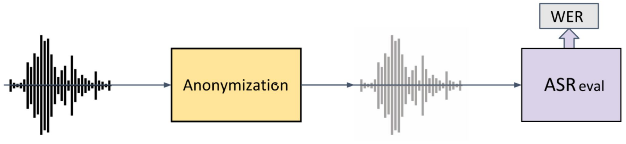
- Privacy: reduce speaker linkability
- Utility: allows the speech to be used for downstream task such as speech recognition

Privacy:



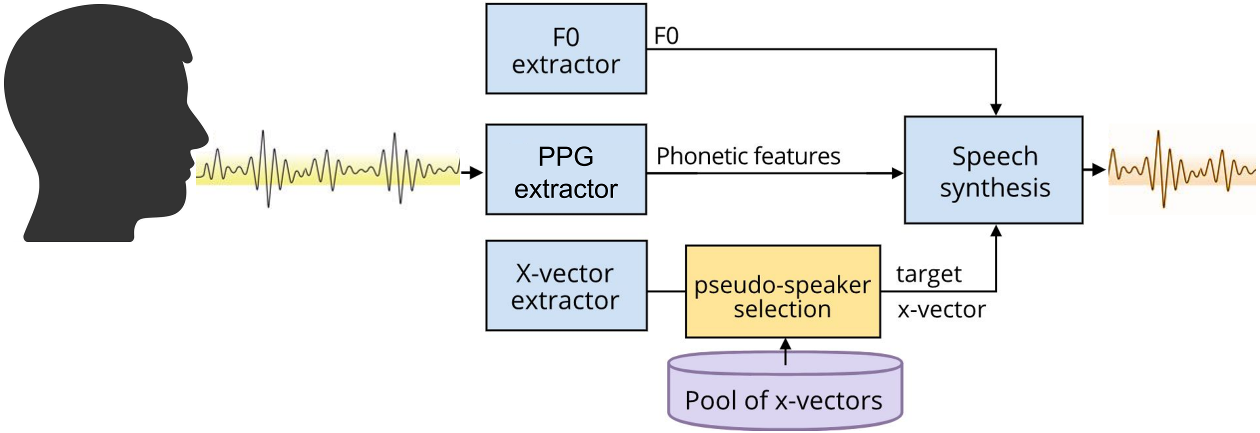
Privacy evaluation using
Automatic Speaker Verification
Metric: **EER** (maximize)

Utility:



Utility evaluation using
Automatic Speech Recognition
Metric: **WER** (minimize)

Voice Privacy: Speaker anonymization framework

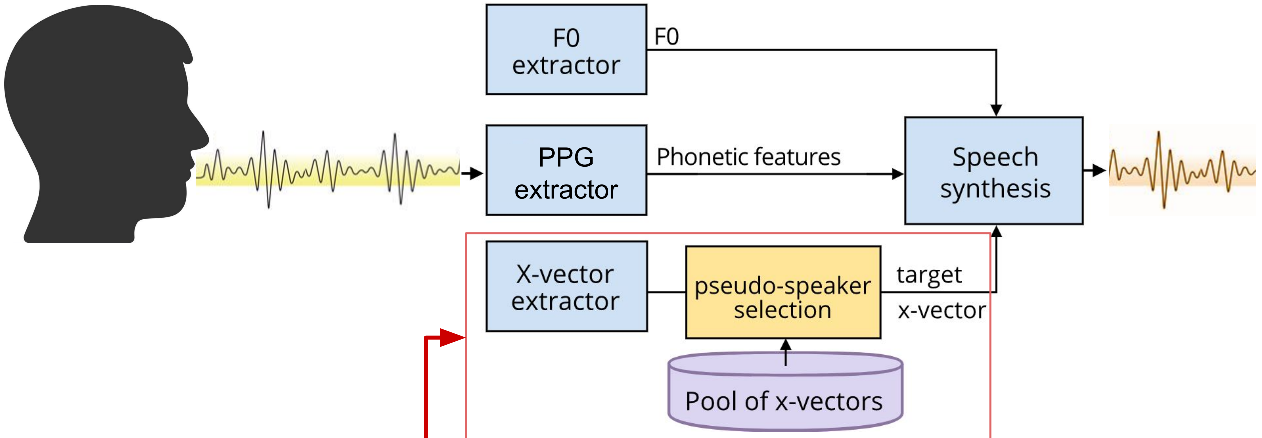


The baseline speaker anonymization system

Speaker representation and modification

Speaker representation and modification

Baseline voice privacy

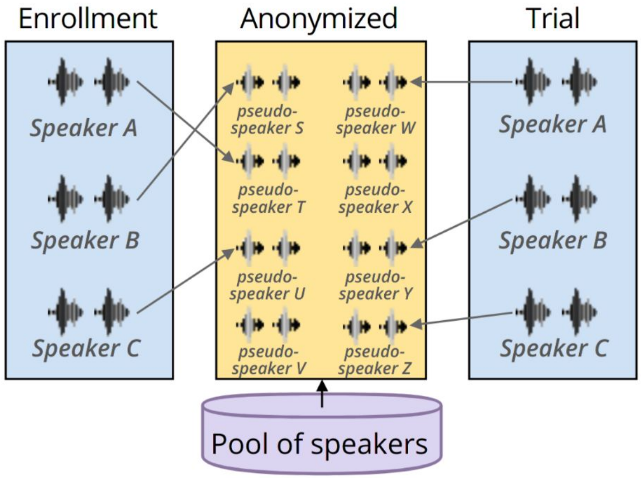


The baseline speaker anonymization system

Performs the speaker modification using x-vector

1. FANG, Fuming et al., « Speaker Anonymization Using X-vector and Neural Waveform Models », in : *10th ISCA Speech Synthesis Workshop*, 2019.
2. SRIVASTAVA, Brij Mohan Lal et al., « Design Choices for X-vector Based Speaker Anonymization », in : *Interspeech* (2020).

Baseline Voice Privacy: Speaker modification

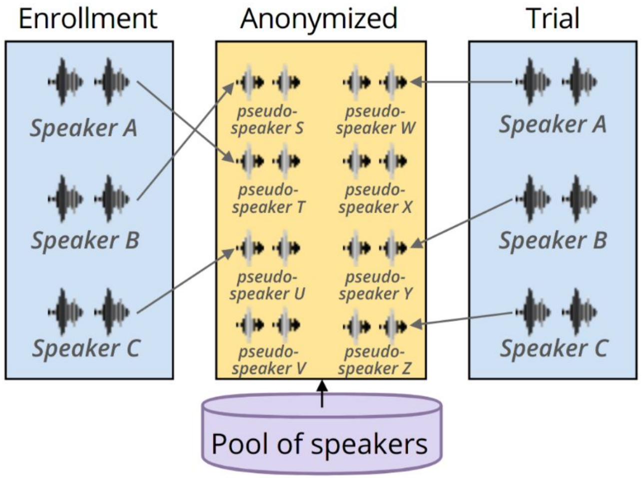


[1,2] Pseudo random target speaker selection

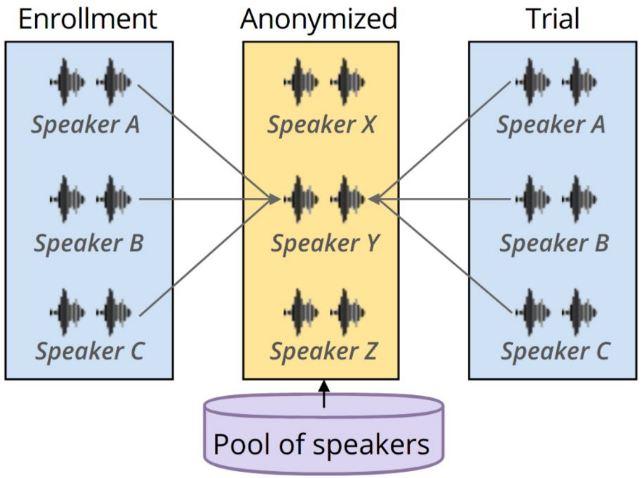
Any to Any voice conversion

1. FANG, Fuming et al., « Speaker Anonymization Using X-vector and Neural Waveform Models », in : *10th ISCA Speech Synthesis Workshop*, 2019.
2. SRIVASTAVA, Brij Mohan Lal et al., « Design Choices for X-vector Based Speaker Anonymization », in : *Interspeech* (2020).

Voice Privacy: Speaker modification



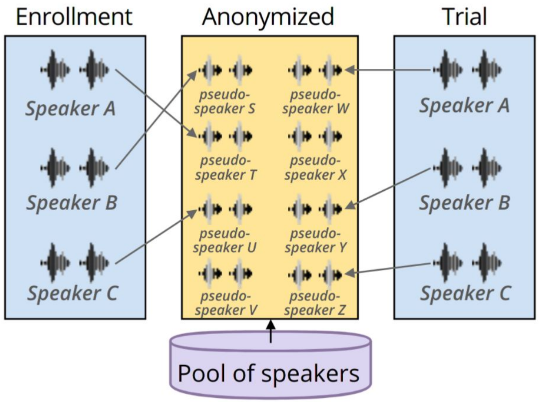
[1,2] Pseudo random target speaker selection
Any to Any voice conversion



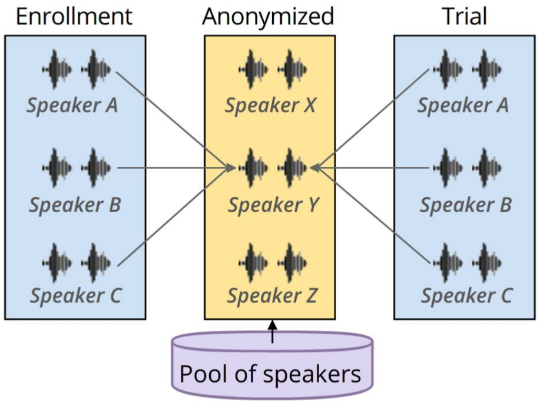
[3] Unique speaker target selection
Any to One voice conversion

1. FANG, Fuming et al., « Speaker Anonymization Using X-vector and Neural Waveform Models », in : *10th ISCA Speech Synthesis Workshop*, 2019.
2. SRIVASTAVA, Brij Mohan Lal et al., « Design Choices for X-vector Based Speaker Anonymization », in : *Interspeech* (2020).
3. CHAMPION, Pierre, Denis JOUVET et Anthony LARCHER, « Evaluating X-vector-based Speaker Anonymization under White-box Assessment », in : *SPECOM*, 2021.

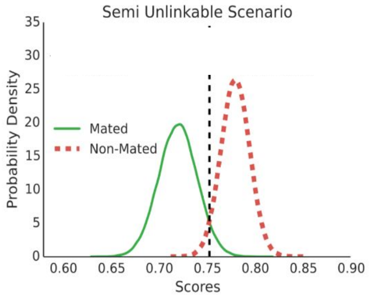
Voice Privacy: Speaker modification



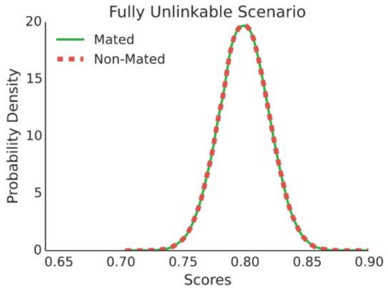
Pseudo random target speaker selection
Any to Any voice conversion



Unique speaker target selection
Any to One voice conversion



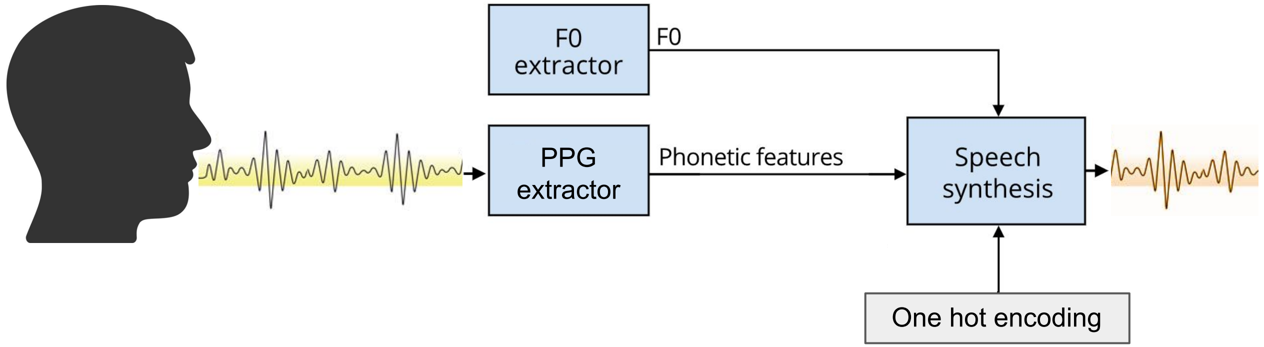
- Not all target selection algorithm fulfill unlinkability
- Overestimation of privacy protection are more susceptible due to weak attacker



- Fulfill unlinkability
- Symplicity
- Better guarantee to train a powerful attacker

Speaker representation and modification

Proposed



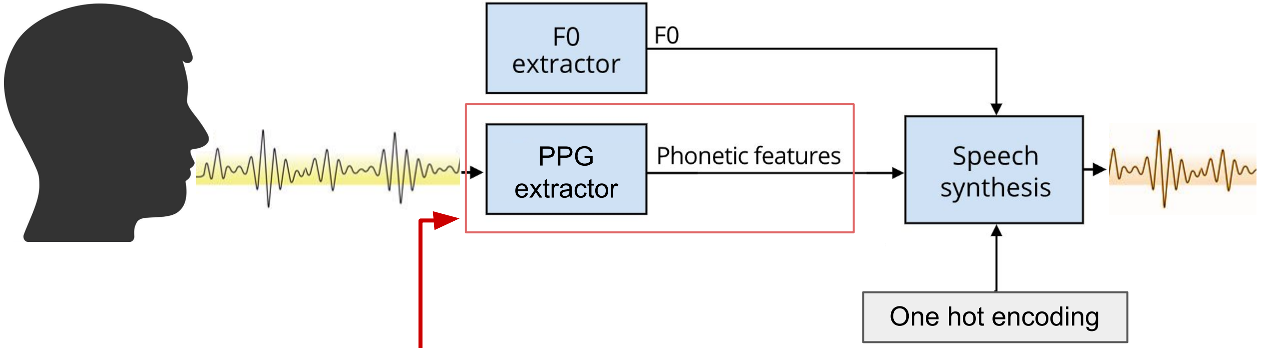

```
array([0, 1, 0, 0])
```

Performs the speaker modification using a **one hot encoding of a single speaker identity**

1. Pipeline simplification
2. Unlinkability guarantee
3. No overestimation of privacy protection

Phonetic PosteriorGrams

Phonetic PosteriorGrams representation



Encodes:
- Speech sounds corresponding to the content
- Speaker Identity

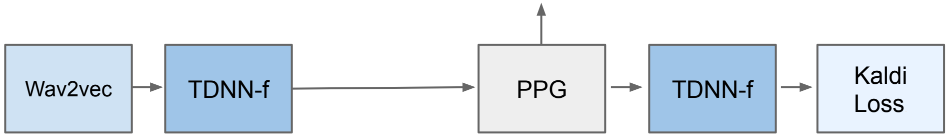
1. ADI, Y. et al., « To Reverse the Gradient or Not : an Empirical Comparison of Adversarial and Multi-task Learning in Speech Recognition », in : *IEEE ICASSP*, 2019.
2. CHAMPION, Pierre, Denis JOUVET et Anthony LARCHER, « Privacy-Preserving Speech Representation Learning using Vector Quantization », in : *Journées d'Études sur la Parole (JEP, 34e édition)*, 2022.
3. SHAMSABADI, Ali Shahin et al., « Differentially Private Speaker Anonymization », in : *arXiv* (2022).

PPG extractor: acoustic model

Explored with multiple acoustic model:

- **Wav2vec 2.0** pre-trained with VoxPopuli
  Wav2vec 2.0-TDNN-f trained with
  librispeech train-100

=> extract **continuous** PPG

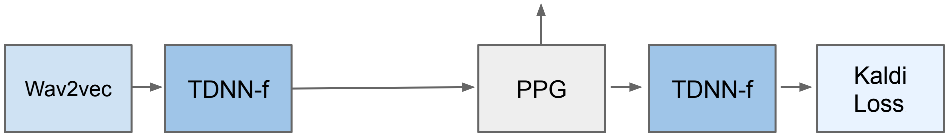


PPG extractor: acoustic model

Explored with multiple acoustic model:

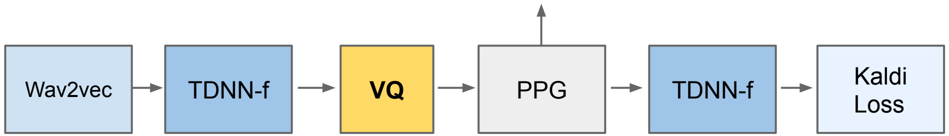
- **Wav2vec 2.0** pre-trained with VoxPopuli
  Wav2vec 2.0-TDNN-f trained with
  librispeech train-100

  => extract **continuous** PPG



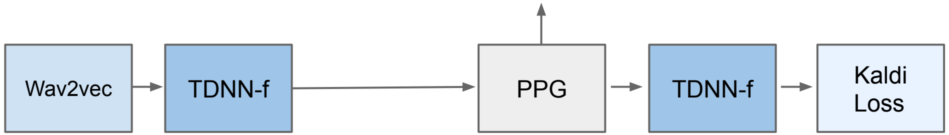
- **Wav2vec 2.0-TDNN-f** trained with
  librispeech train-100
  \+ **vector quantization layer**

  => extract **discrete** PPG



Baseline performance

Dataset	LibriSpeech test-clean		VCTK test	
Method	Privacy EER% ↑	Utility WER% ↓	Privacy EER% ↑	Utility WER% ↓
Clean speech	4.1	4.1	2.7	12.8
Anonymized (Wav2Vec 2.0 - No VQ)	↑ 7.7	3.8 ↓	↑ 12.1	7.8 ↓



Perfect privacy protection:
50% EER
0% WER

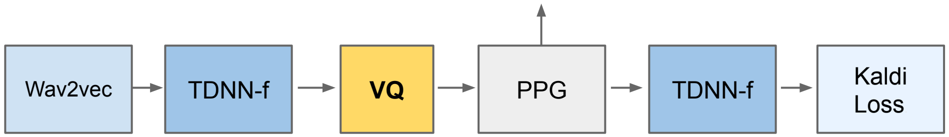
Wav2vec 2.0-TDNN-f

Small privacy improvement for both datasets
Utility improvement for both datasets

Speaker leakage occurs in the pipeline as the EER are still very low

Vector quantized PPG performance

Dataset	LibriSpeech test-clean		VCTK test	
Method	Privacy EER% ↑	Utility WER% ↓	Privacy EER% ↑	Utility WER% ↓
Clean speech	4.1	4.1	2.7	12.8
Anonymized (Wav2Vec 2.0 - No VQ)	↑ 7.7	3.8 ↓	↑ 12.1	7.8 ↓
Anonymized (Wav2Vec 2.0 - VQ 48)	↑ 17.5	4.5		↑ 28.0	10.0 ↓



Perfect privacy protection:
50% EER
0% WER

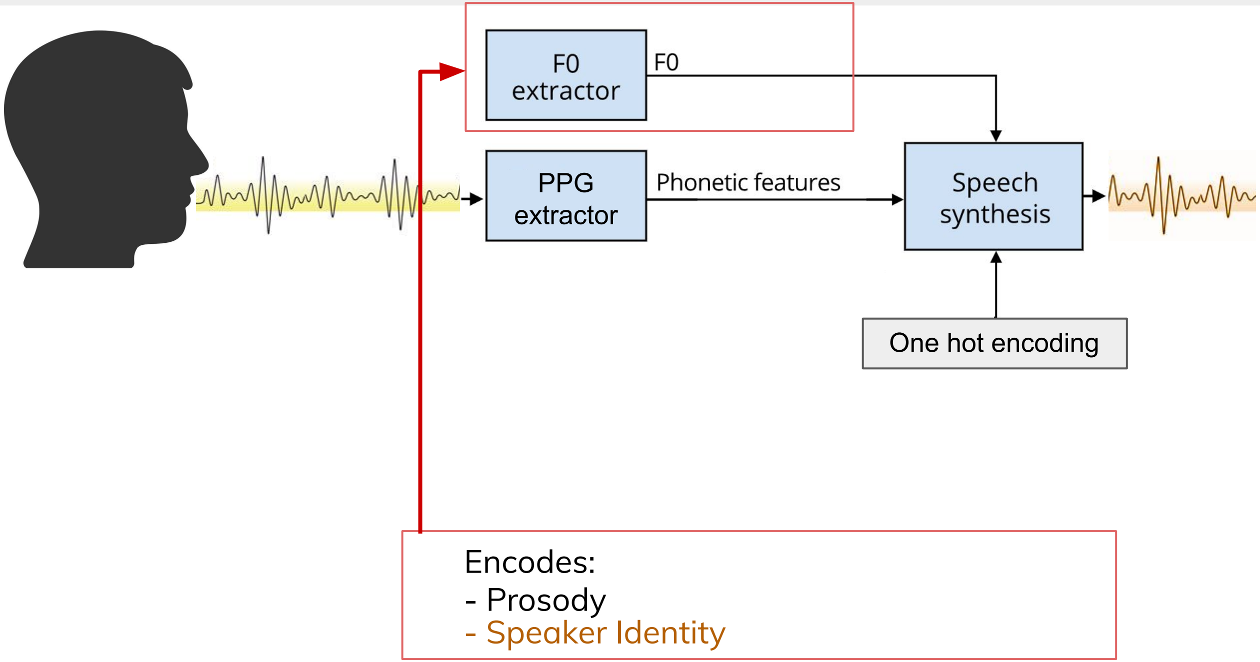
Wav2vec 2.0-TDNN-f VQ

Higher privacy improvement for both datasets
Same utility improvement for both datasets

Vector quantization applies a constraint on the PPG representation space, making them more private without losing too much utility

Fundamental frequency

Fundamental frequency modification



1. CHAMPION, Pierre, Denis JOUVET et Anthony LARCHER, « A Study of F0 Modification for X-Vector Based Speech Pseudonymization Across Gender », in : *The Second AAAI Workshop on Privacy-Preserving Artificial Intelligence*, 2021.
2. GAZNEPOGLU, Ünal Ege et Nils PETERS, « Exploring the Importance of F0 Trajectories for Speaker Anonymization using X-vectors and Neural Waveform Models », in : *Workshop on Machine Learning in Speech and Language Processing*, 2021.
3. SHAMSABADI, Ali Shahin et al., « Differentially Private Speaker Anonymization », in : *arXiv* (2022).

F0 modified + Vector quantized PPG performance

Dataset	LibriSpeech test-clean		VCTK test	
Method	Privacy EER% ↑	Utility WER% ↓	Privacy EER% ↑	Utility WER% ↓
Clean speech	4.1	4.1	2.7	12.8
Anonymized	↑ 7.7	3.8 ↓	↑ 12.1	7.8 ↓
Anonymized VQ 48	↑ 17.5	4.5	↑ 28.0	10.0 ↓
Anonymized VQ 48 + F_0 NOISE	↑ 23.4	4.6	↑ 40.8	10.3 ↓

Perfect privacy protection:
50% EER
0% WER

F0 noise + Wav2vec 2.0-TDNN-f VQ

Highest privacy improvement for both datasets
Same utility improvement for both datasets

Adding White Gaussian noise to the F0 trajectory allows to hide the speaker information that it contained

F0 modified + Vector quantized PPG performance

Dataset	LibriSpeech test-clean		VCTK test	
Method	Privacy EER% ↑	Utility WER% ↓	Privacy EER% ↑	Utility WER% ↓
Clean speech	4.1	4.1	2.7	12.8
Anonymized	↑ 7.7	3.8 ↓	↑ 12.1	7.8 ↓
Anonymized VQ 48	↑ 17.5	4.5	↑ 28.0	10.0 ↓
Anonymized VQ 48 + F_0 NOISE	↑ 23.4	4.6	↑ 40.8	10.3 ↓
Anonymized VPC 2022 baseline	13.5	5.1	20.6	13.0

Perfect privacy protection:
50% EER
0% WER

Significantly better than the VPC 2022 baseline

Conclusion

Q: Are disentangled representations all you need to build speaker anonymization systems?
A: Yes, but how?

- One hot encoding is all we need, targeting a single identity => simplifying the pipeline with guarantee
- Vector quantized PPG has some limitation => Can we annotate the anonymized speech to retrain ASR system?
- F0 modification with noise has intelligibility limitation

Thank for your attention
Email: pierre.champion@inria.fr

Live Demo / Shared Models / Code Source at:
https://colab.research.google.com/github/deep-privacy/SA-toolkit/blob/master/SA-colab.ipynb

Wav2vec 2.0 Vector quantized PPG performance

Dataset	LibriSpeech test-clean		VCTK test	
Method	Privacy EER% ↑	Utility WER% ↓	Privacy EER% ↑	Utility WER% ↓
Clean speech	4.1	4.1	2.7	12.8
Ours TDNNF NO VQ	↓ 8.7	6.9 ↑	↓ 10.8	19.1 ↑
Ours TDNNF VQ 256	↓ 16.2	9.9 ↑	↓ 22.9	24.1 ↑
Ours TDNNF VQ 128	↓ 17.7	10.4 ↑	↓ 24.0	26.3 ↑
Ours TDNNF VQ 64	↓ 21.1	12.4 ↑	↓ 30.0	29.1 ↑
Ours WAV2VEC2 TDNNF NO VQ	↓ 7.7	3.8 ↓	↓ 12.1	7.8 ↓
Ours WAV2VEC2 TDNNF VQ 48	↓ 17.5	4.5		↓ 28.0	10.0 ↓

Great privacy improvement for both datasets
And better utility for both datasets

Perfect privacy protection:
50% EER
0% WER

With the correct architecture, depth and amount of unsupervised training data,
vector quantization can apply a high constraint on PPG, making them more private without losing utility

F0 modified + Wav2vec 2.0 Vector quantized PPG performance

Dataset	LibriSpeech test-clean		VCTK test	
Method	Privacy EER% ↑	Utility WER% ↓	Privacy EER% ↑	Utility WER% ↓
Clean speech	4.1	4.1	2.7	12.8
Ours TDNNF NO VQ	8.7	6.9	10.8	19.1
Ours TDNNF VQ 256	16.2	9.9	22.9	24.1
Ours TDNNF VQ 128	17.7	10.4	24.0	26.3
Ours TDNNF VQ 64	21.1	12.4	30.0	29.1
Ours WAV2VEC2 TDNNF NO VQ	7.7	3.8	12.1	7.8
Ours WAV2VEC2 TDNNF VQ 48	17.5	4.5	28.0	10.0
Ours WAV2VEC2 TDNNF VQ 48 + F_0 $AWGN_{15dB}$	23.4	4.6	40.8	10.3
VPC 2022 baseline	13.5	5.1	20.6	13.0

Perfect privacy protection:
50% EER
0% WER

Significantly better than the VPC 2022 baseline