PhD-Defense

Anonymizing Speech: Evaluating and Designing Speaker Anonymization Techniques

Pierre Champion

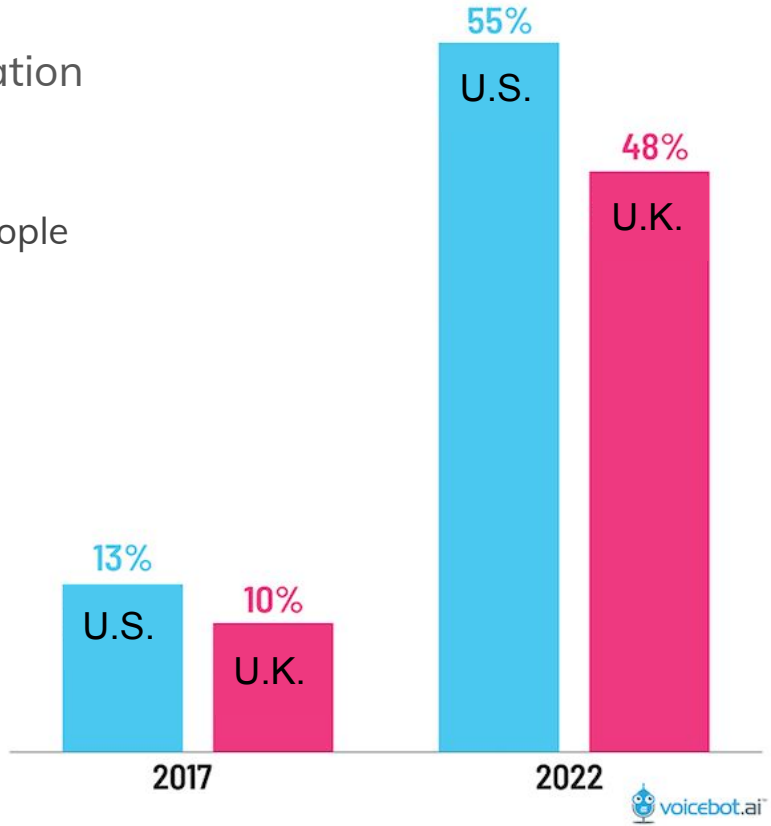
Supervised by:
Dr. Denis Jouvet (Nancy)
Prof. Anthony Larcher (Le Mans)
Prof. Slim Ouni (Nancy)

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
Le Mans Université, LIUM, France

Context

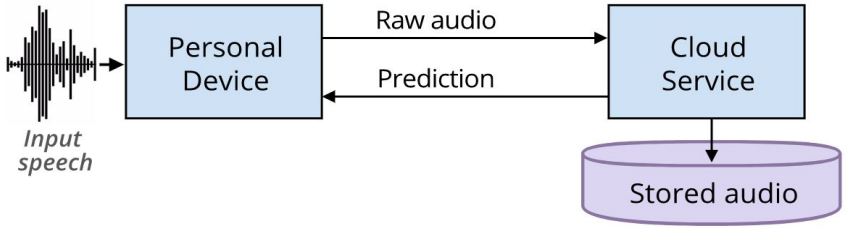
Speech is a natural form of communication

Human-computer interfaces make
the use of technology more natural for people



Voice assistant adoption in U.S. & U.K.

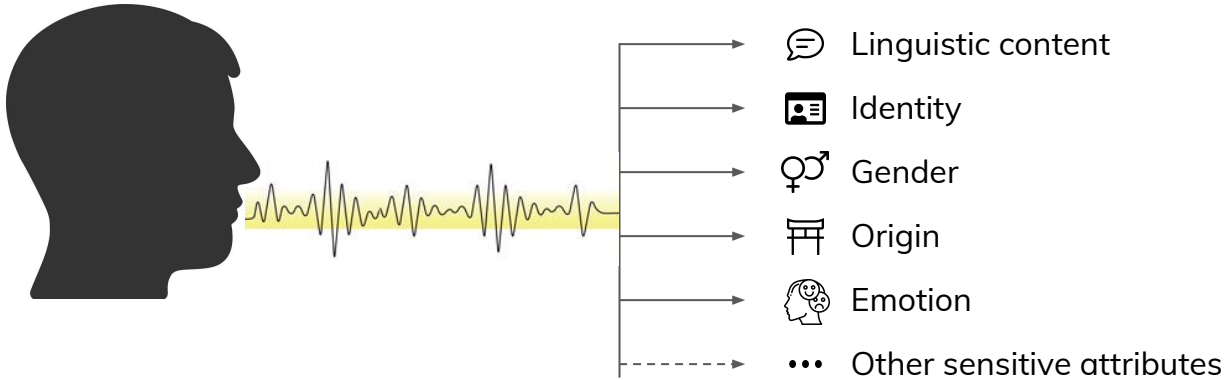
Context: Speech recognition in the cloud



In the cloud, a classifier uses raw audio to make predictions.

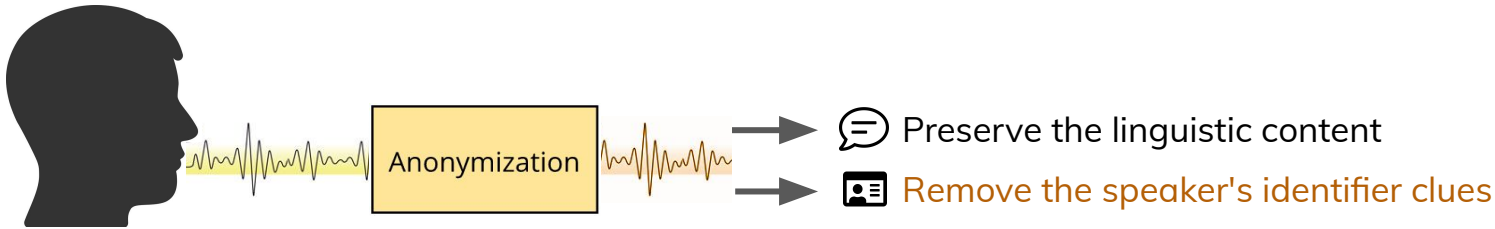
- Produce a lot of data
- Stored in centralized servers

Issue: Privacy



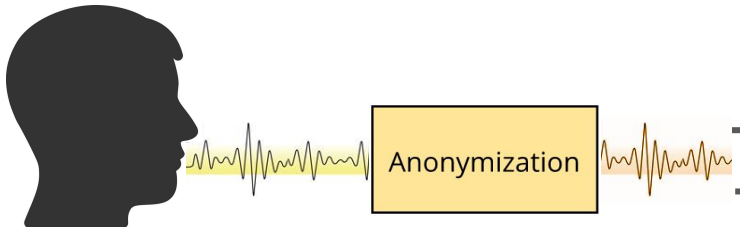
Subject of this Thesis

How to create and evaluate anonymization systems that removes the speaker identifier clues while preserving the linguistic content?



Subject of this Thesis

How to create and evaluate anonymization systems that removes the speaker identifier clues while preserving the linguistic content?



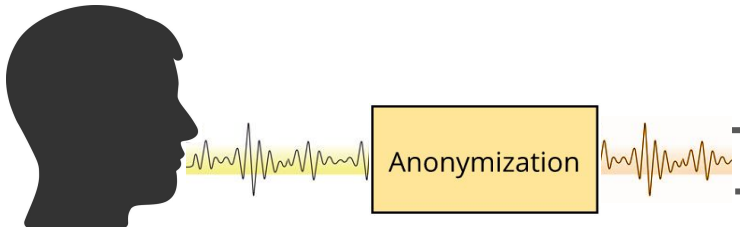
Preserve the linguistic content

Remove the speaker's identifier clues

Preserve or remove intonation and other attributes, depending on application and user choices

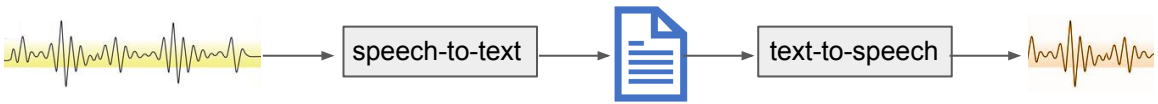
Subject of this Thesis

How to create and evaluate anonymization systems that removes the speaker identifier clues while preserving the linguistic content?



- Preserve the linguistic content
- Remove the speaker's identifier clues
- ... Preserve or remove intonation and other attributes, depending on application and user choices

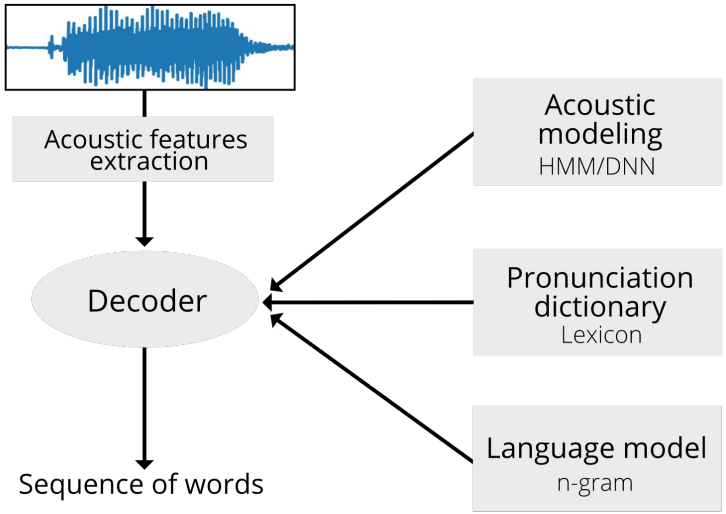
Transcription + synthesis is not the most broadly applicable solution



Outline:

- **Background**
  - Automatic speech recognition
  - Automatic speaker verification
  - Voice conversion
  - Threat model
  - Evaluation methods
- Impact of the target speaker parameter in VC anonymization
- Anonymization with feature-level disentanglement
- Conclusion and perspectives

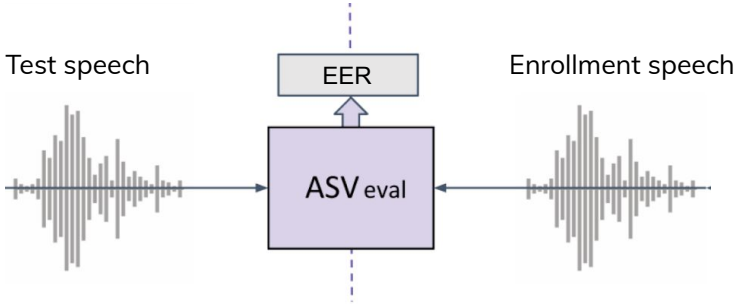
Automatic speech recognition (ASR)



ASR is used for:

- Automatic objective evaluation of the preservation of the linguistic content
- Feature extraction for voice conversion (ASR-bottleneck)

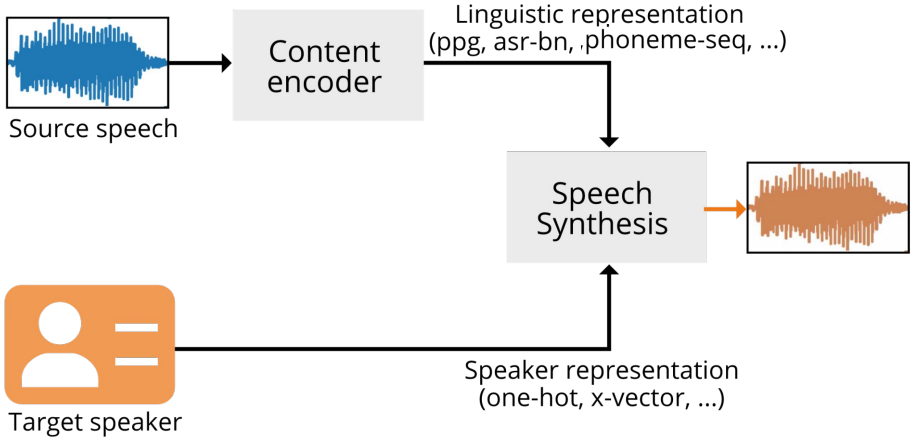
Automatic speaker verification (ASV)



ASV is used for:

- Automatic objective evaluation of the removal of speaker's identifier clues

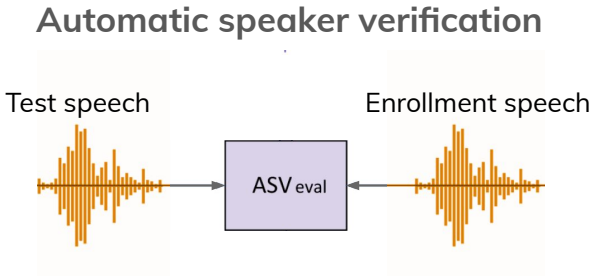
Voice conversion (VC)



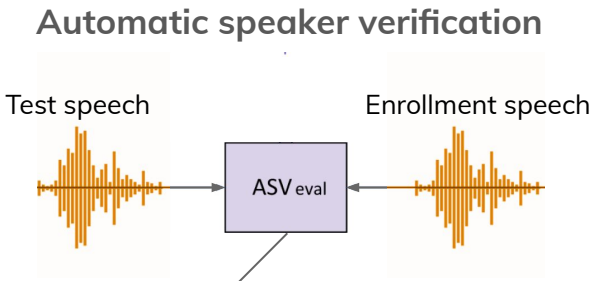
VC is used for:

- Replacing the source speaker's identity with a target speaker
  → Can be used for speaker anonymization

Threat model - Linkability assessment



Threat model - Linkability assessment



Genuine scores: speakers are the same

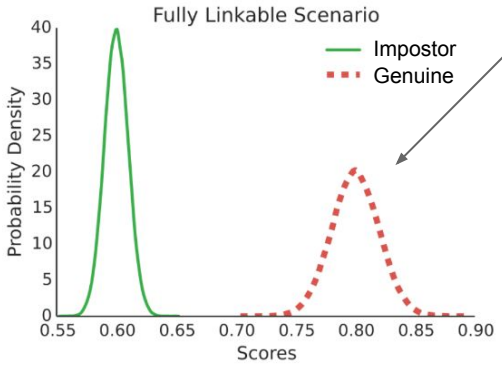
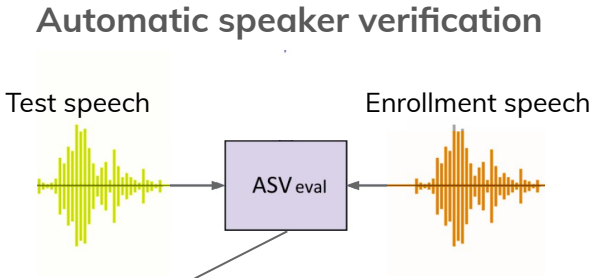


Image from: General Framework to Evaluate Unlinkability in Biometric Template Protection Systems, Gomez-Barrero et al

Threat model - Linkability assessment



Impostor scores: speakers are different
Genuine scores: speakers are the same

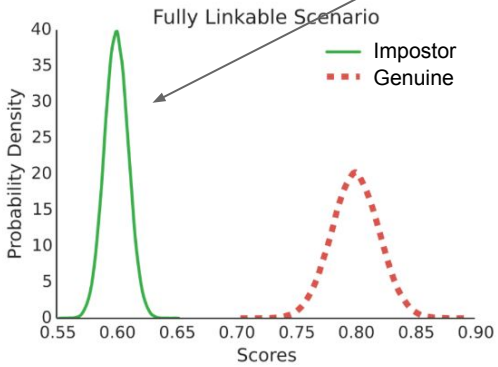
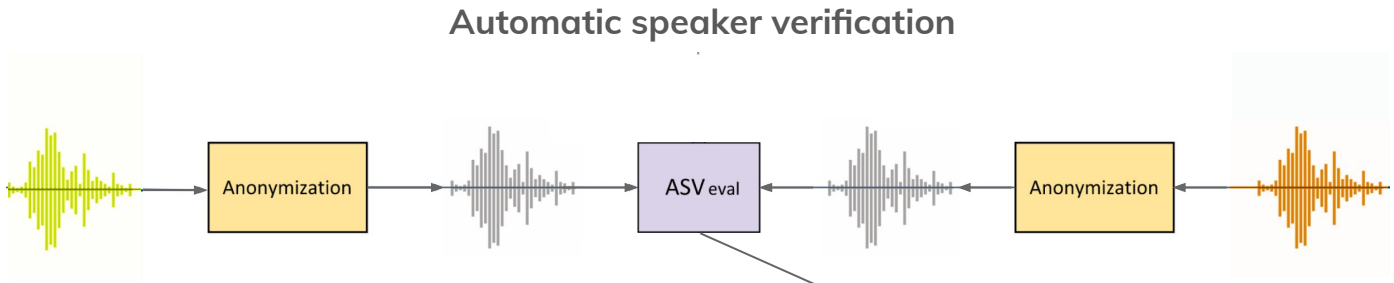
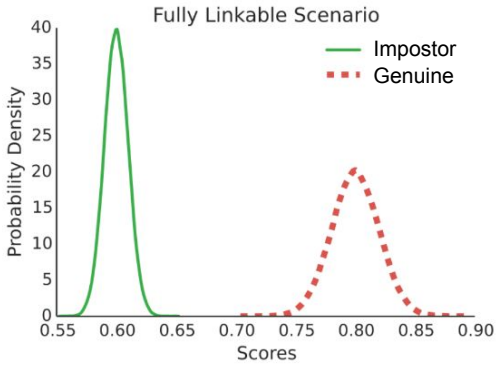


Image from: General Framework to Evaluate Unlinkability in Biometric Template Protection Systems, Gomez-Barrero et al

Threat model - Linkability assessment



Impostor scores: speakers are different
Genuine scores: speakers are the same



Anonymization

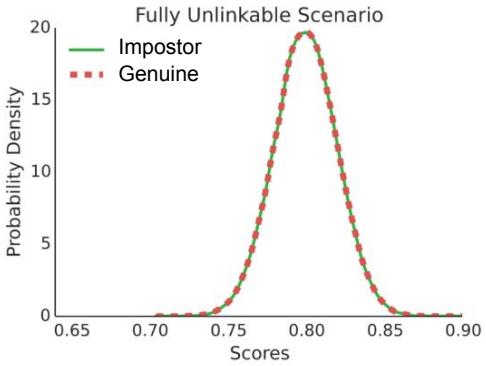
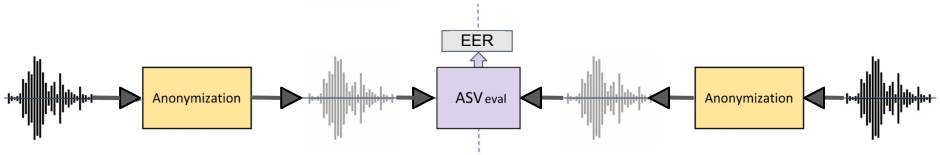


Image from: General Framework to Evaluate Unlinkability in Biometric Template Protection Systems, Gomez-Barrero et al

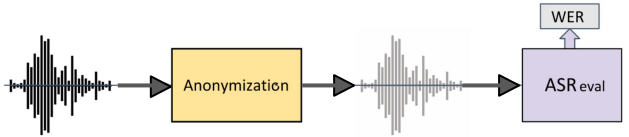
Evaluation methods - VoicePrivacy protocol [1]

Privacy:



Privacy evaluation using
Automatic speaker verification
Metric: **EER** (maximize) **Linkability** (minimize)

Utility:

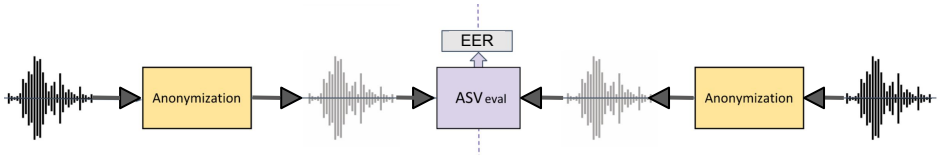


Utility evaluation using
Automatic speech recognition
Metric: **WER** (minimize)

1. Tomashenko, Natalia, Brij Mohan Lal Srivastava, et al., « Introducing the VoicePrivacy Initiative », *in*: *Interspeech*, 2020

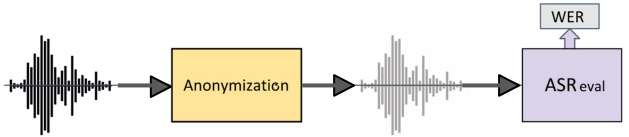
Evaluation methods - VoicePrivacy protocol [1]

Privacy:



Privacy evaluation using
Automatic speaker verification
Metric: **EER** (maximize) **Linkability** (minimize)

Utility:



Utility evaluation using
Automatic speech recognition
Metric: **WER** (minimize)

Both ASV and ASR models are trained on anonymized speech

Baseline clear speech	1.1% EER
Not trained on anonymized speech	36.7% EER
Trained on anonymized speech	10.7% EER

Baseline clear speech	5.5% WER
Not trained on anonymized speech	11.7% WER
Trained on anonymized speech	5.8% WER

1. Tomashenko, Natalia, Brij Mohan Lal Srivastava, et al., « Introducing the VoicePrivacy Initiative », *in*: *Interspeech*, 2020

Outline:

- Background
- **Impact of the target speaker parameter in VC anonymization**
  - How does the target selection algorithms impact the generation of anonymized speech and the evaluation pipeline?
  - Is there a relationship between the speaker and target parameters that maximizes privacy?
- Anonymization with feature-level disentanglement
- Conclusion and perspectives

Impact of the target speaker parameter in VC anonymization

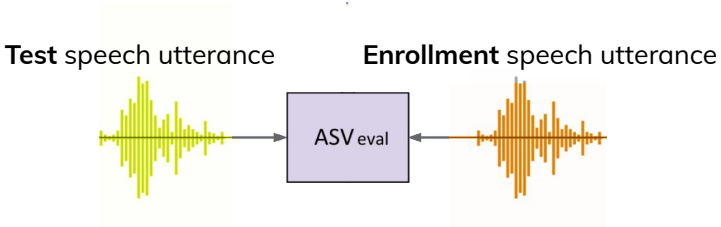
Test dataset: Librispeech test-clean male

Set	Speakers	Utterances
Enrollment	13	184
Test	20	762

140 208 scores

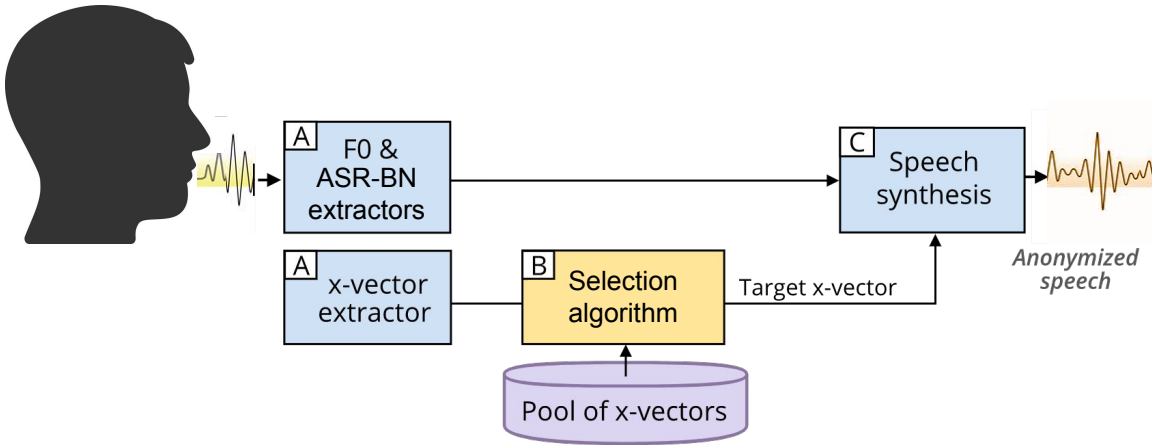
Statistics of the VoicePrivacy Librispeech evaluation dataset

ASV scoring is done on **test** and **enrollment** pairs by combining all possible combinations of **utterance-to-utterance** representations
Average length of segments: ~= 7 sec



Impact of the target speaker parameter in VC anonymization

Voice conversion model

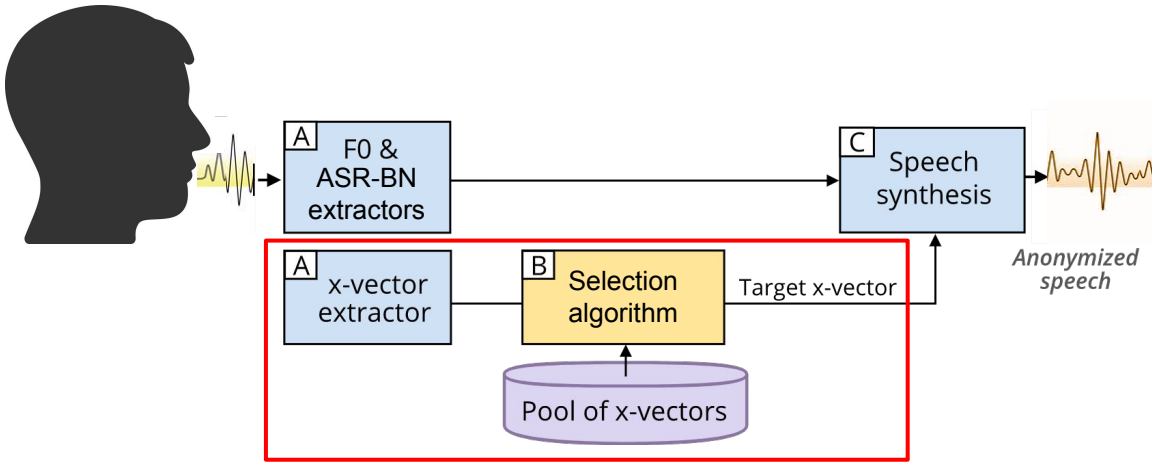


Speaker anonymization system [1]

1. Tomashenko, Natalia, Brij Mohan Lal Srivastava, et al., « Introducing the VoicePrivacy Initiative », *in*: *Interspeech*, 2020

Impact of the target speaker parameter in VC anonymization

To generate anonymized speech using VC, a target speaker is used



Speaker anonymization system

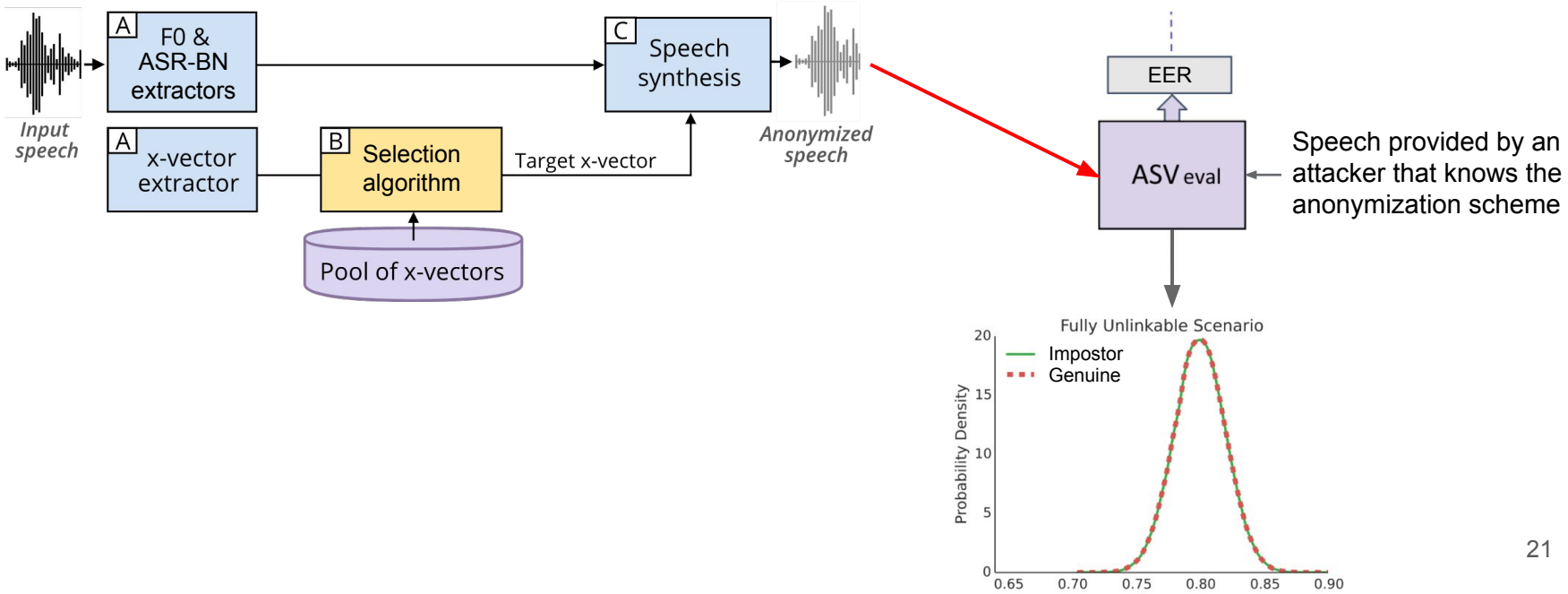
How does the target selection algorithms impact the generation of anonymized speech and the evaluation pipeline?

Impact of the target speaker parameter in VC anonymization

Evaluation procedure

The **utterances are anonymized independently** using the target selection algorithm

The voice conversion anonymization system is the VoicePrivacy 2022 baseline system (HiFi-GAN-based model)

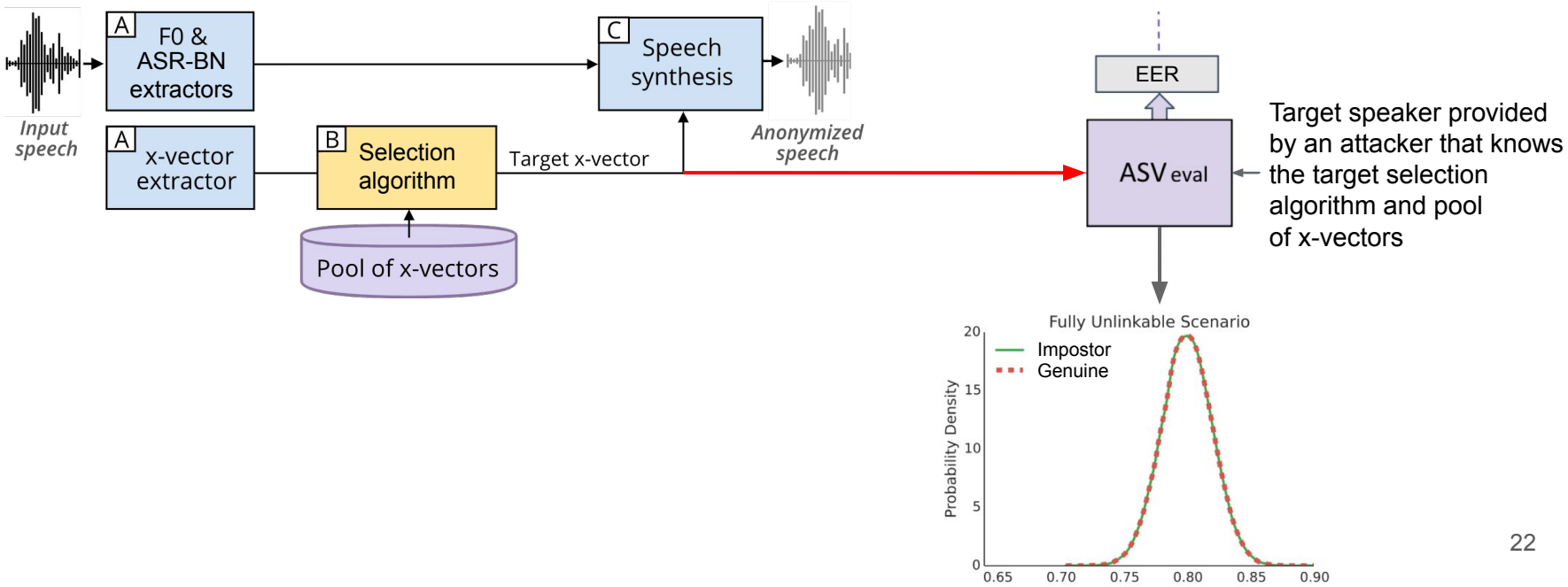


Impact of the target speaker parameter in VC anonymization

Evaluation procedure

The **utterances are anonymized independently** using the target selection algorithm

The voice conversion anonymization system is the VoicePrivacy 2022 baseline system (HiFi-GAN-based model)



Impact of the target speaker parameter in VC anonymization

We have identified and analyzed multiple target selection algorithms

- VoicePrivacy (VPC) baseline target selection [1]
- Random vector target selection
- Random speaker target selection [2]
- Constant speaker target selection [2]
- Dense area target selection [2]

1. Tomashenko, Natalia, Brij Mohan Lal Srivastava, et al., « Introducing the VoicePrivacy Initiative », *in*: *Interspeech*, 2020
2. Srivastava, Brij Mohan Lal, Natalia Tomashenko, et al., « Design Choices for X-vector Based Speaker Anonymization », *in*: *Interspeech*, 2020

Impact of the target speaker parameter in VC anonymization

Privacy metrics computed using directly the target x-vector of the VC system, and using the x-vector extracted from the anonymized speech.
The confidence interval stays within ± 0.4% EER for all experiments [feerci].

Subject	target x-vector	speech x-vector
Metric	Privacy EER% ↑	Privacy EER% ↑
Clear speech		3.1

[feerci] E. Haasnoot, A. Khodabakhsh, C. Zeinstra, L. Spreeuwers and R. Veldhuis, "FEERCI: A Package for Fast Non-Parametric Confidence Intervals for Equal Error Rates in Amortized O(m log n)," 2018 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 2018

Impact of the target speaker parameter in VC anonymization

Privacy metrics computed using directly the target x-vector of the VC system, and using the x-vector extracted from the anonymized speech.
The confidence interval stays within ± 0.4% EER for all experiments.

Subject	target x-vector	anonymized speech x-vector
Metric	Privacy EER% ↑	Privacy EER% ↑
Clear speech		3.1
VPC farther 200 random 100	4.8	

Impact of the target speaker parameter in VC anonymization

Privacy metrics computed using directly the target x-vector of the VC system, and using the x-vector extracted from the anonymized speech.
The confidence interval stays within ± 0.4% EER for all experiments.

Subject	target x-vector	anonymized speech x-vector
Metric	Privacy EER% ↑	Privacy EER% ↑
Clear speech		3.1
VPC farther 200 random 100	4.8	19.5

Impact of the target speaker parameter in VC anonymization

Privacy metrics computed using directly the target x-vector of the VC system, and using the x-vector extracted from the anonymized speech.
The confidence interval stays within ± 0.4% EER for all experiments.

Subject	target x-vector	anonymized speech x-vector
Metric	Privacy EER% ↑	Privacy EER% ↑
Clear speech		3.1
VPC farther 200 random 100	4.8	19.5
Random vector	50.0	
Random speaker	50.0	
Constant speaker	50.0	

Impact of the target speaker parameter in VC anonymization

Privacy metrics computed using directly the target x-vector of the VC system, and using the x-vector extracted from the anonymized speech.
The confidence interval stays within ± 0.4% EER for all experiments.

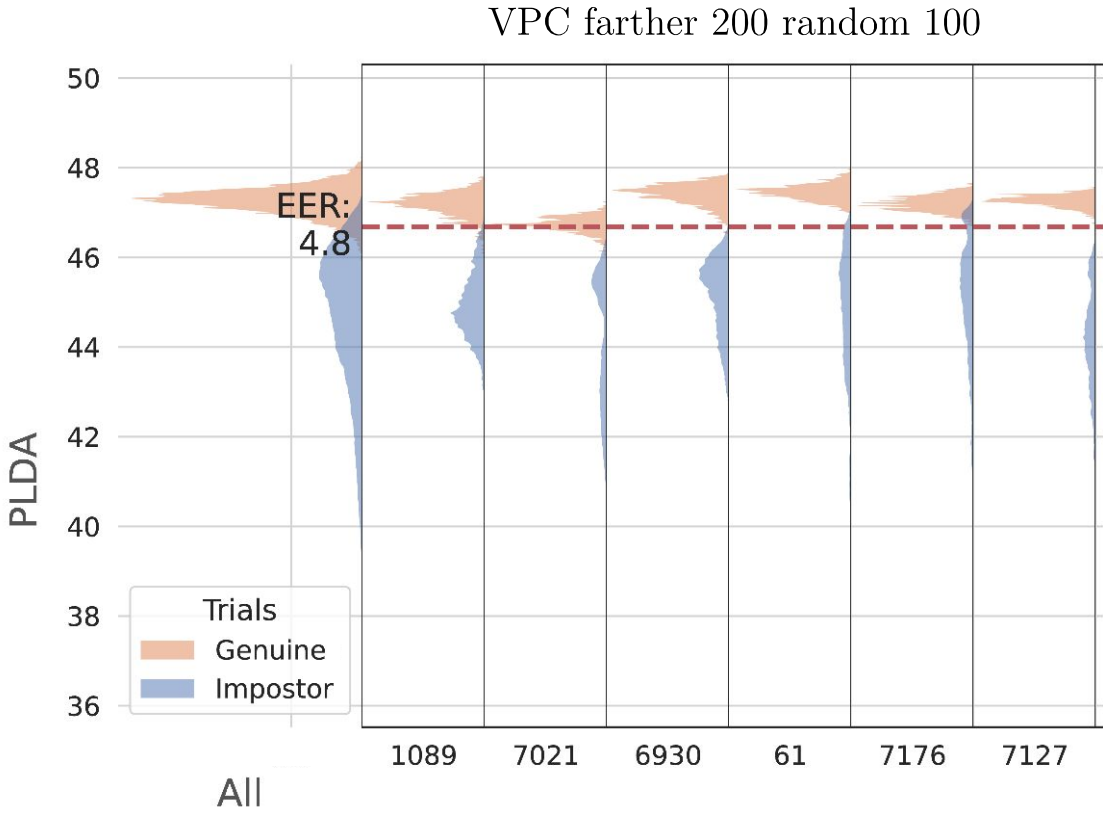
Subject	target x-vector	anonymized speech x-vector
Metric	Privacy EER% ↑	Privacy EER% ↑
Clear speech		3.1
VPC farther 200 random 100	4.8	19.5
Random vector	50.0	23.3
Random speaker	50.0	21.6
Constant speaker	50.0	22.4

Impact of the target speaker parameter in VC anonymization

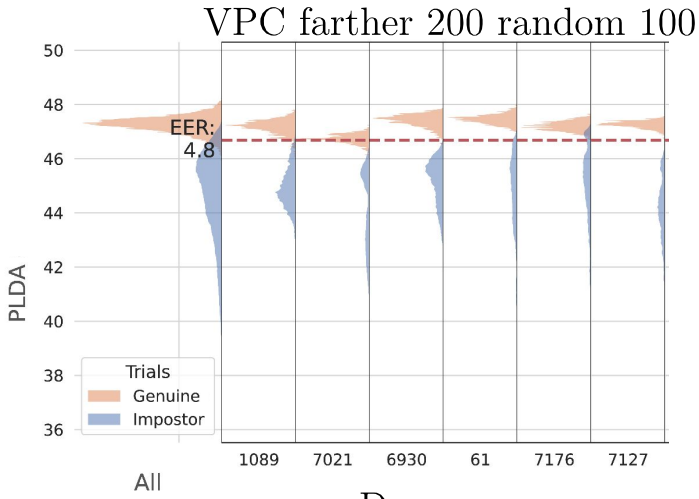
Privacy metrics computed using directly the target x-vector of the VC system, and using the x-vector extracted from the anonymized speech.
The confidence interval stays within ± 0.4% EER for all experiments.

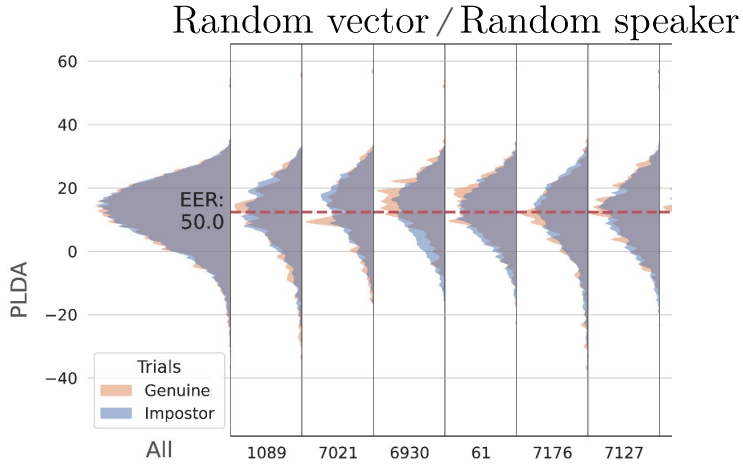
Subject	target x-vector	anonymized speech x-vector
Metric	Privacy EER% ↑	Privacy EER% ↑
Clear speech		3.1
VPC farther 200 random 100	4.8	19.5
Random vector	50.0	23.3
Random speaker	50.0	21.6
Constant speaker	50.0	22.4
Dense area	43.1	39.2

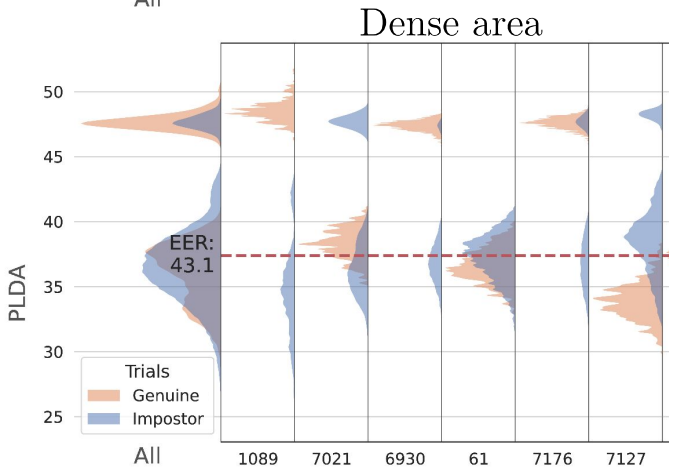
Impact of the target speaker parameter in VC anonymization

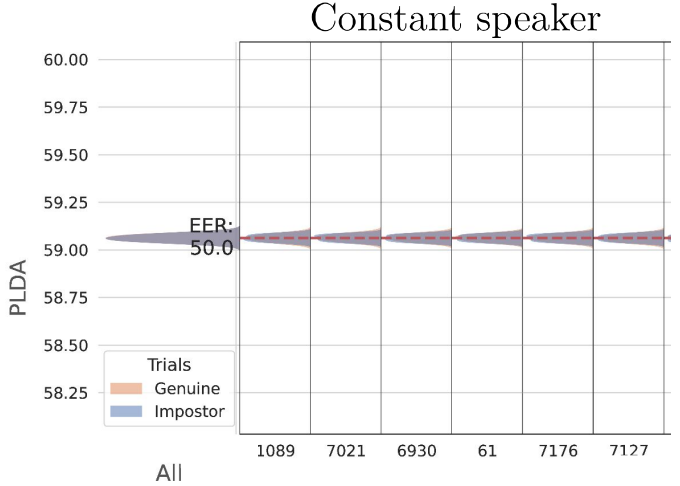


Impact of the target speaker parameter in VC anonymization



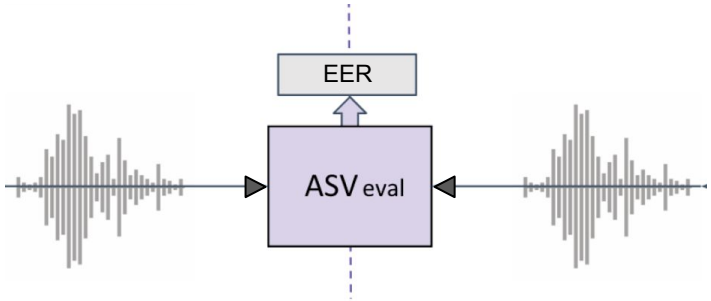






Anonymized data used to train the evaluation model

The anonymized data of the dense target selection is not suitable to train a proper ASV model



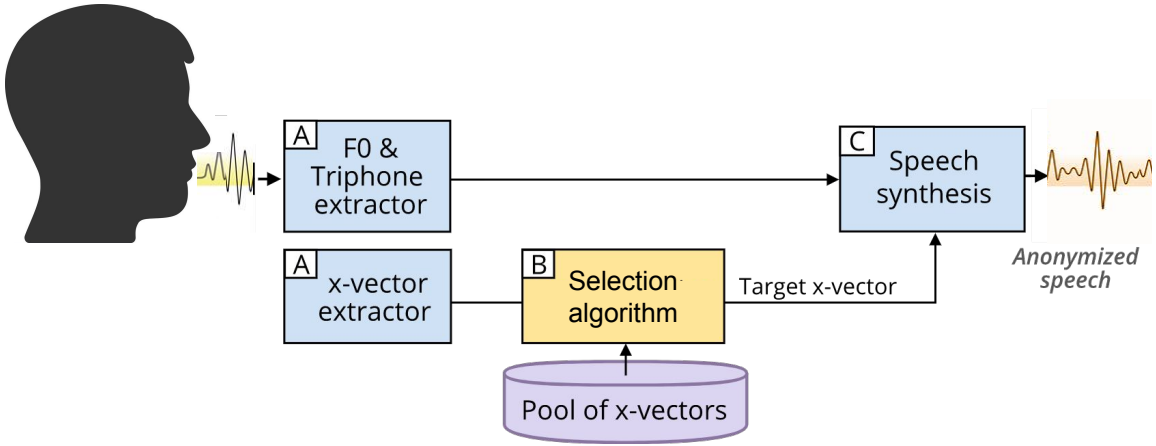
Impact of the target speaker parameter in VC anonymization

Privacy metrics computed using directly the target x-vector of the VC system, and using the x-vector extracted from the anonymized speech.
The confidence interval stays within ± 0.4% EER for all experiments.

Subject	target x-vector	anonymized speech x-vector
Metric	Privacy EER% ↑	Privacy EER% ↑
Clear speech		3.1
VPC farther 200 random 100	4.8	19.5
Random vector	50.0	23.3
Random speaker	50.0	21.6
Constant speaker	50.0	22.4
Dense area, Dense area ASV model	43.1	39.2
Dense area, Random spk ASV model	43.1	20.9

Use target selection algorithms that have **50% of EER on the target**

Impact of the target speaker parameter in VC anonymization



Speaker anonymization system

Is there a relationship between the speaker and target parameters that maximizes privacy?

Select 40 target x-vectors (20 males & 20 females) and evaluate their privacy performances

Impact of the target speaker parameter in VC anonymization

Test dataset: Librispeech test-clean

Set	Speakers	Utterances
Test	40	1496
Enrollment	29	438

43 384 scores

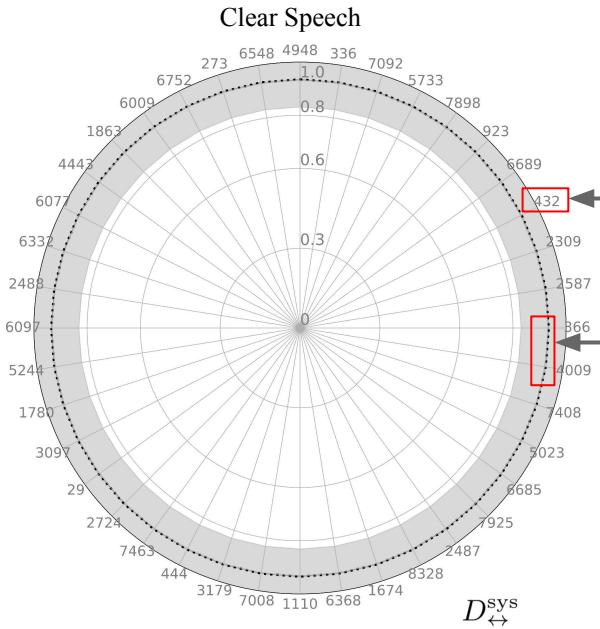
Statistics of the VoicePrivacy Librispeech evaluation dataset

ASV scoring is done on **test** and **enrollment** pairs by combining all possible combinations of **utterance-to-speaker** representations

Average length of segments: ~= 7 sec

Impact of the target speaker parameter in VC anonymization

Detailed analysis



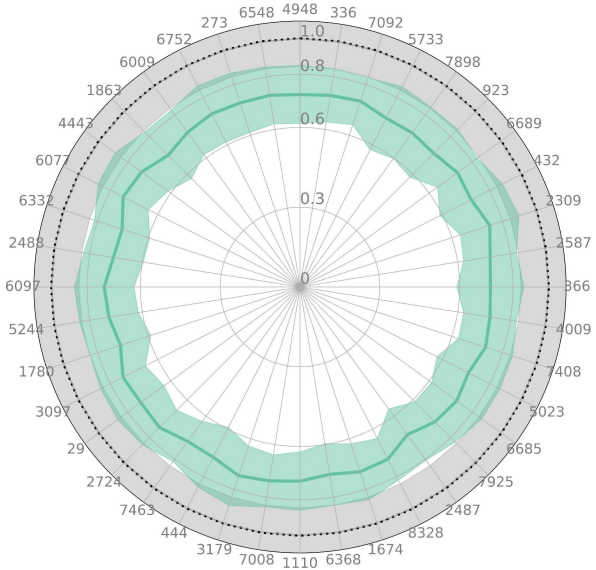
The label of the target speaker

The linkability score on original speech
(No anonymization)

Impact of the target speaker parameter in VC anonymization

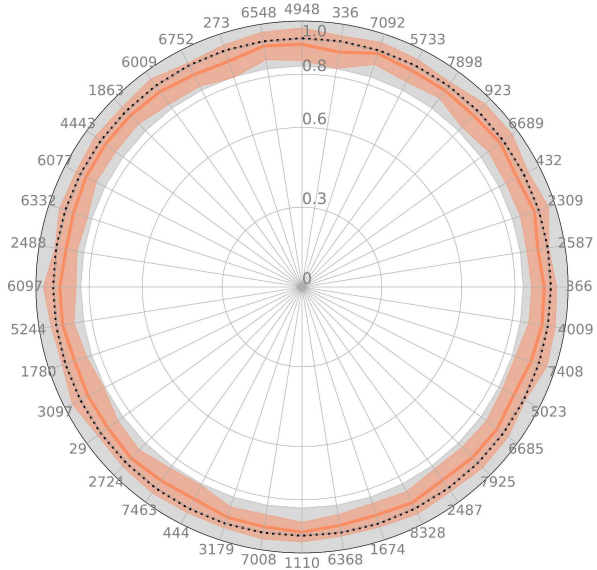
Detailed analysis

Well anonymized speakers



14/29 speakers
Outside of the linkability performance on original speech. The separation is distinct, speaker information was removed by the anonymization system

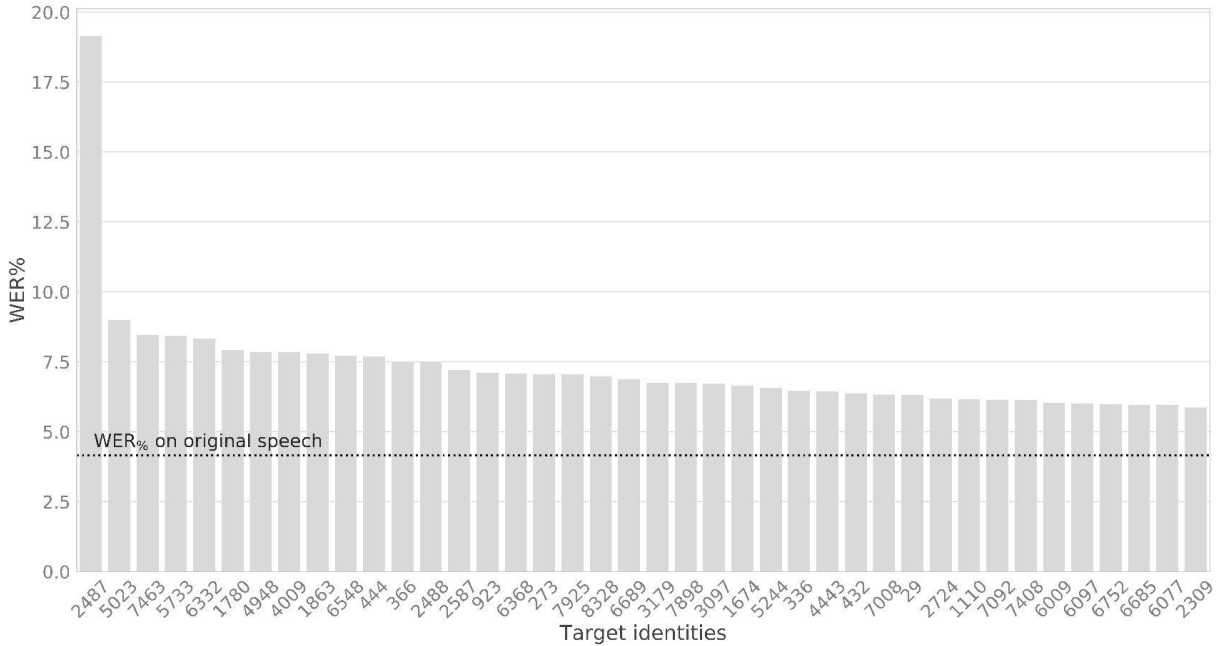
Poorly anonymized speakers



15/29 speakers
Overlap is complete, the anonymization system did not remove speaker information for half of our test speakers

Impact of the target speaker parameter in VC anonymization

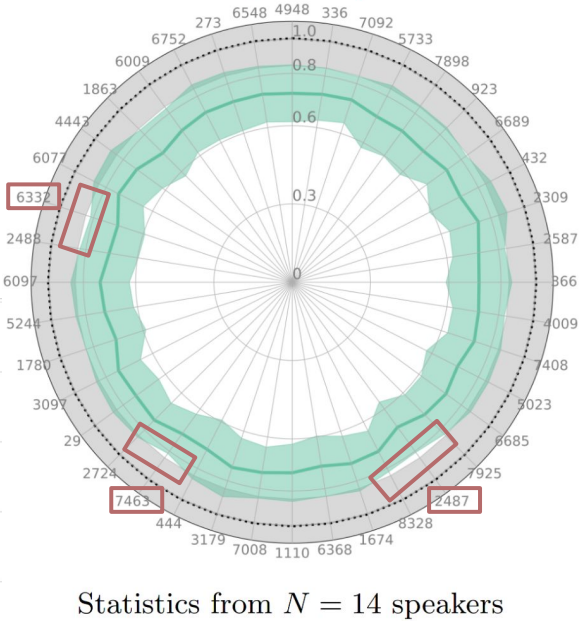
Detailed analysis: utility



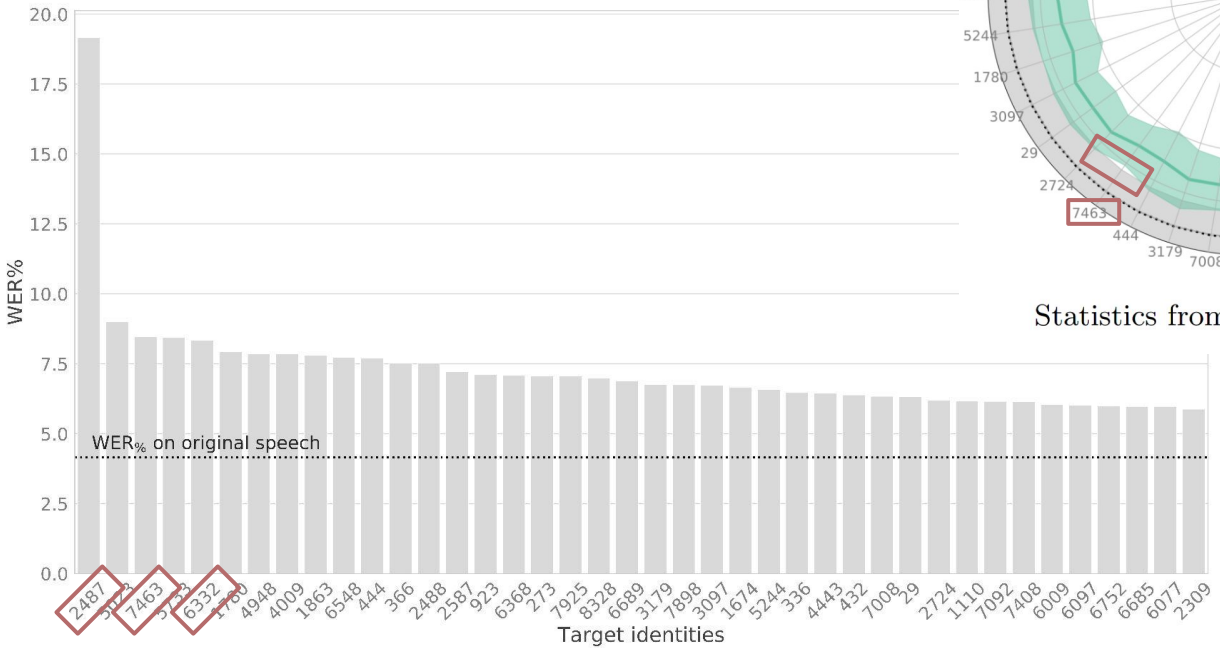
WER% scores obtained on the anonymized speech by the automatic speech recognition evaluation system for each of the 40 targets and original (dotted) line

Impact of the target speaker in VC anonymization

Detailed analysis: utility



Statistics from $N = 14$ speakers

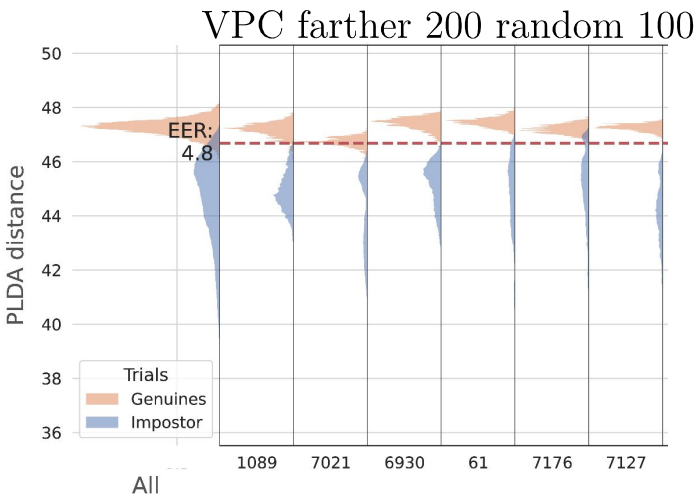


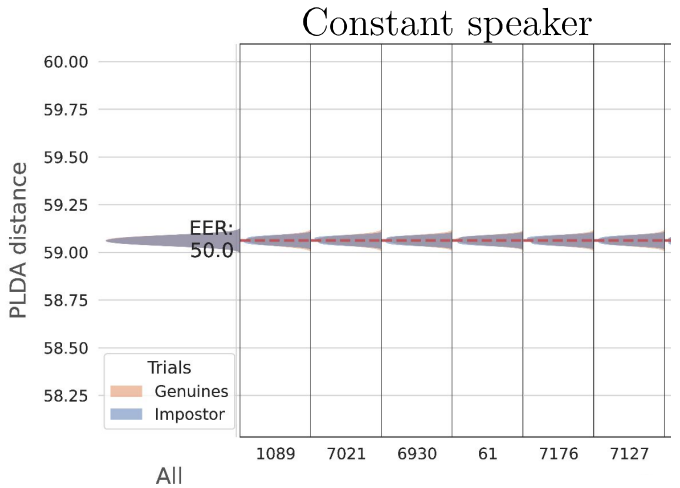
WER% scores obtained on the anonymized speech by a non-adapted automatic speech recognition system for each of the 40 targets and original (dotted) line

Impact of the target speaker parameter in VC anonymization

How does the target selection algorithms impacts the generation of anonymized speech and the evaluation pipeline?

- It impacts the expected privacy results
- It impacts the ASV eval





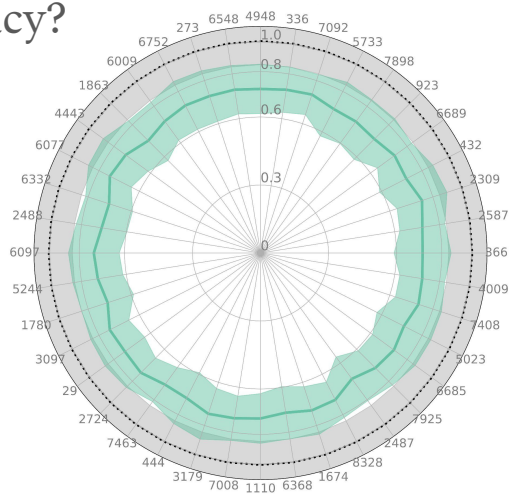
Impact of the target speaker parameter in VC anonymization

How does the target selection algorithms impacts the generation of anonymized speech and the evaluation pipeline?

- It impacts the expected privacy results
- It impacts the ASV eval

Is there a relationship between the speaker and target parameters that maximizes privacy?

- Not to a significant extent for privacy
- But it does impact the utility



Impact of the target speaker parameter in VC anonymization

How does the target selection algorithms impacts the generation of anonymized speech and the evaluation pipeline?

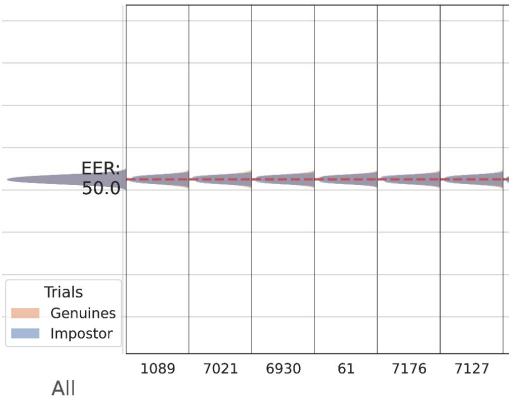
- It impacts the expected privacy results
- It impacts the ASV eval

Is there a relationship between the speaker and target parameters that maximizes privacy?

- Not to a significant extent for privacy
- But it does impact the utility

Constant speaker target selection

- Fulfill unlinkability at the target level
- Allows to pick a target speaker that maximizes the utility
- Simple



Outline:

- Background
- Impact of the target speaker parameter in VC anonymization
- **Anonymization with feature-level disentanglement**
  - Baseline
  - Privacy enhancing with adversarial training
  - Privacy enhancing with inplace modification
- Conclusion and perspectives

Impact of the target speaker parameter in VC anonymization

Test dataset: VCTK

Set	Speakers	Utterances
Test	30	11448
Enrollment	30	600

343 440 scores

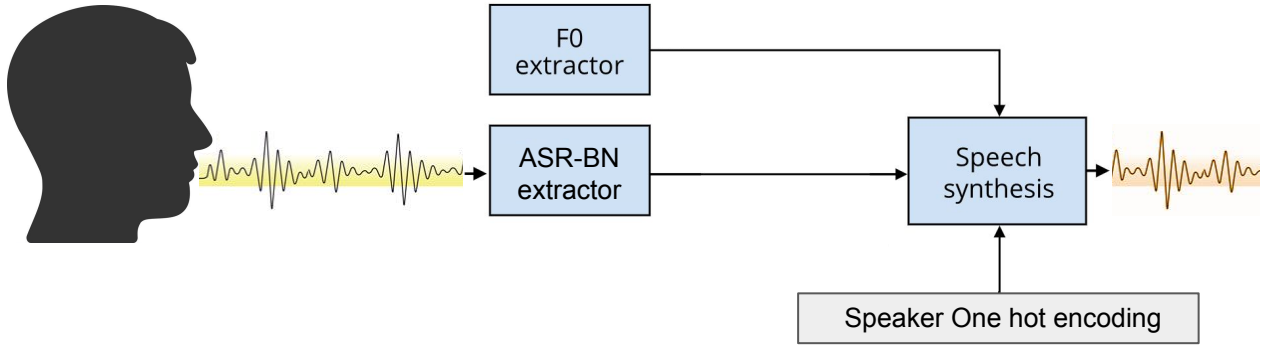
Statistics of the VCTK evaluation dataset

ASV scoring is done on **test** and **enrollment** pairs by combining all possible combinations of **utterance-to-speaker** representations

Average length of segments: ~= 3 sec

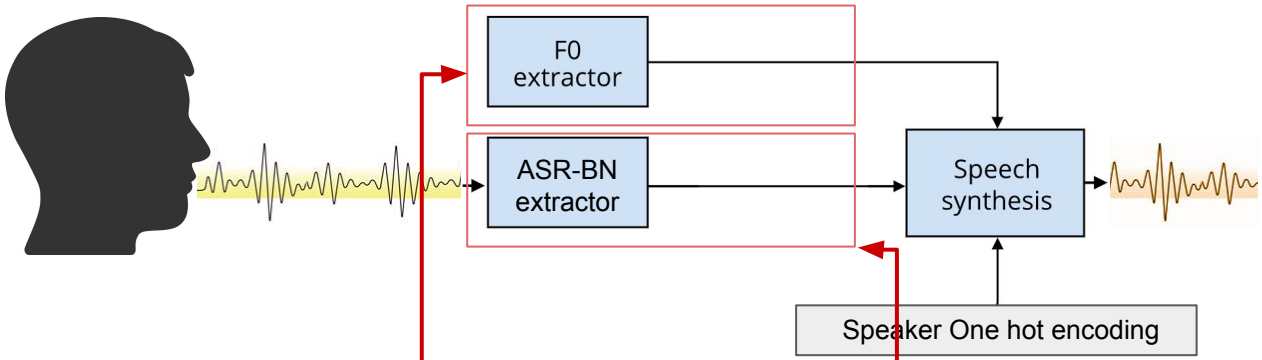
Anonymization with feature-level disentanglement

Baseline



Anonymization with feature-level disentanglement

Baseline feature extractors

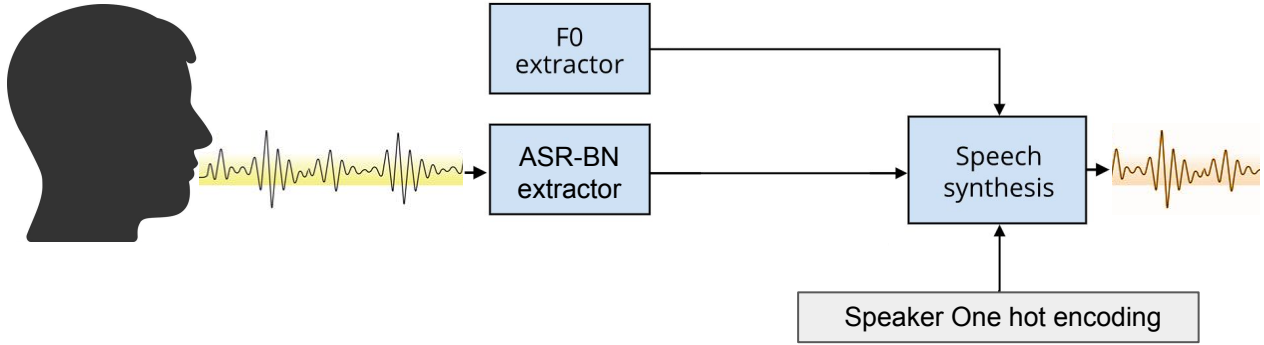


F0 - Encodes:
- Intonation

ASR-Bottleneck (ASR-BN) - Encodes:
- Speech sounds corresponding to the content

Anonymization with feature-level disentanglement

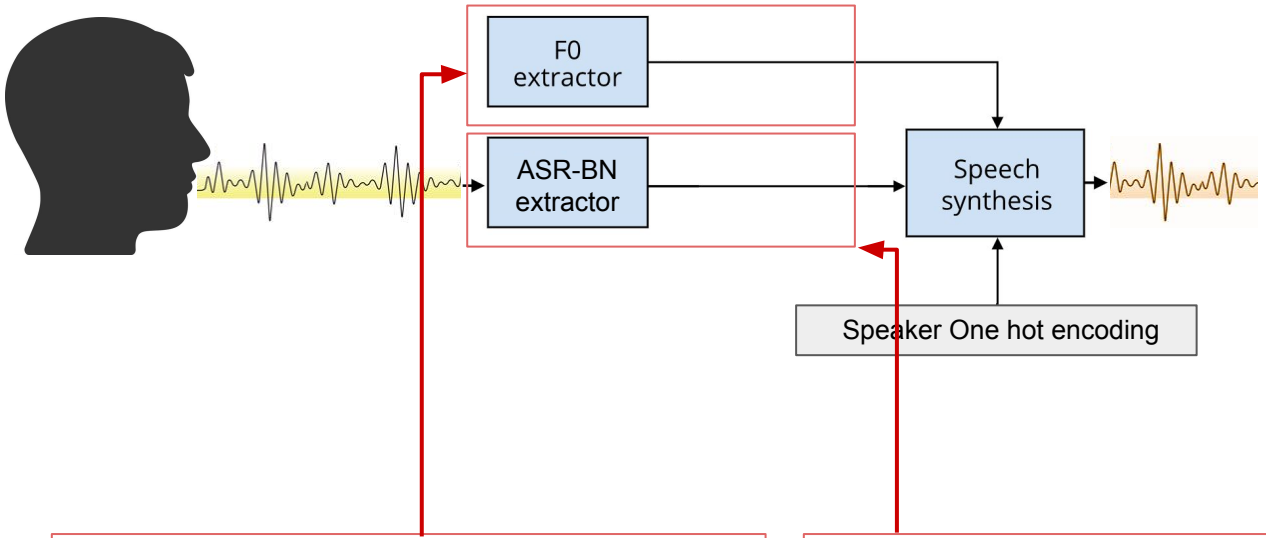
Baseline results



Dataset	VCTK test	
Method	Privacy EER% ↑	Utility WER% ↓
Clear speech	2.7	12.8
ASR-BN TDNNF	10.8	19.1

Anonymization with feature-level disentanglement

Baseline privacy leakage

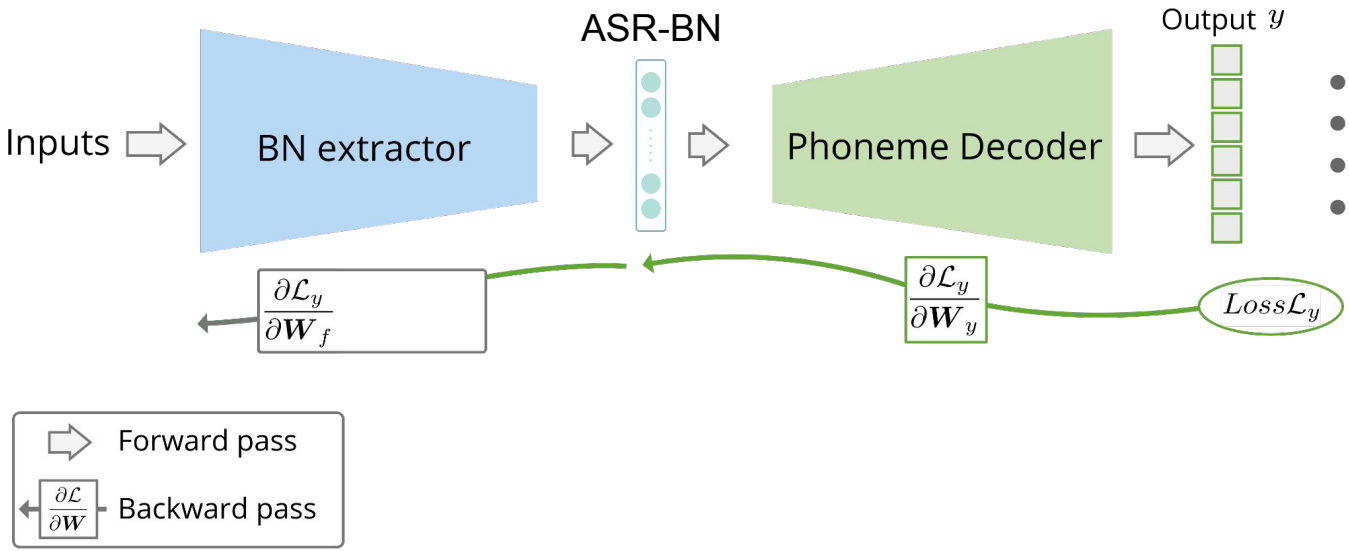


Encodes:
- Intonation
- Speaker Identity EER: 20%

Encodes:
- Speech sounds corresponding to the content
- Speaker Identity EER: 6.7%

Anonymization with feature-level disentanglement

ASR-BN extractor for baseline

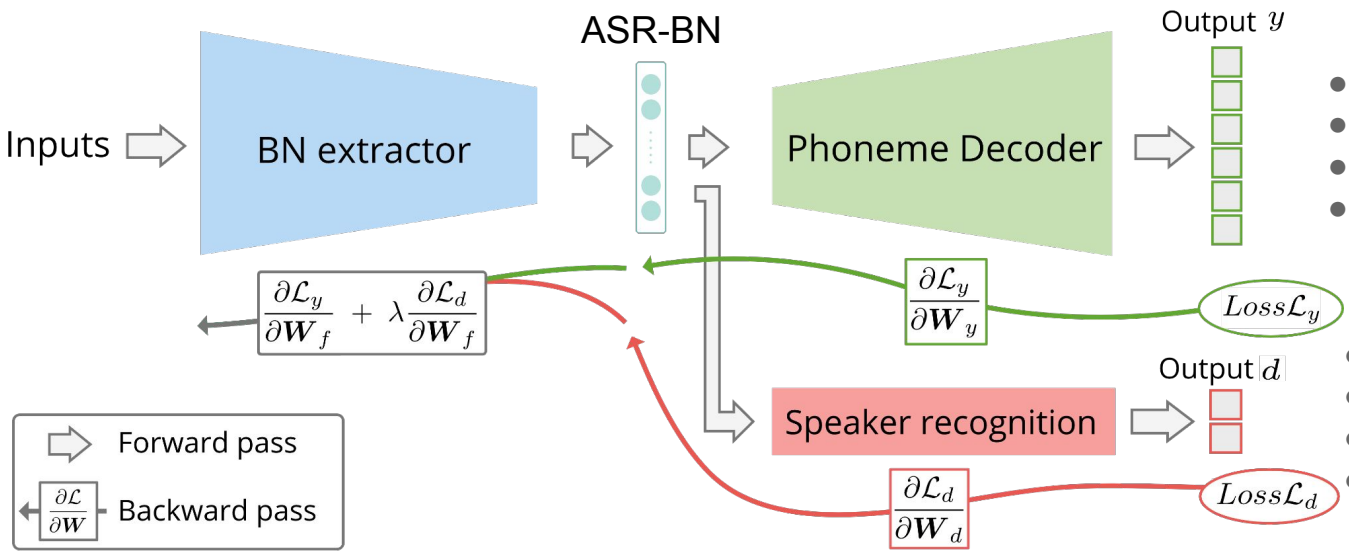


- Pytorch kaldi wrapper
- 15 layers F-TDNN [3]
- Left-biphone classif
- 100h of training data

3. Povey, Daniel, Gaofeng Cheng, et al., « Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks », *Interspeech*, 2018

Anonymization with feature-level disentanglement

Privacy enhanced ASR-BN extractor with adversarial training



- Pytorch kaldi wrapper
- 15 layers F-TDNN
- Left-biphone classif
- 100h of training data

- Sidekit
- 5 layers TDNN
- Speaker classif
- 600h of training data

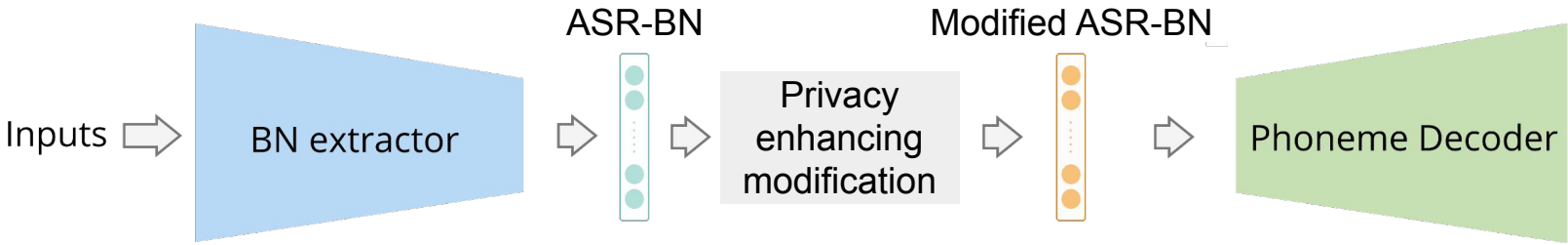
Privacy enhanced ASR-BN extractor with adversarial training

Dataset	VCTK test	
Method	Privacy EER% ↑	Utility WER% ↓
Clear speech	2.7	12.8
ASR-BN TDNNF	10.8	19.1
ASR-BN TDNNF ADVERSARIAL	11.8	14.4

Utility improvement because of the additional training weakly labeled data [4]
Similar privacy: adversarial training only removes
the speaker information that the adversarial network observe

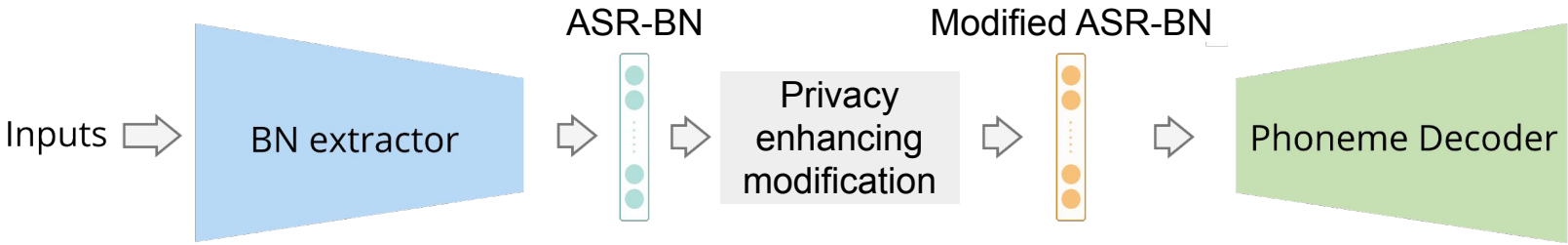
4. Adi, Y., N. Zeghidour, et al., « To Reverse the Gradient or Not: an Empirical Comparison of Adversarial and Multi-task Learning in Speech Recognition », *in*: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* 2019

Privacy enhanced ASR-BN extractor with inplace modification

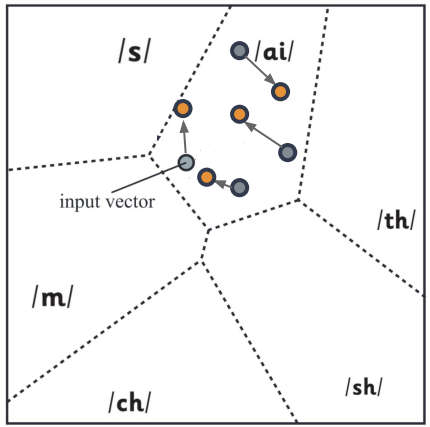


Anonymization with feature-level disentanglement

Privacy enhanced ASR-BN extractor with inplace modification



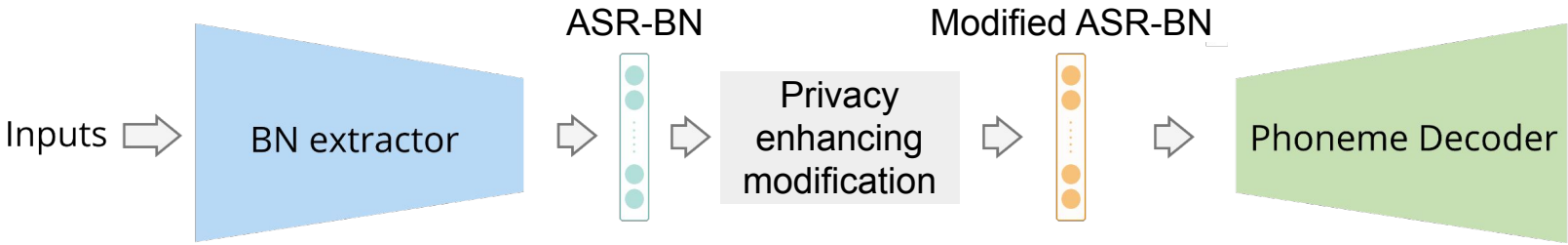
Noise addition [5]



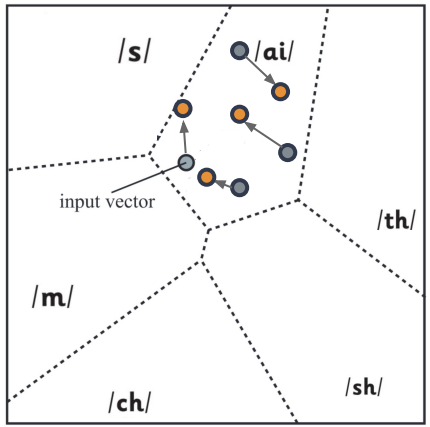
5. Shamsabadi, Ali Shahin, Brij Mohan Lal Srivastava, et al., « Differentially private speaker anonymization », *in*: *Privacy Enhancing Technologies*, 2022

Anonymization with feature-level disentanglement

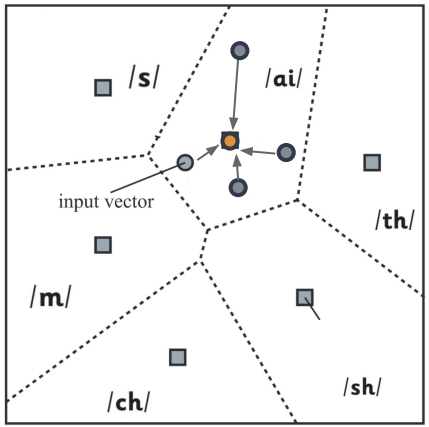
Privacy enhanced ASR-BN extractor with inplace modification



Noise addition [5]



Vector quantization (proposed)



5. Shamsabadi, Ali Shahin, Brij Mohan Lal Srivastava, et al., « Differentially private speaker anonymization », *in*: *Privacy Enhancing Technologies*, 2022

Privacy enhanced ASR-BN extractor

Dataset	VCTK test	
Method	Privacy EER% ↑	Utility WER% ↓
Clear speech	2.7	12.8
ASR-BN TDNNF	10.8	19.1
ASR-BN TDNNF ADVERSARIAL	11.8	14.4
ASR-BN TDNNF VQ 64	30.0	29.1

Privacy improvement because of the vector quantization
Utility degradation because of the smaller encoding capacity of the network

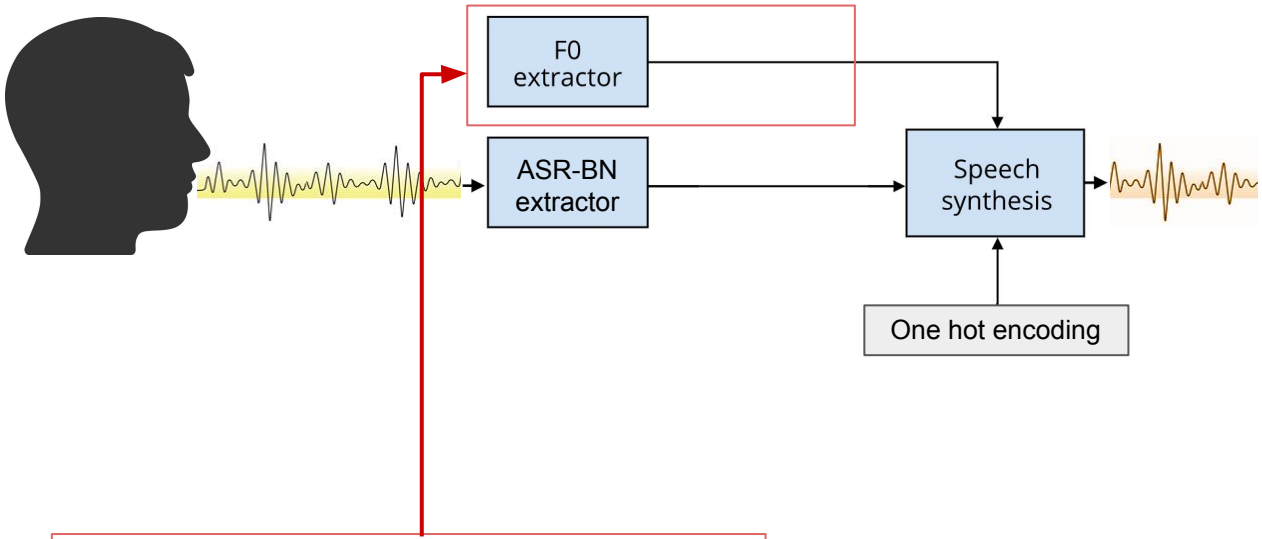
Privacy enhanced ASR-BN extractor

Dataset	VCTK test	
Method	Privacy EER% ↑	Utility WER% ↓
Clear speech	2.7	12.8
ASR-BN TDNNF	10.8	19.1
ASR-BN TDNNF ADVERSARIAL	11.8	14.4
ASR-BN TDNNF VQ 64	30.0	29.1
ASR-BN WAV2VEC2 VQ 48	28.0	10.0

Self-supervised feature extractor such as Wav2vec2 helps to improve the utility

Anonymization with feature-level disentanglement

Baseline privacy leakage



Privacy enhanced ASR-BN and F0

Dataset	VCTK test	
Method	Privacy EER% ↑	Utility WER% ↓
Clear speech	2.7	12.8
ASR-BN TDNNF	10.8	19.1
ASR-BN TDNNF ADVERSARIAL	11.8	14.4
ASR-BN TDNNF VQ 64	30.0	29.1
ASR-BN WAV2VEC2 VQ 48	28.0	10.0
ASR-BN WAV2VEC2 VQ 48 + F_0 VQ	39.8	9.9

Fully vector quantized voice conversion based speaker anonymization achieves the best results in both utility and privacy metric

Privacy enhanced ASR-BN and F0

Dataset	VCTK test	
Method	Privacy EER% ↑	Utility WER% ↓
Clear speech	2.7	12.8
ASR-BN TDNNF	10.8	19.1
ASR-BN TDNNF ADVERSARIAL	11.8	14.4
ASR-BN TDNNF VQ 64	30.0	29.1
ASR-BN WAV2VEC2 VQ 48	28.0	10.0
ASR-BN WAV2VEC2 VQ 48 + F_0 VQ	39.8	9.9
ASR-BN WAV2VEC2 NOISE + F_0 NOISE	41.4	10.2

Similar results as a fully noise based voice conversion speaker anonymization system
VQ transformation easier to train than the noise approach

Outline:

- Context and background
- Impact of the target speaker parameter in VC anonymization
- Anonymization with feature-level disentanglement
- **Conclusion and perspectives**

Impact of the target speaker parameter in VC anonymization

We analyzed in a comprehensive way the role and impact of the choice of the target speaker parameter

- Provide a better understanding of the role of the target speaker for anonymized speech generation and evaluation [1,2]
- Define the most suitables target speaker selection algorithm that maximise privacy, utility and is the less keen to evaluation biases [1,2]

Improvement of privacy evaluation while simplifying the pipeline

1. Champion, Pierre, Denis Jouvet, and Anthony Larcher, « Speaker information modification in the VoicePrivacy 2020 toolchain », *in*: *VoicePrivacy 2020 Virtual Workshop at Odyssey*, 2020
2. — « Evaluating X-vector-based Speaker Anonymization under White-box Assessment », *in*: *International Conference on Speech and Computer (SPECOM)*, 2021

Anonymization with feature-level disentanglement

We analyzed and proposed improvements of the features used by VC to generate anonymized voice

- Adversarial learning to improve utility in restricted data availability scenario [1]
- Vector quantization as an alternative to noise addition privacy modification [3,4,5]
- Use of pre-trained feature (wav2vec2) to improve utility [5]

1. Champion, Pierre, Denis Jouvet, and Anthony Larcher, « Speaker information modification in the VoicePrivacy 2020 toolchain », *in*: *VoicePrivacy 2020 Virtual Workshop at Odyssey*, 2020 [1]
3. — « A Study of F0 Modification for X-Vector Based Sp ch Pseudonymization Across Gender », *in*: *The Second AAAI Workshop on Privacy-Preserving Artificial Intelligence*, 2020
4. — « Privacy-Preserving Speech Representation Learning using Vector Quantization », *in*: *Journées d'Études sur la Parole (JEP, 34e édition)*, 2022
5. — « Are disentangled representations all you need to build speaker anonymization systems? », *in*: *Interspeech*, 2022

Perspectives *(from the report)*

- Proposed an invertibility attack [6]
- Identified an utility evaluation limitation

	Utility WER% ↓
Clear speech	12.8
ASR-BN WAV2VEC2 VQ 48 + F_0 VQ	9.9

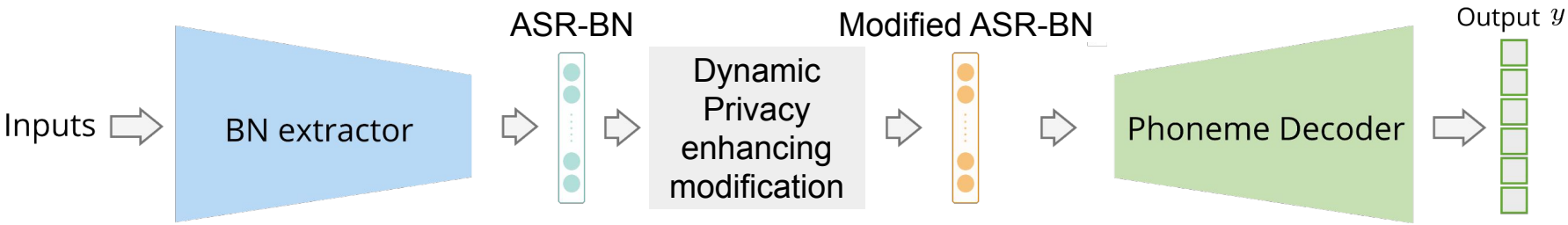
ASR eval decoding errors

+ VC anonymization mispronunciation errors

6. Champion, Pierre, Thomas Thebaud, et al., « On the invertibility of a voice privacy system using embedding alignment », *in*: *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021

Perspectives

- Dynamically adapt for each frame the privacy modification to maximize privacy/utility



Thank for your attention!

Question slides!

Other privacy and utility measurements *(from the report)*

- Proposed an invertibility attack [6]
- Identified an utility evaluation limitation

		Utility WER% ↓
ASR: 360h	Clear speech	12.8
VC: 22k+100h ASR: 360h	BN WAV2VEC2 VQ 48 + F_0 VQ	9.9

Decoding errors

+ errors due to mispronunciation

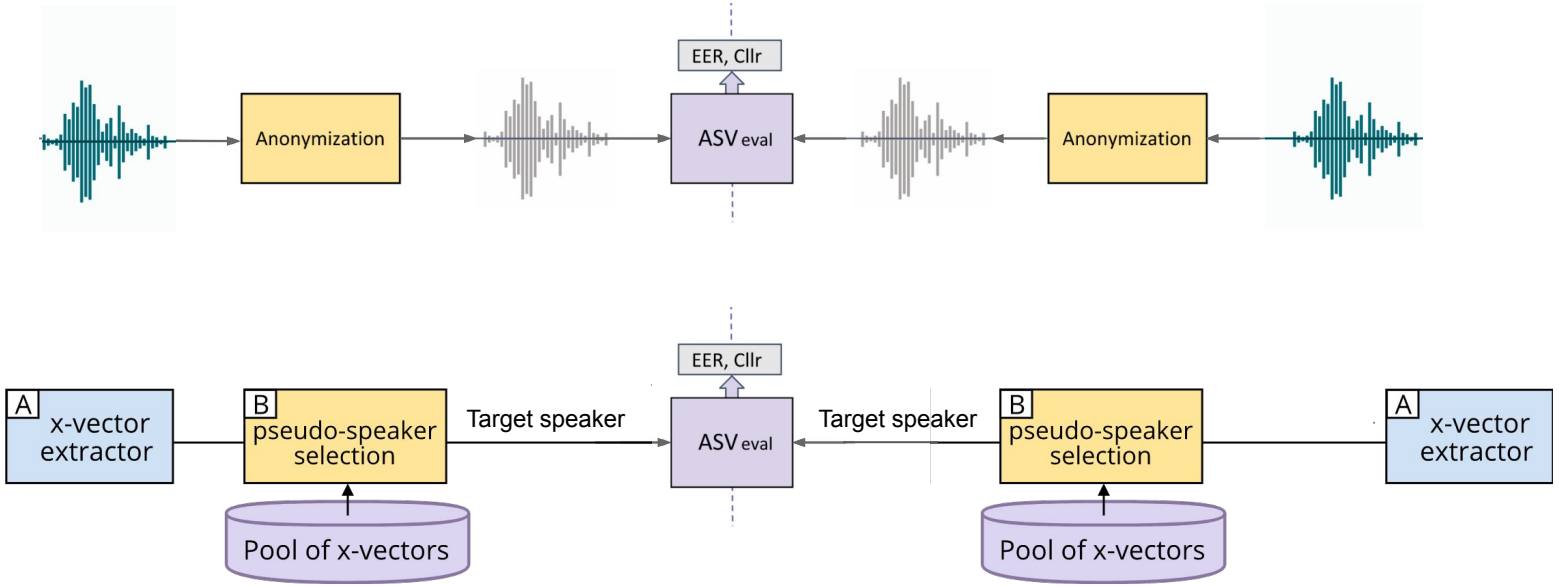
6. Champion, Pierre, Thomas Thebaud, et al., « On the invertibility of a voice privacy system using embedding alignment », *in*: *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021

The impact of the target speaker in VC anonymization

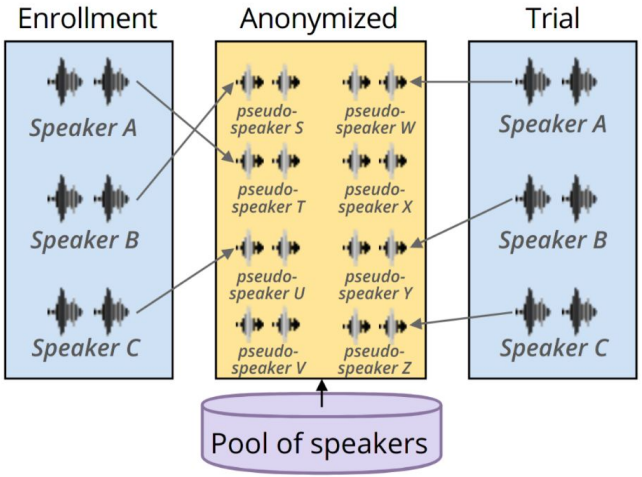
Evaluation procedure

For the *librispeech test dataset*, we **anonymize every utterances independently** using the target selection algorithm.

The voice conversion anonymization system is the VoicePrivacy 2022 baseline system (Unified HiFi-GAN NSF model).



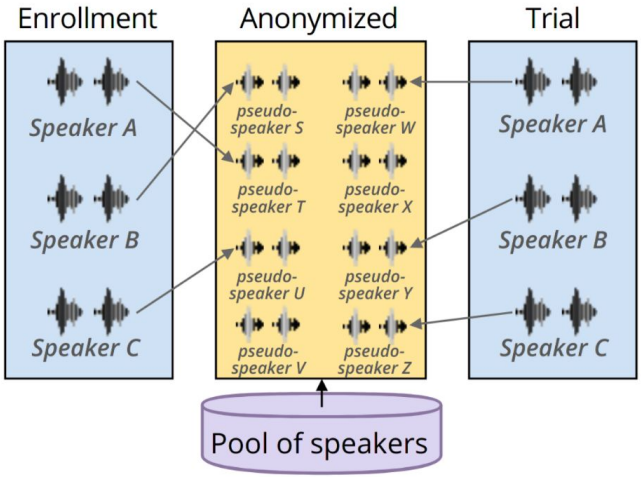
Baseline Voice Privacy: Speaker modification



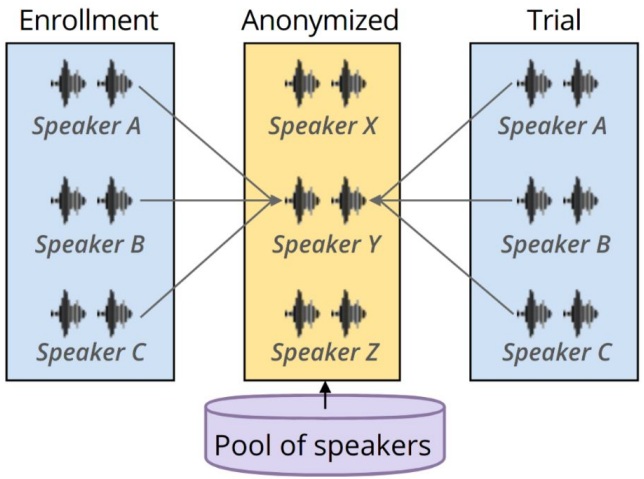
[1,2] Pseudo random target speaker selection
Any to Any voice conversion

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



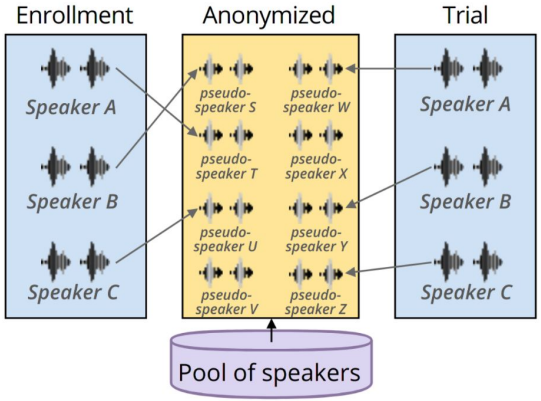
[1,2] Pseudo random target speaker selection
Any to Any voice conversion



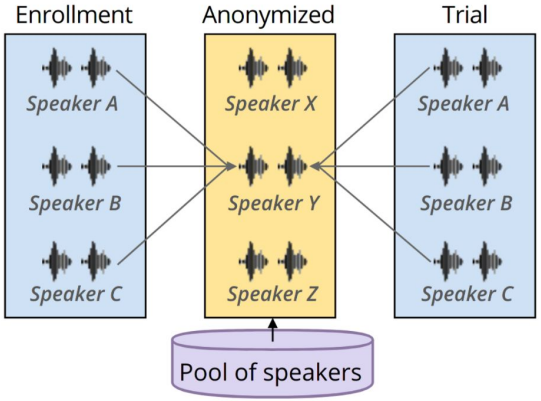
[3] Unique speaker target selection
Any to One voice conversion

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.

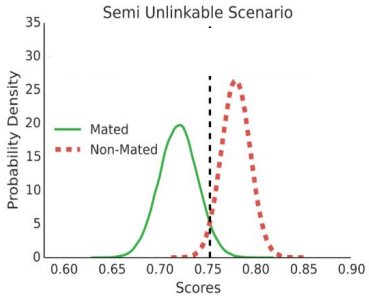
Voice Privacy: Speaker modification



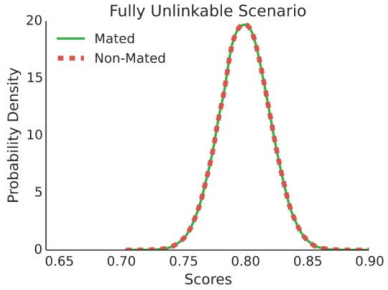
Pseudo random target speaker selection
Any to Any voice conversion



Unique speaker target selection
Any to One voice conversion



- Not all target selection algorithm fulfill unlinkability



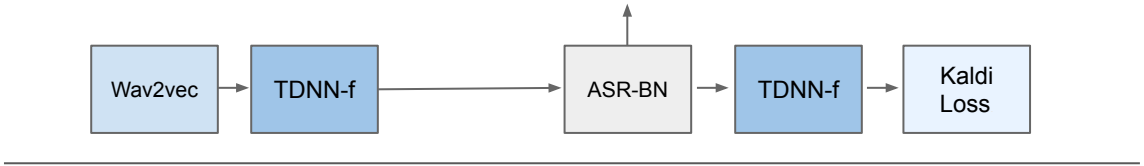
- Fulfill unlinkability
- Symplicity

ASR-BN 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

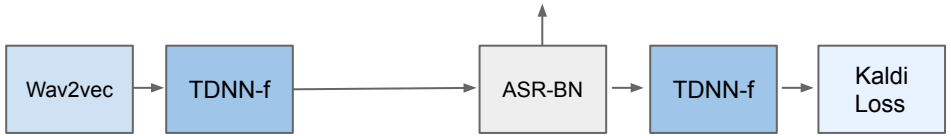
  => extract **continuous** ASR-BN



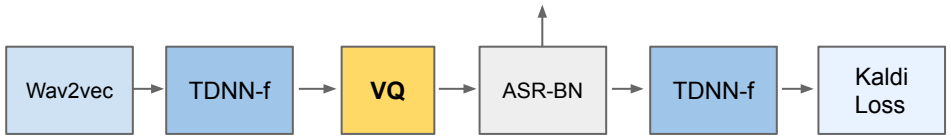
ASR-BN extractor: acoustic model

Explored with multiple acoustic model:

- **Wav2vec 2.0** pre-trained with VoxPopuli
  Wav2vec 2.0-TDNN-f trained with librispeech train-100
  => extract **continuous** ASR-BN

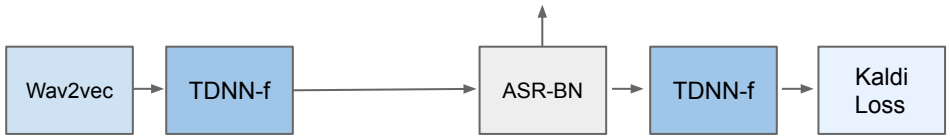


- **Wav2vec 2.0-TDNN-f** trained with librispeech train-100
  + **vector quantization layer**
  => extract **discrete** ASR-BN



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

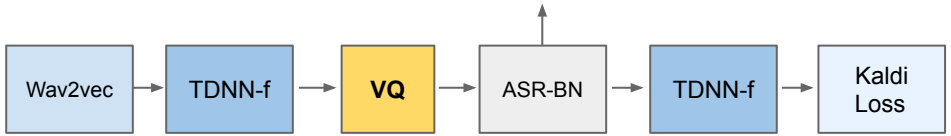
Wav2vec 2.0-TDNN-f

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 ASR-BN 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 scores:
50% EER
0% WER

Wav2vec 2.0-TDNN-f VQ

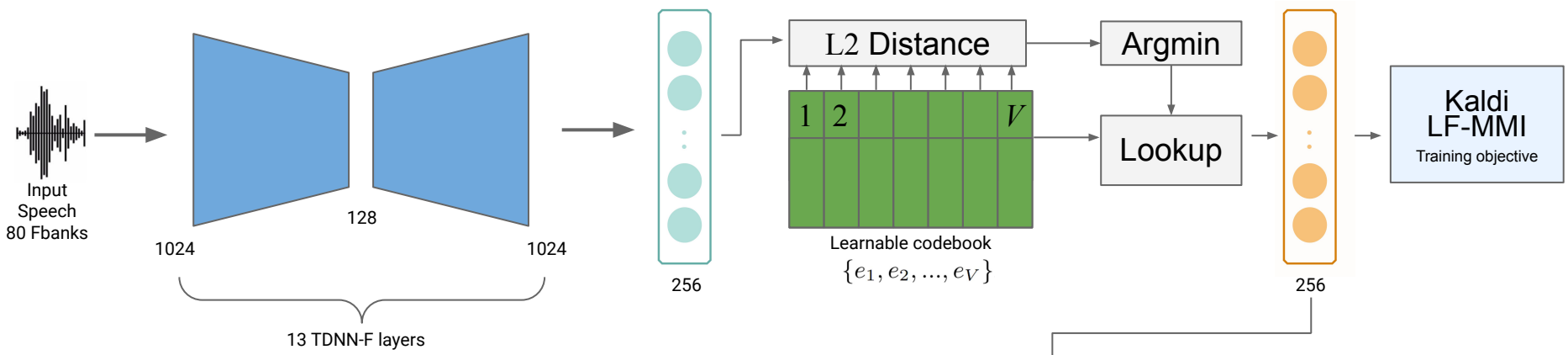
Higher privacy improvement for both datasets

Same utility improvement for both datasets

Vector quantization applies a constraint on the ASR-BN representation, making them more private without losing much utility

Vector quantized ASR-BN extractor

Codebook E, with size V $E = \{e_1, e_2, ..., e_V\}$
Vector quantization replaces the bottleneck representation by its nearest neighbor in the codebook E.
The size V, regulates the constraint



Constrains the representation space with the V parameter

256 dimension **discrete** bottleneck representation

Differentially Private Speaker Anonymization

The ε we report for BN features is frame-level and should be multiplied by the utterance length to obtain an utterance-level differential privacy guarantee

-> DP composition which makes the utterance level guarantee decrease linearly with the utterance length

Differentially Private Speaker Anonymization

ASR-BN extractor wav2vec 2.0: Bottleneck feature from an acoustic model

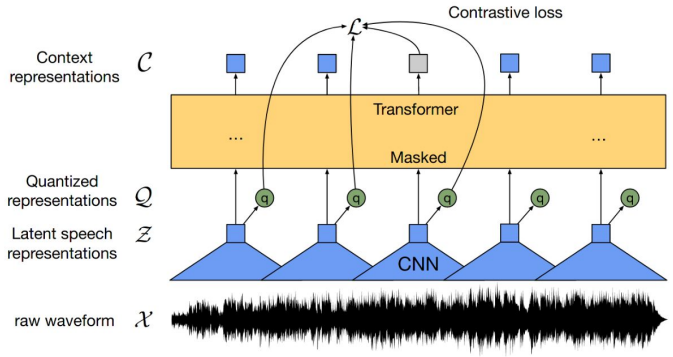
- Only trained on LibriSpeech train-clean-100
- Hypothesize that **privacy improvement** comes from the **vector quantification** while the **utility loss** comes from the **small size of the network and training data**

Solution:

Replaces fbanks by wav2vec 2.0 features

Wav2vec 2.0 pre-trained on 24.1K hours of unlabeled multilingual west Germanic speech from VoxPopuli

wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations



ISO/IEC international Standard 24745 on biometric data protection

Some guidelines:

- Confidentiality → Security / Cryptography

  "...property that **protects** information **against unauthorized access** or disclosure..."

- Integrity → Utility

  "...property of safeguarding the **accuracy and completeness of assets**."

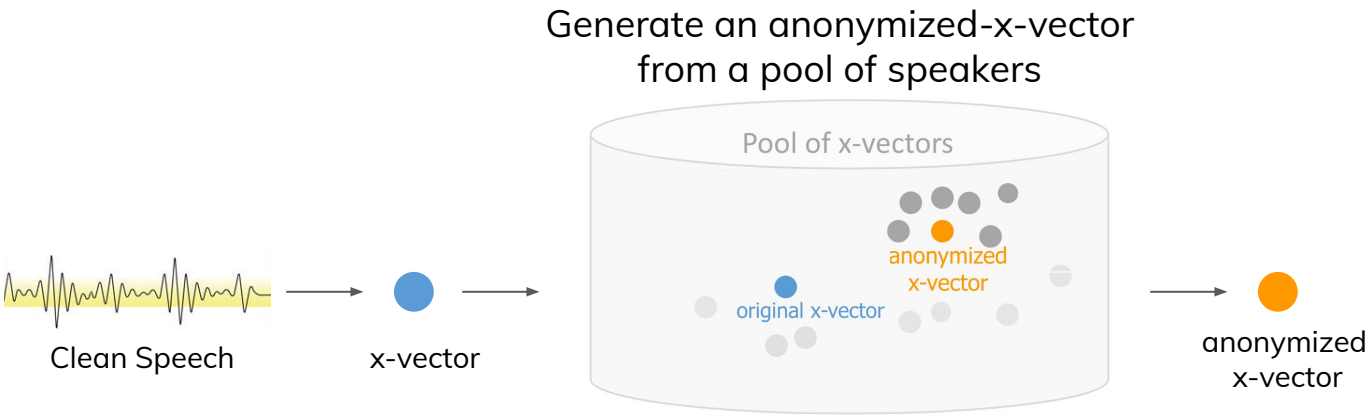
- Irreversibility → Privacy

  "To **prevent** the use of biometric data for **any purpose other** than originally intended, biometric data shall be processed by **irreversible transforms before storage.**

- Unlinkability → Privacy

  "The stored biometric references shall **not be linkable across applications or databases**".

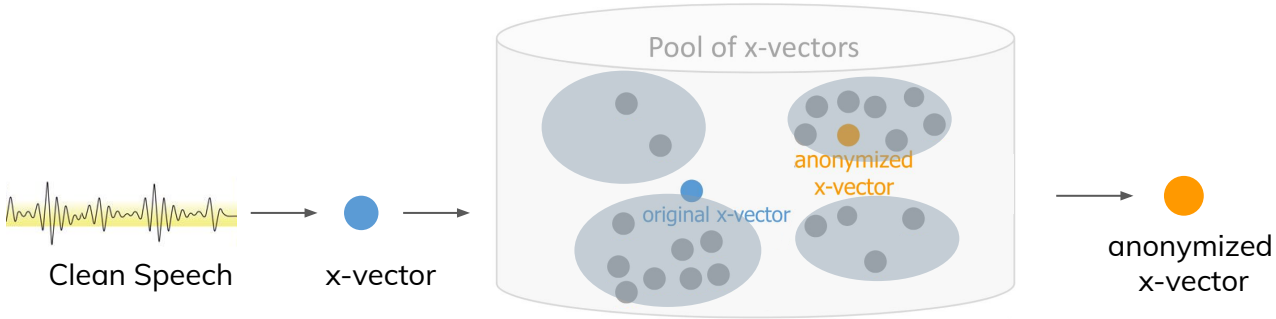
VPC x-vector selection



VPC Baseline target selection:
- Choose 200 x-vectors farthest from the original one
- Choose 100 of the 200 randomly
- Average of the 100 to obtain the anonymized x-vector

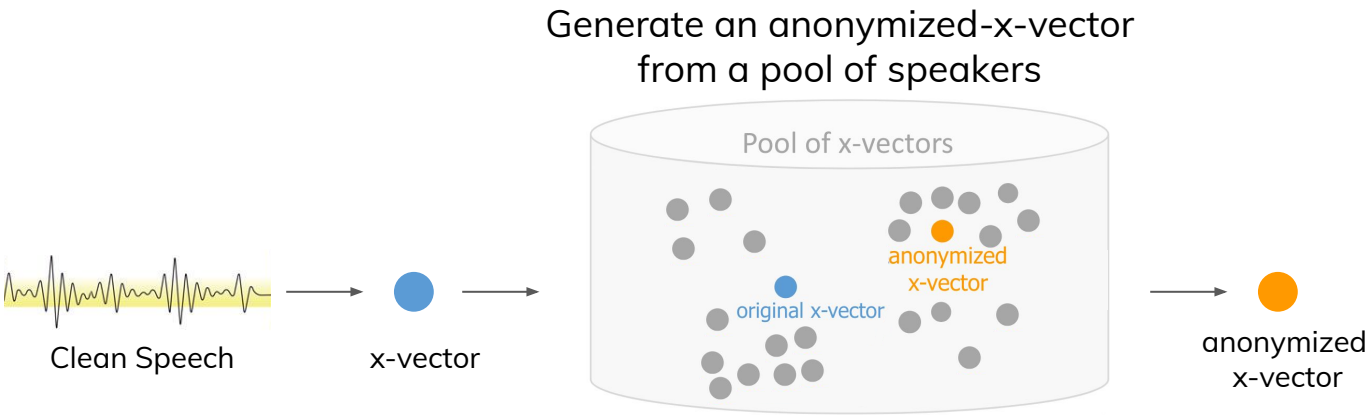
From the paper: The method enables the distance between the anonymized speaker and the original input speaker to be flexibly controlled.

X-vector target selection - dense target selection



Dense target selection:
- Create dense x-vector clusters
- remove the clusters of the source x-vector
- Randomly select one cluster from the 10 largest
- Choose haft of the members randomly
- Average of the remaining to obtain the pseudo-speaker (anonymized x-vector)

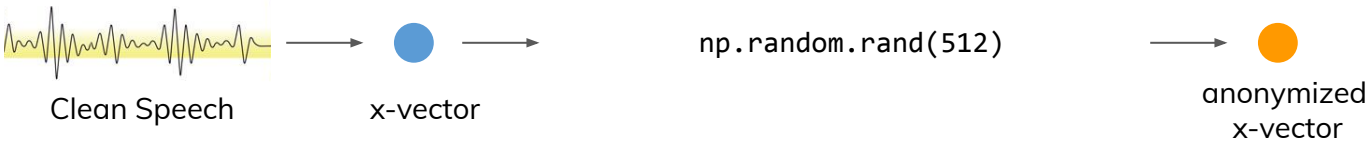
X-vector target selection - Random from pool



Dense target selection:
- Randomly select one x-vector from the pool

X-vector target selection - Random from uniform distribution

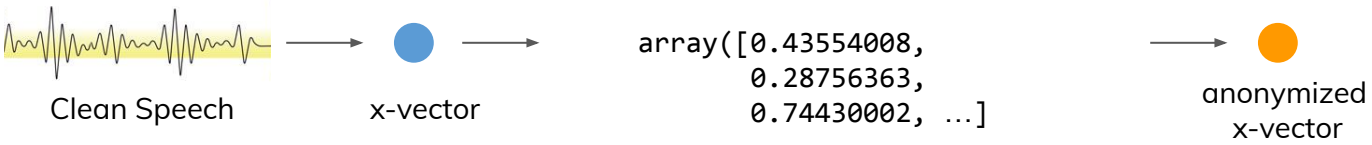
Generate an anonymized-x-vector
from random generator



Target selection:
- Randomly select one x-vector from a uniform distribution

X-vector target selection - constant

Select a single anonymized-x-vector
from everyone



Target selection:
- Select one x-vector

Evaluation of Speaker Anonymization on Emotional Speech

Results form the paper

Table 1: *WER and UAR results on IEMOCAP. The LibriSpeech results from VPC are presented for comparison purposes. The first line shows the WER and UAR results when no anonymization is performed. The second line shows the corresponding results when speech is anonymized and evaluated using an ASR system retrained on anonymized speech and an Informed attacker scenario.*

	$WER_{\%}$		$UAR_{\%}$
	LibriSpeech	IEMOCAP	IEMOCAP
Original speech data Model trained on original speech	4.15	34.62	44.48
Anonymized speech data Model trained on anonymized speech	4.77	38.97	37.92
Difference Anonymized / Original	15% degradation	13% degradation	15% degradation

Hubert Nourtel, Pierre Champion, Denis Jouvet, Anthony Larcher, Marie Tahon. Evaluation of Speaker Anonymization on Emotional Speech. *SPSC 2021 - 1st ISCA Symposium on Security and Privacy in Speech Communication*, Nov 2021, Virtual, Germany.

A Comparative Study of Speech Anonymization Metrics

Results form the paper

Each utterance of speaker A has been randomly mapped to the left or the right cluster, while the utterances of speaker B have been mapped to the center cluster. The resulting score distributions match the non-mated in-between case above. As expected, the two metrics strongly disagree: $D_{sys\leftrightarrow} = 0.99$ (low privacy) and $C_{min\ llr} = 0.81$ (high privacy).

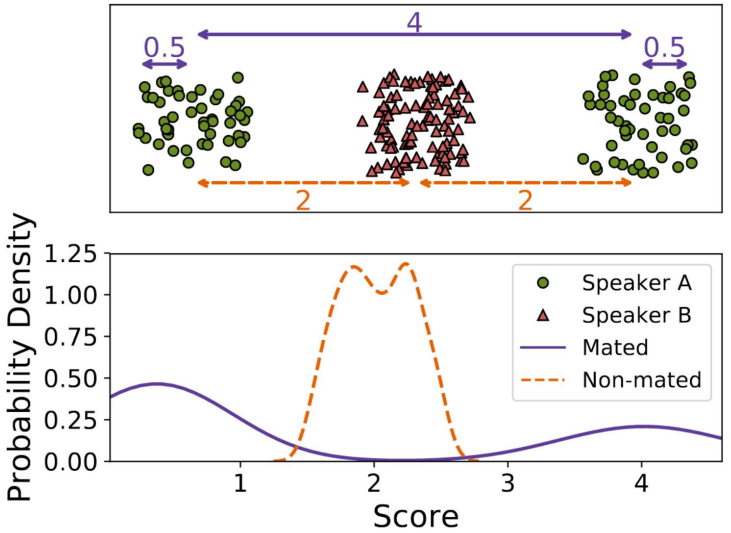


Figure 3: *Simulated 'non-mated in-between' data. Top: x-vectors visualized in 2D. Bottom: resulting score distributions.*

Mohamed Maouche, Brij Mohan Lal Srivastava, Nathalie Vauquier, Aurélien Bellet, Marc Tommasi, et al..
A comparative study of speech anonymization metrics. *INTERSPEECH 2020*, Oct 2020, Shanghai, China.

A Comparative Study of Speech Anonymization Metrics

Results form the paper

The comparison on real speech data processed via 4 anonymization techniques with different target selection strategies and with 9 attackers suggests that these metrics behave similarly

$C_{\text{min llr}}$ becomes less informative than $D_{\text{sys}\leftrightarrow}$ when the mated scores are lower or interleaved with non-mated scores

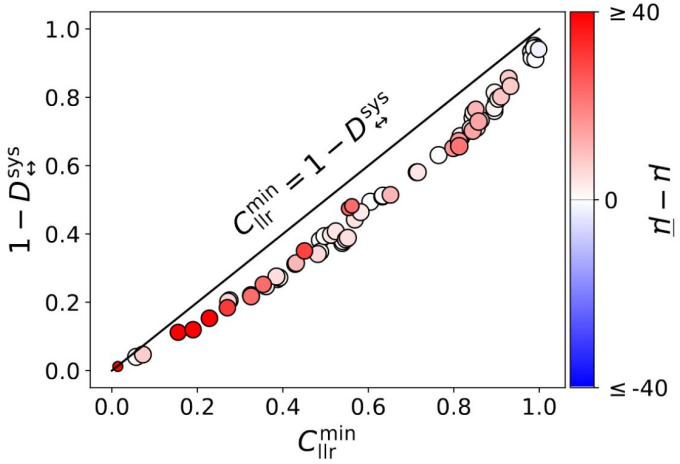


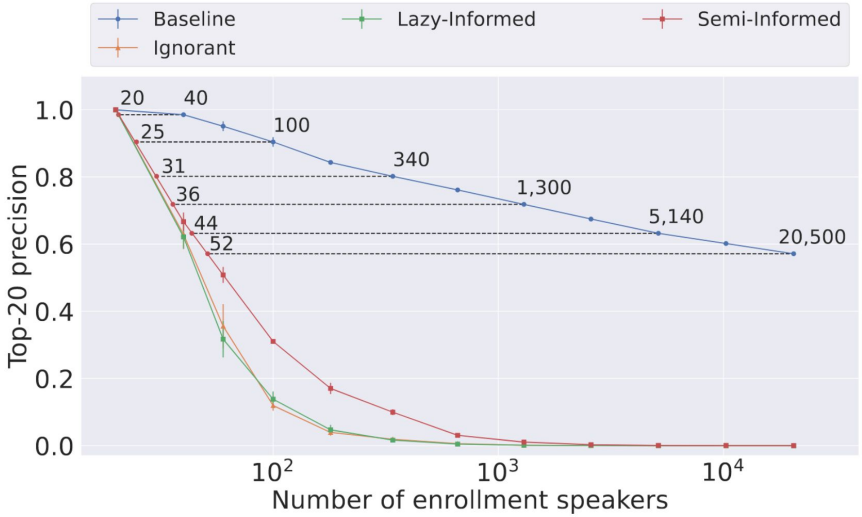
Figure 5: $C^{\text{min}}_{\text{llr}}$ *vs.* $1 - D^{\text{sys}}_{\leftrightarrow}$ *on real data. The color scale* $\mu - \bar{\mu}$ *is the difference of the means of mated and non-mated scores.*

Each dot corresponds to one of the 72 combinations of anonymization techniques, target selection strategies, attacker knowledge levels, and linkage functions

Mohamed Maouche, Brij Mohan Lal Srivastava, Nathalie Vauquier, Aurélien Bellet, Marc Tommasi, et al.. A comparative study of speech anonymization metrics. *INTERSPEECH 2020*, Oct 2020, Shanghai, China.

Privacy and utility of x-vector based speaker anonymization

Results form the paper



(b) $k = 20$

Fig. 15. Top-k ASI precision of *Ignorant*, *Lazy-Informed* and *Semi-Informed* attackers as a function of the number of speakers, compared to original speech (Baseline). The numbers of speakers needed to achieve an equivalent drop in precision before vs. after anonymization are highlighted.

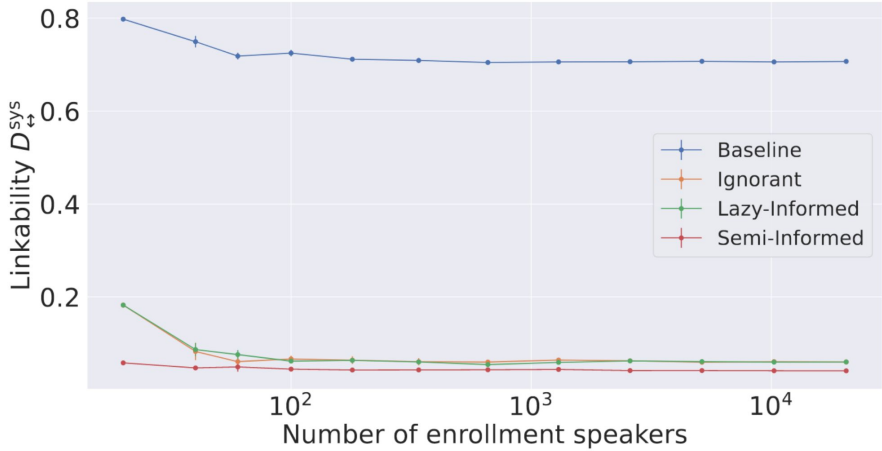


Fig. 13. Open-set ASV performance of *Ignorant*, *Lazy-Informed* and *Semi-Informed* attackers as a function of the number of enrolled speakers, compared to original speech (Baseline).

Brij Mohan Lal Srivastava, Mohamed Maouche, Md Sahidullah, Emmanuel Vincent, Aurélien Bellet, et al.. Privacy and utility of x-vector based speaker anonymization. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2022

Results form the paper

New data, h	Word error rate %	
	Original speech	Transformed speech
0	13.6	13.6
10	12.8	12.7
20	12.5	13.3
30	12.5	13.1
40	12.4	13.1
50	11.9	13.0
60	11.9	13.0
70	12.1	12.6
80	12.2	12.7
90	11.8	12.4
100	11.8	12.8

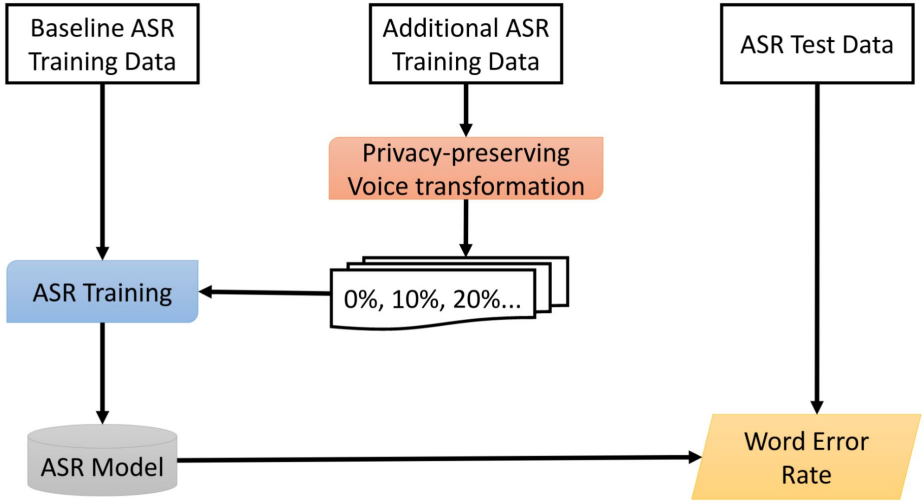


Figure 2. Experimental setup for the evaluation of the ASR trained using privacy-transformed data.

Additional data improve the WER and the difference between adding transformed (anonymized) and untransformed (original) data is small (5% relative between best results of both methods).

Askars Salimbajevs. Using privacy-transformed speech in the automatic speech recognition acoustic model training. 9th International Conference on Human Language Technologies - the Baltic Perspective (Baltic HLT 2020), Sep 2020, Kaunas, Lithuania.

Open question: can we train self supervised (wav2vec2) models on anonymized speech?

Results form the paper

Table 4: *ASR and F_0-PCC results for one-shot VC.*

Methods	CER	WER	F_0-PCC
Source (Oracle)	3.5%	9.0%	1.0
AutoVC	15.7%	30.5%	0.455
AdaIN-VC	27.1%	47.1%	0.346
VQVC+	35.5%	59.5%	0.237
VQMIVC (proposed)	**14.9%**	**29.3%**	**0.781**
w/o MI (proposed)	38.0%	62.1%	**0.781**

VoicePrivacy 2020 baseline 28.2 %

Pearson correlation coefficient (PCC)
ASR system not retrained on transformed speech

Disong Wang, Liqun Deng, Yu Ting Yeung, Xiao Chen, Xunying Liu, Helen Meng. VQMIVC: Vector Quantization and Mutual Information-Based Unsupervised Speech Representation Disentanglement for One-shot Voice Conversion. Interspeech 2021

Vector Quantization and Mutual Information-Based Unsupervised Speech Representation for Voice Conversion

Architecture form the paper

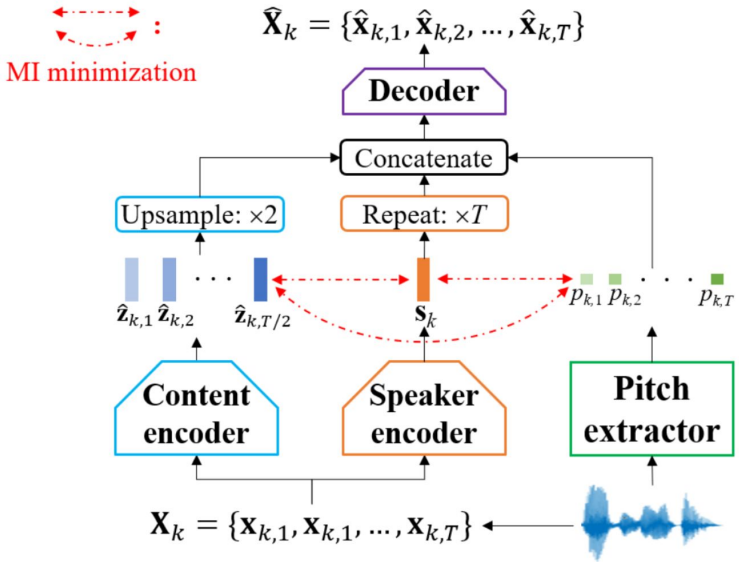


Figure 1: *Diagram of the proposed VQMIVC system.*

Disong Wang, Liqun Deng, Yu Ting Yeung, Xiao Chen, Xunying Liu, Helen Meng. VQMIVC: Vector Quantization and Mutual Information-Based Unsupervised Speech Representation Disentanglement for One-shot Voice Conversion. Interspeech 2021

Evaluation setup

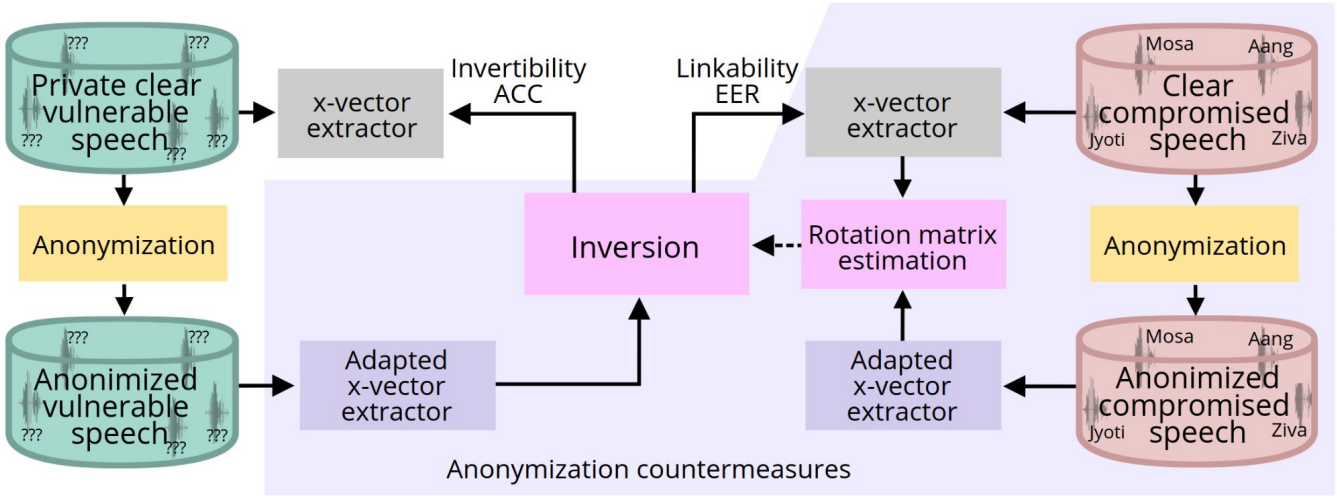


Figure 6.1 – Illustration of training the rotation matrix and inverting anonymized x-vectors. Invertibility and linkability measurements are then performed with clear and inverted x-vectors.

Evaluation setup

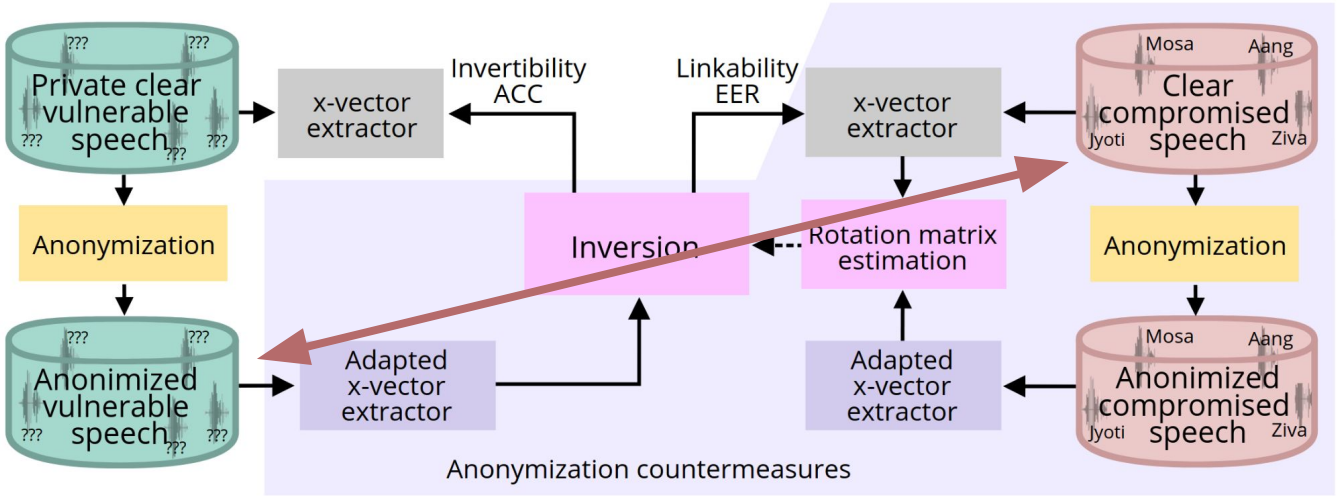


Figure 6.1 – Illustration of training the rotation matrix and inverting anonymized x-vectors. Invertibility and linkability measurements are then performed with clear and inverted x-vectors.

Evaluation setup

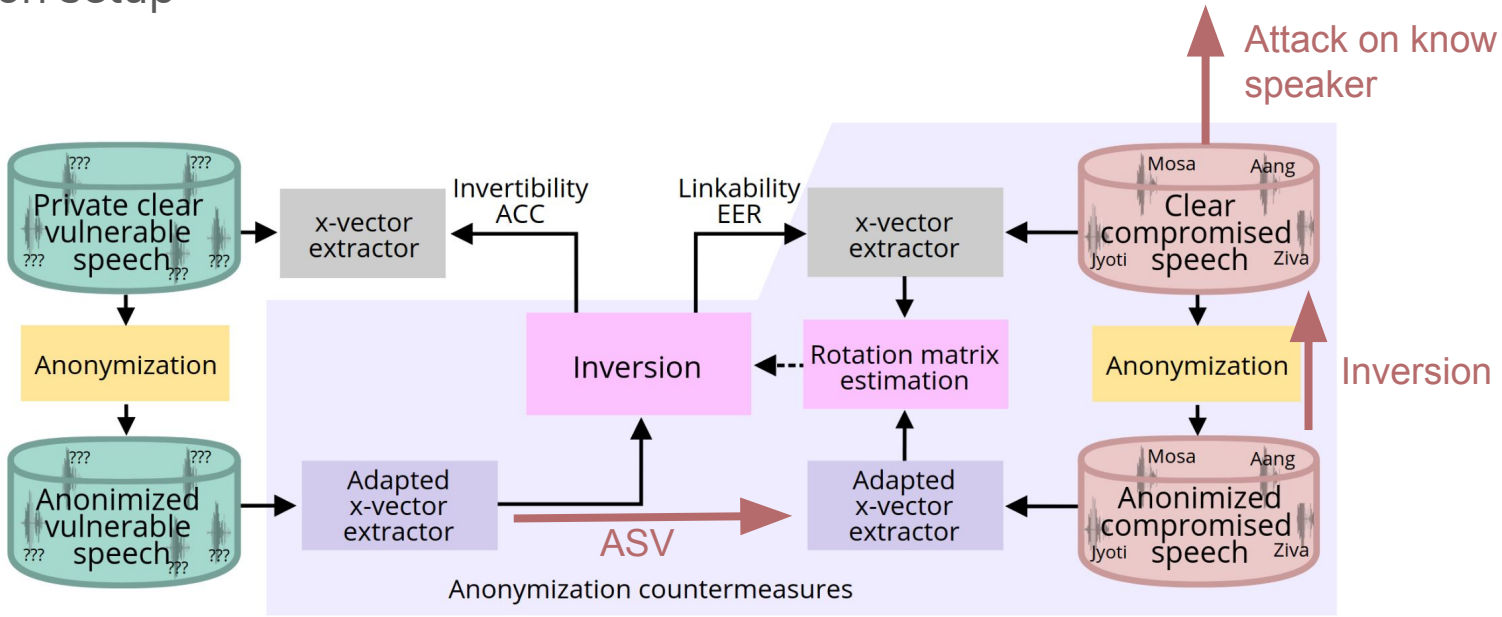


Figure 6.1 – Illustration of training the rotation matrix and inverting anonymized x-vectors. Invertibility and linkability measurements are then performed with clear and inverted x-vectors.

Evaluation setup

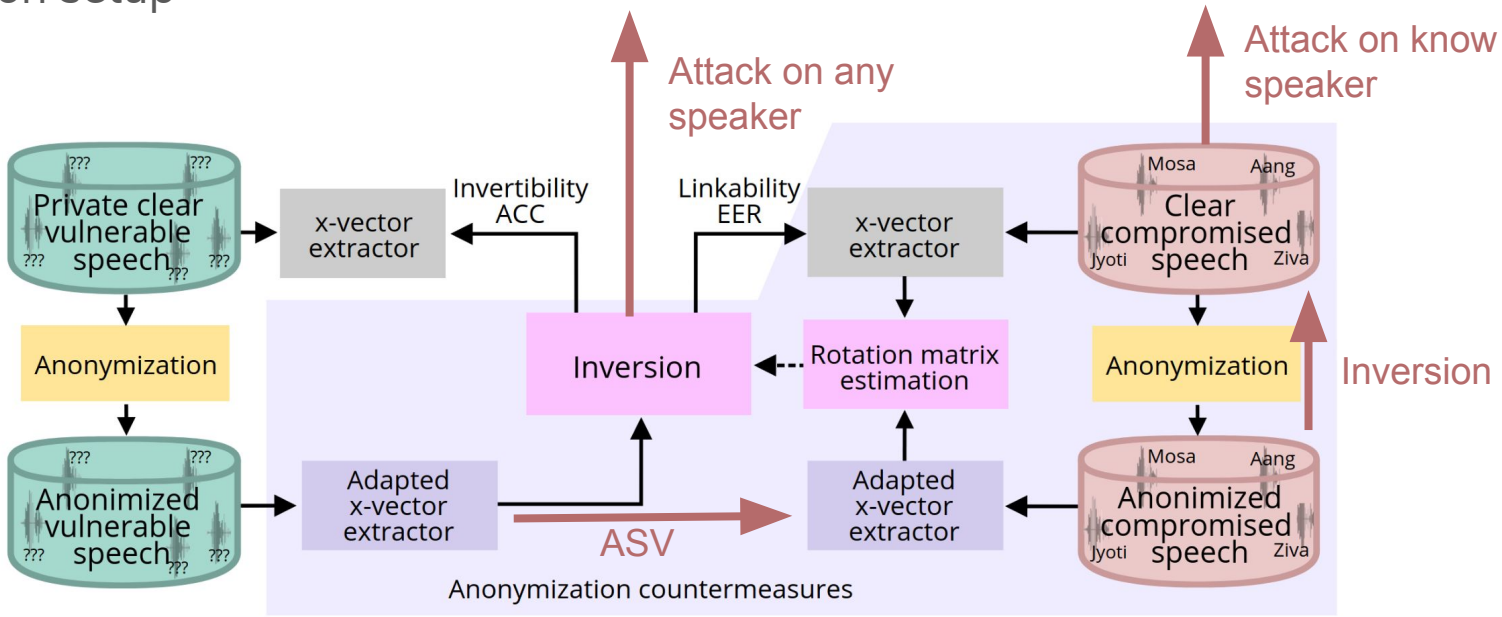


Figure 6.1 – Illustration of training the rotation matrix and inverting anonymized x-vectors. Invertibility and linkability measurements are then performed with clear and inverted x-vectors.

Evaluation setup

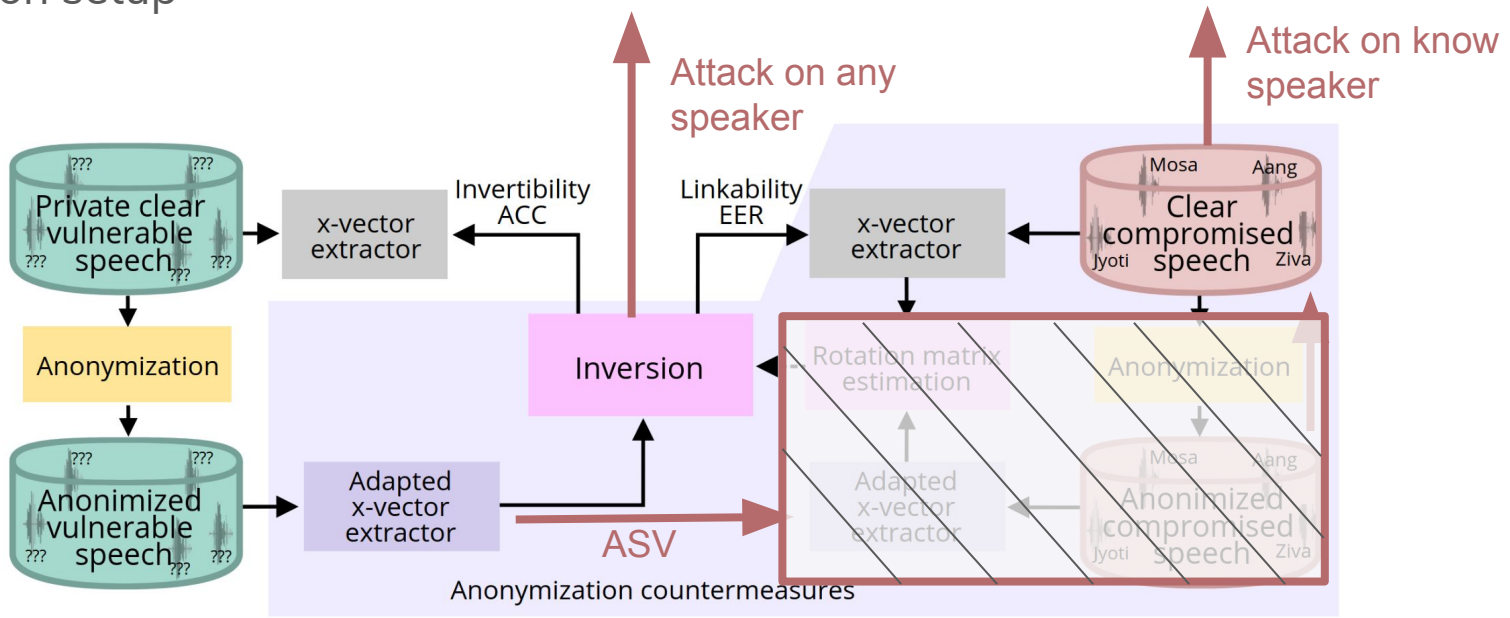


Figure 6.1 – Illustration of training the rotation matrix and inverting anonymized x-vectors. Invertibility and linkability measurements are then performed with clear and inverted x-vectors.

Invertibility evaluation using embedding alignment (Librispeech test)

Table 6.1 – Experimental results for the rotation-based invertibility attack scenarios on the VPC 2022 baseline. The scoring function for ASV linkability assessment is cosine similarity.

		EER ↑		ACC ↓	
		Female	Male	Female	Male
1	Clear speech	10.3	2.9		
2	White-box ASV	27.9	27.8		
3	Procrustes	19.5	21.4	67.7	51.8
4	Wasserstein Procrustes	20.1	22.9	64.7	50.0
5	Oracle Procrustes	17.1	12.0	98.3	97.0
6	Oracle Wasserstein Procrustes	17.7	12.8	99.0	97.2

Table 6.2 – Experimental results for the rotation-based invertibility attack scenarios on our fully quantization-based anonymization. The scoring function for ASV assessment is cosine similarity.

		EER ↑		ACC ↓	
		F	M	F	M
1	Clear speech	10.3	2.9		
2	White-box ASV	37.8	36.4		
3	Procrustes	32.3	35.0	32.7	24.7
4	Wasserstein Procrustes	37.9	36.7	23.7	15.9
5	Oracle Procrustes	29.6	29.4	85.3	84.1
6	Oracle Wasserstein Procrustes	37.8	33.4	87.8	82.7

PPG vs Phoneme classes vs ASR-BN

Compared to phoneme classes, PPGs encode more information about the temporal and spectral characteristics of speech sounds. PPGs capture not only the presence or absence of phonemes, but also the relative timing, duration, and frequency content of speech sounds. For example, PPGs can distinguish between different vowel formants, which are important for distinguishing between similar-sounding vowels.

Additionally, PPGs can be used to model context-dependent variations in speech sounds. For example, the same phoneme may be pronounced differently depending on the surrounding phonetic context. PPGs can capture these context-dependent variations by modeling the conditional distribution of acoustic units given the preceding and following context.

The WER of the natural speech is 9.49% while those of the anonymized speech are between 10% and 30% when using the ASR-BN from the 6th layer and between 25% and 45% when using the PPG from the softmax layer.

In: Fuming Fang, Xin Wang, Junichi Yamagishi, Isao Echizen, Massimiliano Todisco, Nicholas Evans, Jean-Francois Bonastre, Speaker Anonymization Using X-vector and Neural Waveform Models. 10th ISCA Speech Synthesis Workshop.

VQ transform results

Table 5.7 – Privacy and utility results for Vector Quantization-based anonymization, using a Wav2Vec-2.0 feature extractor. Three kinds of F_0 transformations were also tested, note that the linear shift is included by default in our models.

	Dataset	*LibriSpeech test-clean*			*VCTK test*		
	Method: ASR-BN + F_0 transformations	Privacy $D_{\leftrightarrow}^{\text{sys}} \downarrow$	Privacy EER ↑	Utility WER ↓	Privacy $D_{\leftrightarrow}^{\text{sys}} \downarrow$	Privacy EER ↑	Utility WER ↓
1.	Clear speech	0.93	4.1	4.1	0.93	2.7	12.8
2.	VPC 2022 baseline	0.67	13.5	5.1	0.49	20.6	13.0
8.	WAV2VEC2 TDNNF NO VQ	0.83	7.7	3.8	0.69	12.1	7.8
9.	WAV2VEC2 TDNNF VQ 48	0.57	17.5	4.5	0.34	28.0	10.0
10.	WAV2VEC2 TDNNF VQ 48 + F0 AWGN_{15dB}	0.44	23.4	4.6	0.12	40.8	10.3
11.	WAV2VEC2 TDNNF VQ 48 + F0 $\text{LP}_{\epsilon 1}$	0.46	22.5	4.6	0.30	30.2	9.9
12.	WAV2VEC2 TDNNF VQ 48 + F0 QUANT_{4bits}	0.45	23.0	4.4	0.14	39.8	9.9

Computation time

Model	Time
ASR-BN extractor (100h)	5h
ASR-BN extractor (600h)	30h
ASR-BN adversarial model (600h)	12h
VC model (100h)	48h
ASR_eval (360h)	20h
ASV_eval (360h)	5h
Data anonymization (test 80h and asv/asr train 360h)	5h-24h

Time for non adversarial experiments: ~85h
Time the adversarial experiment: ~130h
Time the radar experiment: ~1200h

Dataset

Table 3.1 – Statistics of the datasets.

	Dataset	Usage	# Speakers	# Utterances
Train	*LibriSpeech train-clean-100*	Linguistic extractor	251	28 539
	LibriSpeech train-other-500	Linguistic extractor	1 166	148 688
	LibriTTS train-clean-100	Speech synthesizer	247	33 236
	LibriTTS train-other-500	Pool of x-vectors	1 160	205 044
	VoxCeleb1,2	Speaker extractor	7 363	1 281 762
Test	*LibriSpeech test-clean*	Compromised speech	29	438
		Vulnerable speech	40	1 496
	VCTK test	Compromised speech	30	600
		Vulnerable speech	30	11 448
Eval	*LibriSpeech train-clean-360*	Train privacy/utility eval models	921	104 014