Speech anonymization

Emmanuel Vincent
Inria Nancy – Grand Est

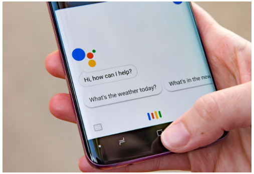
Acknowledgments

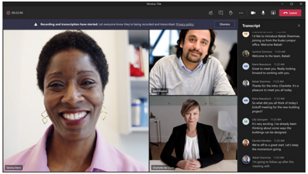
Joint work with

- **Inria**: B. M. L. Srivastava, M. Maouche, N. Vauquier, H. Nourtel, A. Bellet, M. Tommasi, D. Jouvet
- **LIA**: N. Tomashenko, J.-F. Bonastre, P.-G. Noé
- **EURECOM**: A. Nautsch, N. Evans, M. Todisco, J. Patino
- **LIUM**: P. Champion
- **NII**: X. Wang, X. Miao, J. Yamagishi
- **Saarland University**: D. Adelani, A. Davody, T. Kleinbauer, D. Klakow
- **Vector Institute**: A. S. Shamsabadi

Speech technologies

Most speech technologies process and (under certain circumstances) store speech data remotely for inference and training purposes.





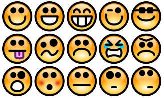
Which information is conveyed?

Speech conveys several pieces of information:

- **verbal content**:
  words, possibly including identifiers and private (phone number, preferences, etc.) or business information

- **speaker**:
  identity, age, gender, ethnic origin, etc.

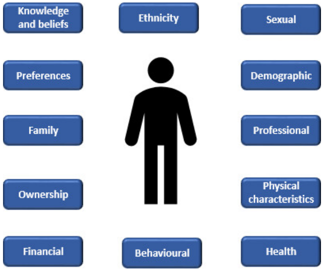
- **nonverbal content**:
  emotions, health status, etc.



- **acoustic environment**:
  acoustics, ambient noise, other speakers

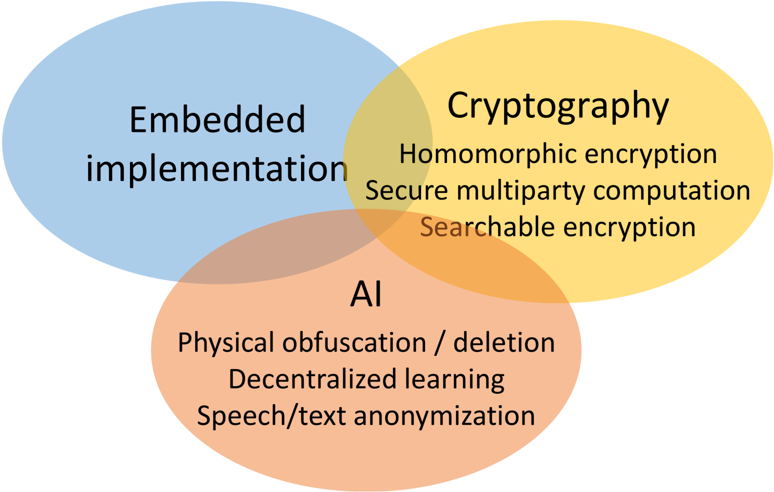
What are the risks?

- **Personal** or even **sensitive** data.
- Collection and processing governed by **privacy laws** such as the General Data Protection Regulation (GDPR) in Europe or the Privacy Act in the USA.
- Legal bases: **user consent** for one or more specific purposes, contractual or legal obligations, protection of vital interests, and public or legitimate interest.



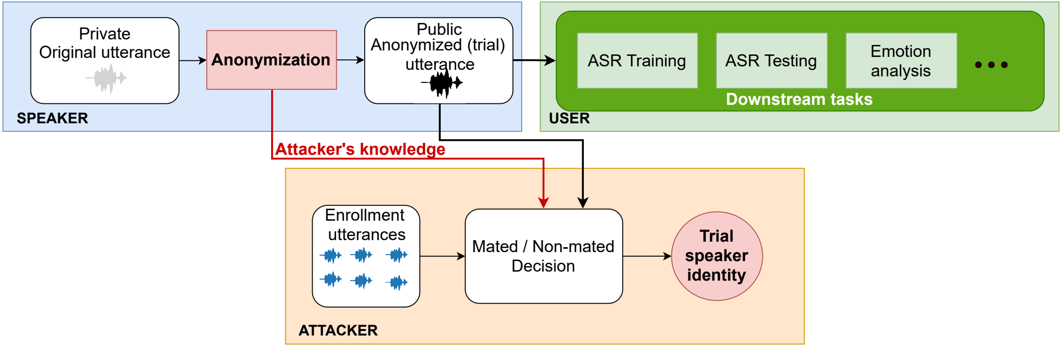
- In practice, users cannot always choose the purposes they accept or not.
- In some situations, **risks** may include
  - user profiling
  - user identification
  - voice cloning or information leakage in case of security breach

How to protect privacy?

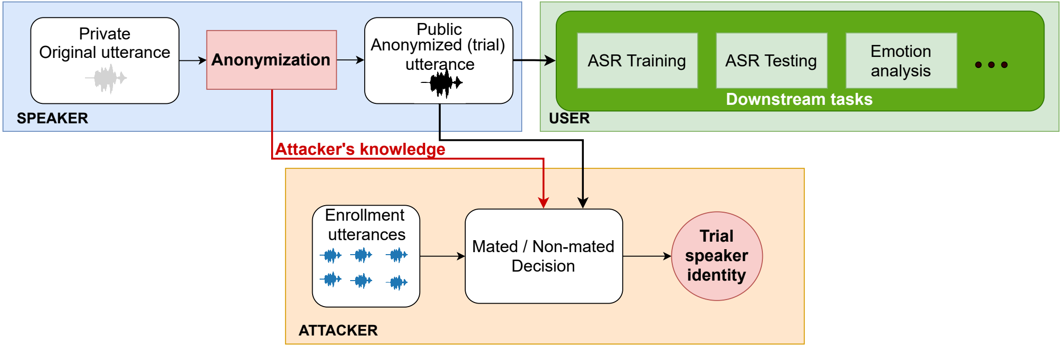


- Anonymization:
  - Transform speech to **hide speaker identity**
  - **Leave other information unchanged**, so that it's useful for downstream tasks
- Defines the goal, even when it's not achieved ($\neq$ strict legal definition)
- Achieving this goal requires:
  - **voice anonymization** (aka **de-identification**) by voice transformation/conversion,
  - hiding identifiable nonverbal attributes but preserving others (ASR+TTS not OK)
  - **verbal content anonymization**.
- Only approach compatible with privacy preservation at both training and test time. Can be complemented by encryption & decentralized learning.
- Assumption: **no metadata** (often does not hold in practice).

Threat model



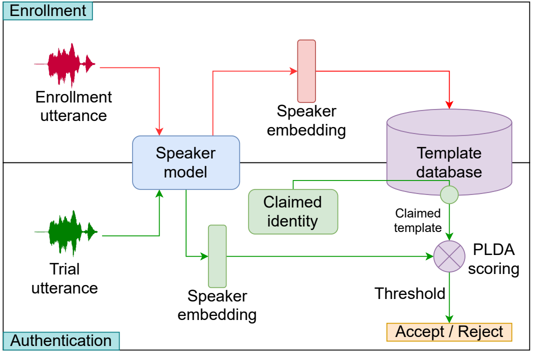
Threat model



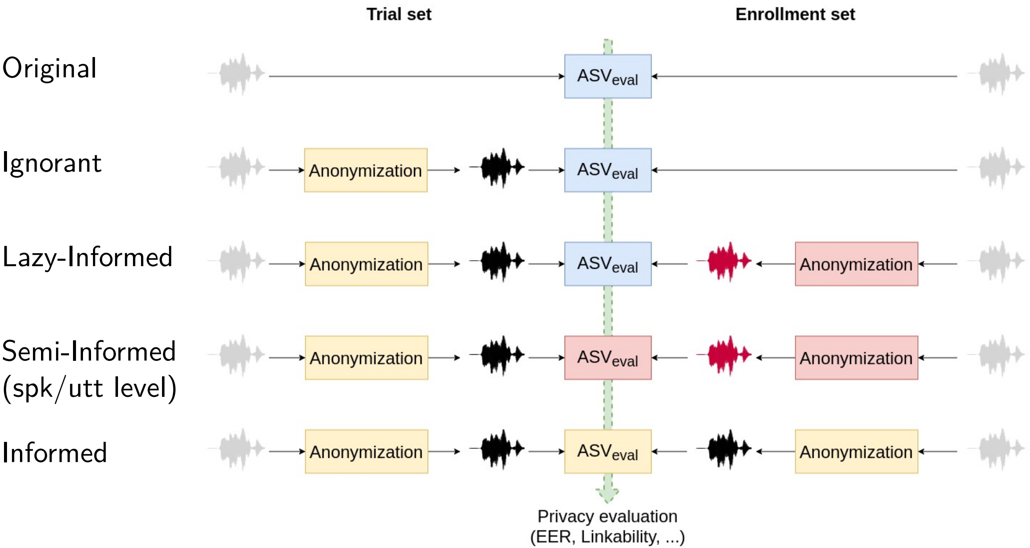
From now on, **focus on voice anonymization** by voice transformation or conversion.

Privacy assessment

- The success or failure of voice anonymization can be evaluated via **speaker verification**.
- In practice, speaker embeddings = x-vectors.
- Higher score $\Rightarrow$ greater chance of being from the same speaker

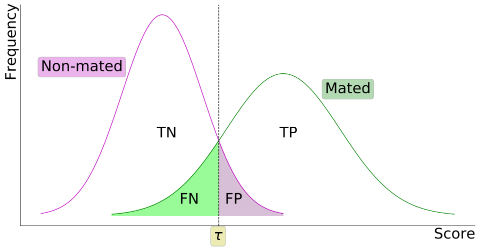


Attacker's knowledge



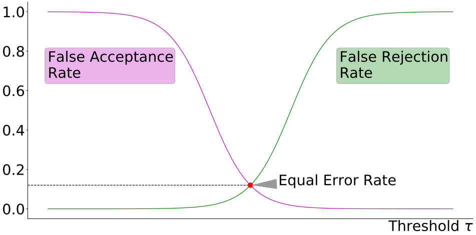
Privacy metrics

Compare same- and different-speaker score distributions with a threshold.



Derive the **equal error rate** (EER). Varies from 0 to 50%, higher is better.

Other metrics include **linkability** (varies from 0 to 1, lower is better) and ZEBRA.

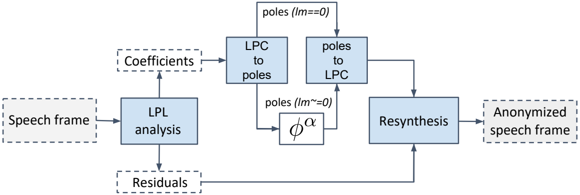


Simple transformation approaches such as

- **pitch shifting** (often used on TV/radio)

  Original -3 tone shift Multiple shifts

- **spectral envelope warping**
  - Baseline B2 of the VoicePrivacy 2022 Challenge
  - VoiceMask
  - VTLN

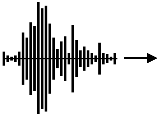


EER (Librispeech)

Attacker	Baseline-2	VoiceMask	VTLN
Original speech		4.3%	
Ignorant	26.2%	28.7%	27.4%
Semi-Informed (utt-level)	5.3%	5.0%	6.3%

Simple transformations **fail against non-ignorant attackers**.

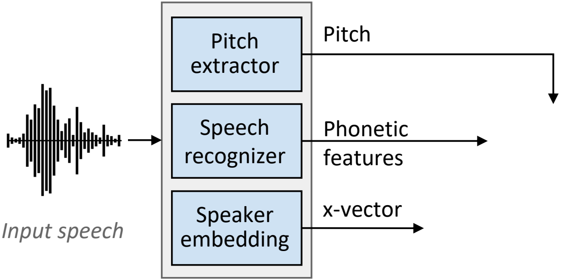
- Idea: **replace user's voice** by that of a target speaker
- Baseline B1.a of the VoicePrivacy 2022 Challenge



Input speech

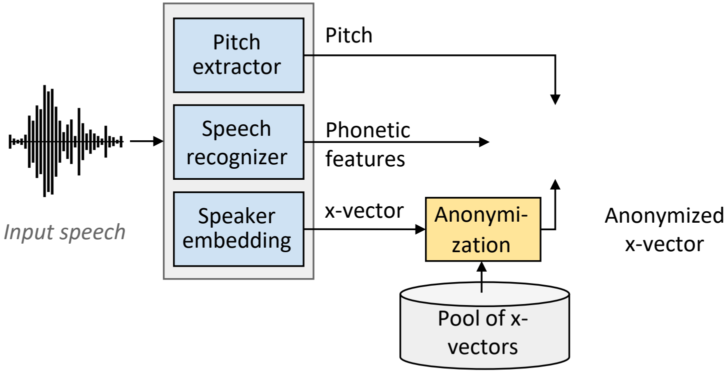
Phonetic features = bottleneck (BN)

- Idea: **replace user's voice** by that of a target speaker
- Baseline B1.a of the VoicePrivacy 2022 Challenge



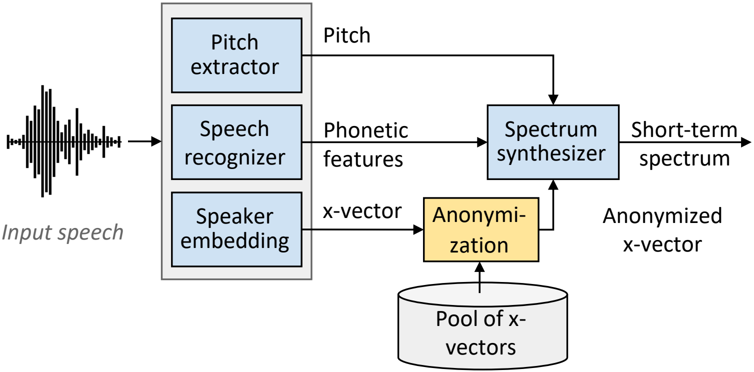
Phonetic features = bottleneck (BN)

- Idea: **replace user's voice** by that of a target speaker
- Baseline B1.a of the VoicePrivacy 2022 Challenge



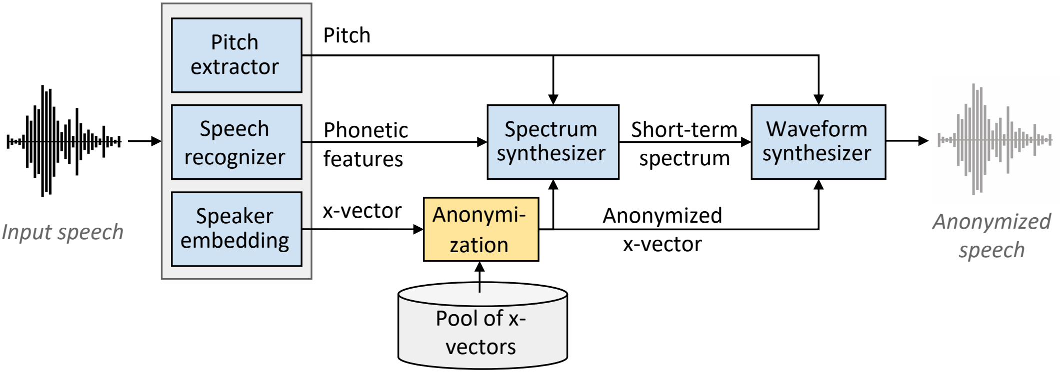
Phonetic features = bottleneck (BN)

- Idea: **replace user's voice** by that of a target speaker
- Baseline B1.a of the VoicePrivacy 2022 Challenge



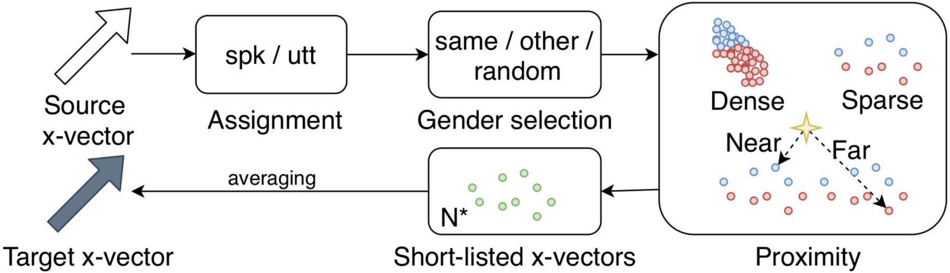
Phonetic features = bottleneck (BN)

- Idea: **replace user's voice** by that of a target speaker
- Baseline B1.a of the VoicePrivacy 2022 Challenge



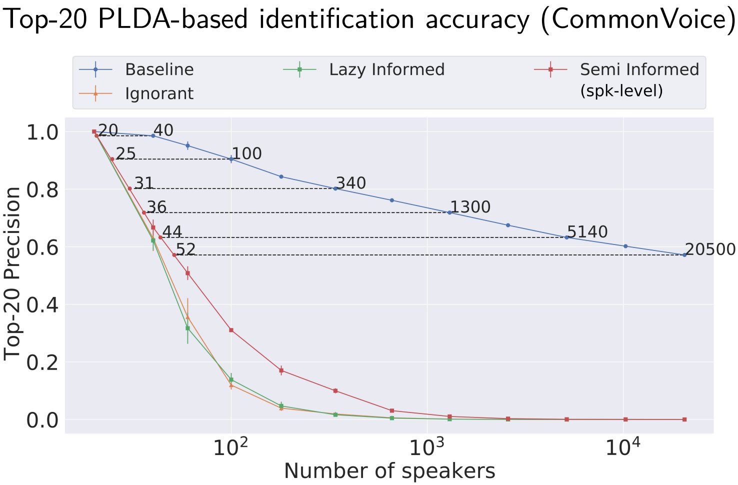
Phonetic features = bottleneck (BN)

- Target selection procedure:



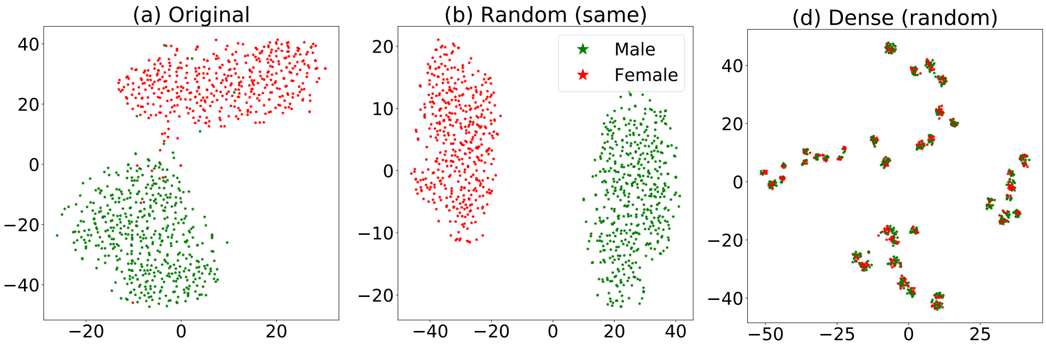
- Retained choice: random gender + dense

Original Modified



Re-identification risk $\rightarrow 0$ with **2,000+ speakers** with best (Semi-Informed) attack.

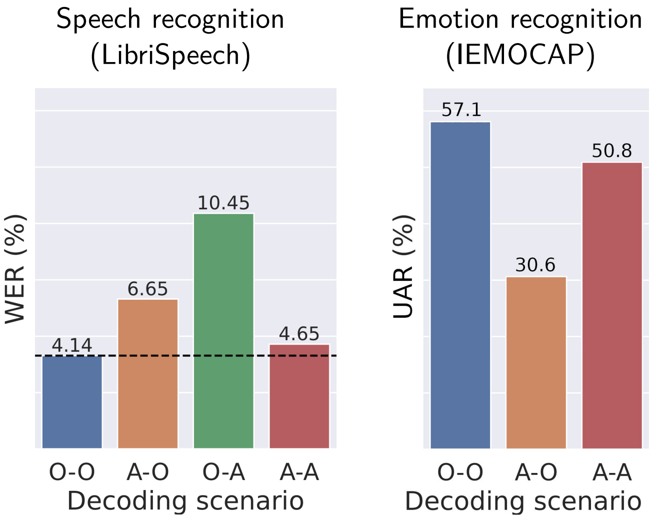
Besides identity, voice conversion can **hide** (or not) **speaker traits** such as gender.



98.6% 100.0% 53.3%

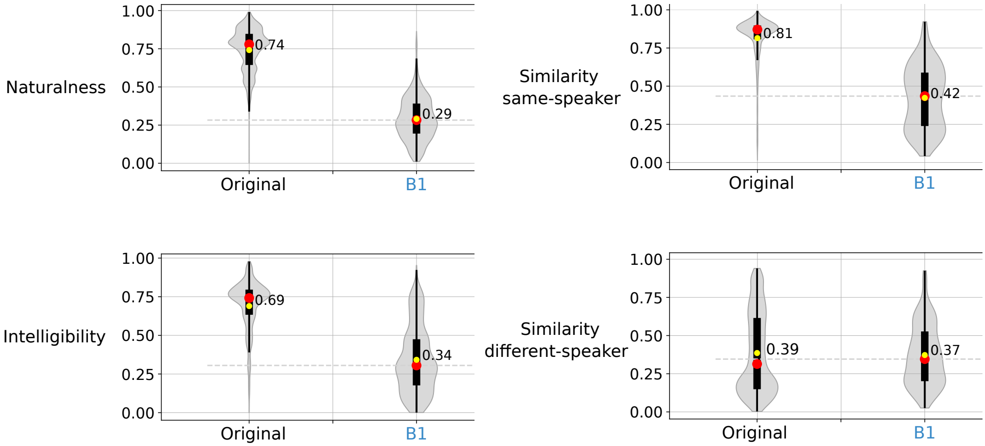
Gender classification accuracy

Voice conversion — Utility results



Small or negligible loss of utility after **retraining on anonymized data** (A-A).

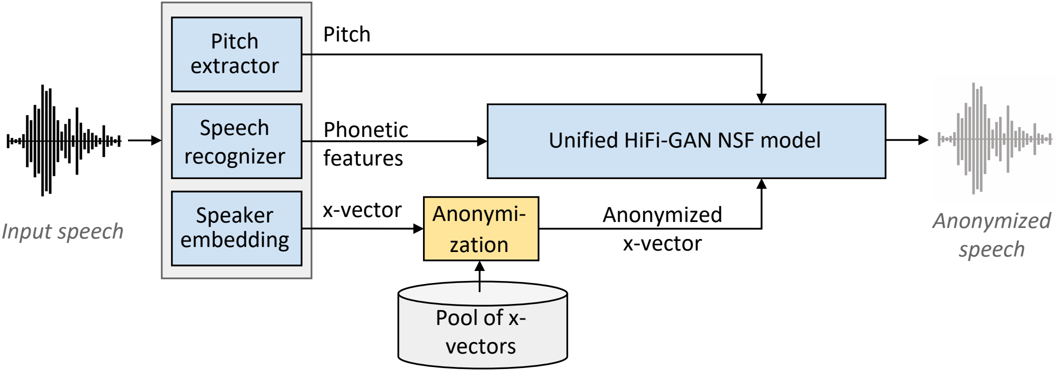
Voice conversion — Subjective results



Human listeners are **easily fooled**.

Voice conversion — Limitations (1)

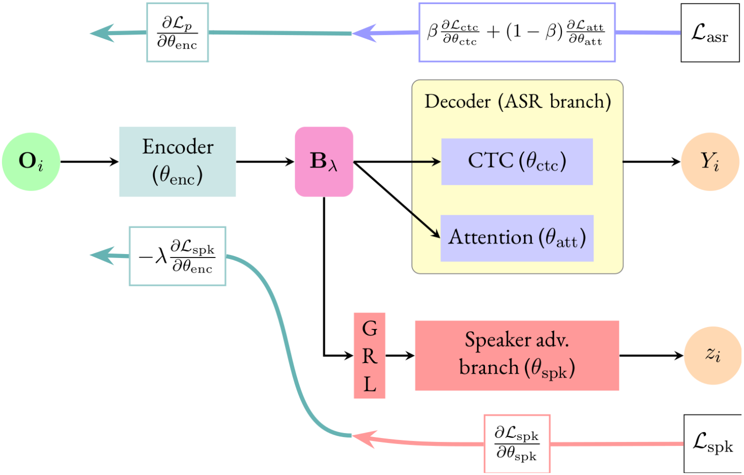
- Two-step synthesis in Baseline B1.a yields low naturalness/intelligibility
- Idea: simply replace by a **better synthesis model**
- Baseline B1.b of the VoicePrivacy 2022 Challenge



Voice conversion — Limitations (2)

- Key limitation:
  - pitch and phonetic features contain **residual speaker information**
  - this information remains after resynthesis and can be captured by the attacker
- Some ideas explored:
  - better input features (e.g., wav2vec2.0)
  - better F0/BN models, trained on more data
  - **adversarial representation learning**
  - attribute-aligned representation learning (e.g., attention-based)
  - vector quantization
  - **additive noise** (local differential privacy)
  - slicing utterances into shorter segments

Adversarial learning of phonetic features for speech recognition

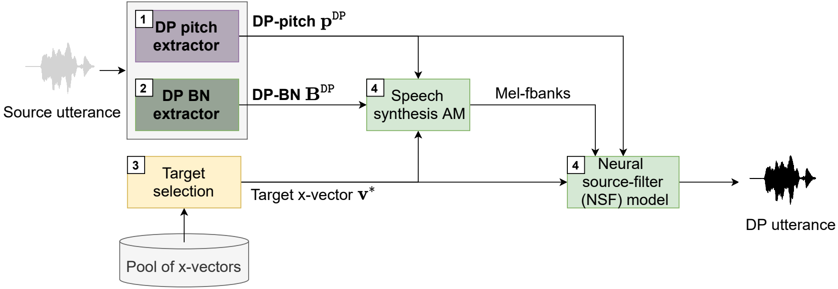


Accuracy, EER and WER (Librispeech)

	Spec. feat.	$\lambda = 0$	$\lambda = 0.5$	$\lambda = 2$
Speaker identification accuracy	93.1%	46.3%	6.4%	2.5%
Speaker verification EER	5.7%	23.1%	22.0%	19.6%
Speech recognition WER	–	10.9%	12.5%	12.5%

Adversarial learning **generalizes poorly to unseen speakers**.

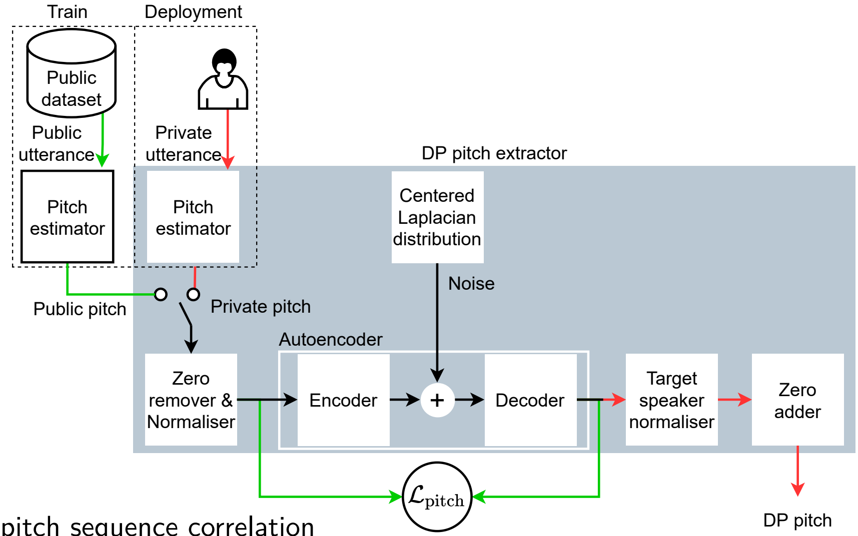
DP voice anonymization — Overall approach



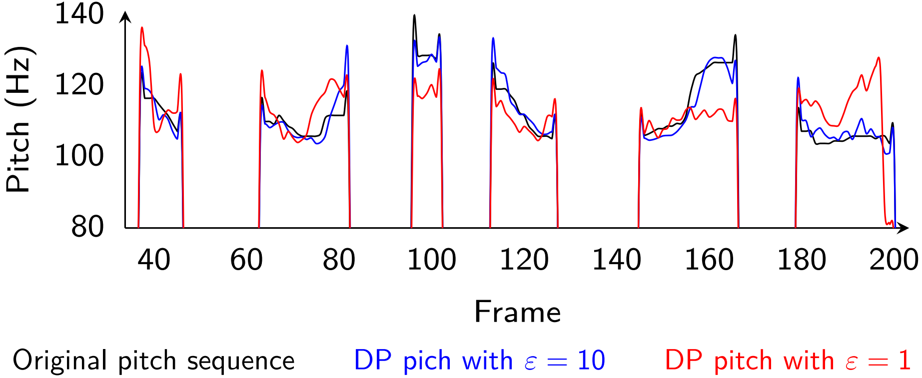
Local differential privacy (DP) principle:

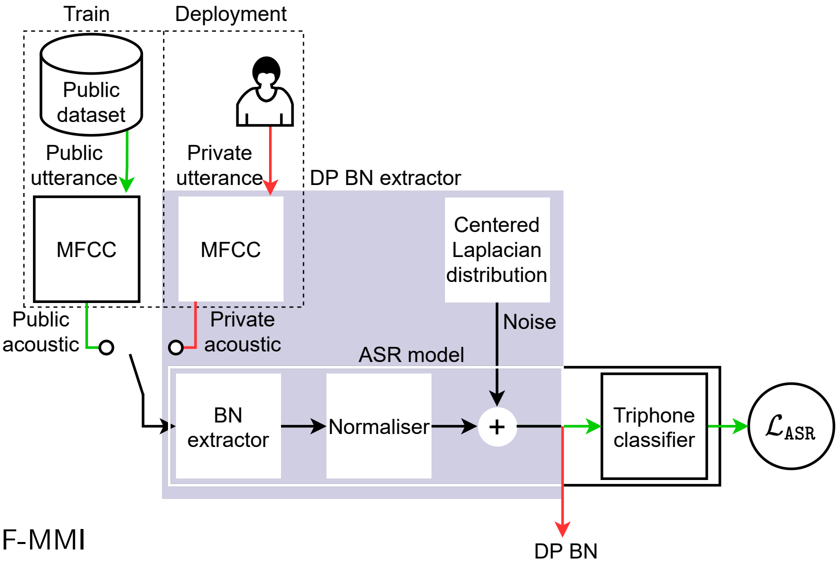
- add **Laplacian noise** to pitch and phonetic features
- noise scale $\propto \Delta/\epsilon$ with Δ maximum absolute difference between two data points
- if $\epsilon \ll 1$, **formal privacy guarantees** against any attack
- popular for tabular data (e.g., Apple uses $2 \leq \epsilon \leq 8$)

DP voice anonymization — DP pitch



Loss = pitch sequence correlation


Pitch (Hz)
140
120
100
80
40
60
80
100
120
140
160
180
200
Frame
Original pitch sequence
DP pich with $\varepsilon = 10$
DP pitch with $\varepsilon = 1$


Train
Deployment
Public dataset
Public utterance
Private utterance
DP BN extractor
MFCC
MFCC
Centered Laplacian distribution
Public acoustic
Private acoustic
Noise
ASR model
BN extractor
Normaliser
+
Triphone classifier
$\mathcal{L}_{ASR}$
DP BN

Loss = LF-MMI

Semi-Informed (utt-level) EER and WER (Librispeech)

Phonetic ϵ	**Pitch ϵ**	**EER**	**WER**
∞	∞	14.6%	5.4%
100	100	24.2%	6.0%
10	10	27.7%	7.0%
1	1	30.0%	7.8%

Laplacian noise **improves privacy**.

No formal guarantee though, because ϵ not small enough.

(Side note: utt-level Semi-Informed attacker stronger than spk-level one.)

Reminder: **voice anonymization is not (always) anonymization** because of

- possibly preserved (quasi-)identifiable **nonverbal attributes**
  - $\Rightarrow$ Many studies on re-identifying citizens from demographic attributes
  - $\Rightarrow$ No comparable study for speech attributes, let alone voice-anonymized speech
  - $\Rightarrow$ Some utterance-level attributes likely already concealed
  - $\Rightarrow$ Main foreseen risk due to utterance aggregation when metadata is included
- preserved **verbal content**
  - $\Rightarrow$ Solution depends on the intended usage

- When running automatic speech recognition (ASR) on the data, the verbal content cannot be changed.
- When using the data to train an acoustic model (AM), identify **named entities** carrying personal information and **discard** them from the speech signal.

Replacement strategy	Transformed text
No Replacement	Hi Mister Miller, the Lufthansa flight from Frankfurt Airport to Rome is leaving by six pm
Redact	Hi Mister IIIII, the IIIII flight from IIIII to IIIII is leaving by IIIII

- Private named entities are **domain-dependent**: person, age, ethnic category, email, licence plate number, occupation, organisation, address, date, calendar event, amount, URL, etc.
- There exists commercial software for legal, health, etc.

- When using the data to train a language model (LM), **replace** words instead

Replacement strategy	Transformed text
No Replacement	Hi Mister Miller, the Lufthansa flight from Frankfurt Airport to Rome is leaving by six pm
Typed-Placeholder	Hi Mister PER, the ORG flight from LOC to LOC is leaving by TIME
Named-Placeholder	Hi Mister Smith, the SAP flight from London to London is leaving by afternoon
Word by word	Hi Mister John, the BOSCH flight from New Boston to Berlin is leaving by eleven morning
Full entity	Hi Mister John, the BOSCH flight from New York to Berlin is leaving by twelve pm

- This also applies to NLP tasks such as named entity recognition (NER), intent detection (ID), or dialog act classification (DAC).

Verbal content anonymization for text processing

Replacement strategy	**VerbMobil NER F1-score**	**ATIS ID Accuracy**	**SNIPS ID Accuracy**	**en-TOD ID Accuracy**	**Restaurant DAC Accuracy**	**Taxi DAC Accuracy**
No replacement	$\mathbf{88.3 \pm 0.2}$	$\mathbf{98.4 \pm 0.2}$	$\mathbf{98.0 \pm 0.2}$	$\mathbf{99.4 \pm 0.0}$	$\mathbf{78.9 \pm 0.1}$	$\mathbf{90.0 \pm 0.1}$
Redact	0.2 ± 0.2	94.8 ± 0.2	89.7 ± 0.8	97.4 ± 0.6	75.9 ± 0.3	88.1 ± 0.2
Typed-Placeholder	0.0 ± 0.0	95.7 ± 0.3	54.1 ± 3.8	97.2 ± 0.7	76.5 ± 0.2	87.9 ± 0.5
Named Placeholder	13.5 ± 1.4	95.9 ± 0.3	76.2 ± 2.9	98.2 ± 0.1	77.3 ± 0.2	89.3 ± 0.1
Word-by-Word	72.6 ± 0.3	$\mathbf{98.6 \pm 0.2}^*$	$97.5 \pm 0.3^*$	$99.2 \pm 0.1^*$	78.4 ± 0.2	$89.9 \pm 0.2^*$
Full Entity	$85.9 \pm 0.3^*$	$\mathbf{98.5 \pm 0.2}^*$	$97.4 \pm 0.3^*$	$99.2 \pm 0.1^*$	$78.5 \pm 0.1^*$	$89.9 \pm 0.1^*$

- Full entity replacement preserves utility.
- However, it may still not result in anonymization due to preserved attributes.

Verbmobil dialog corpus, rephrasing by BART

Test set	Gender	Age
Original training data		
Original test set	70.3	65.4
Paraphrased test set	62.1	60.6
Anonymised training data		
Original test set	68.5	61.1
Paraphrased test set	66.7	60.5

- Hiding attributes such as age ($\leq$ or $>$ 21) or gender is a lot more difficult.

Was anonymization successful?

- Is an EER of xx% enough? What's the threshold?
- The **reduction in re-identification accuracy** after anonymization is more easily interpretable.
- Experiments suggest that for **short sentences**, **if** the dataset has **many speakers**, accurate **text anonymization**, **no metadata**, the answer is probably **yes**.
- This remains to be **legally validated** using, e.g., the three criteria of the Article 29 Working Party (European Data Protection Board)
  - linkability: ability to link records related to an individual or a group → we measured this for individuals, not groups
  - singling out: ability to single out an individual or a group→ TBD
  - inference: ability to re-identify an individual based on observed attributes → TBD

- **Anonymization**:
  - Improved attribute disentanglement and noising/quantization
  - Word replacement inside speech signals (not only text)
- **Selective attribute manipulation**:
  - Privacy w.r.t. other attributes, e.g., gender, age, accent
  - Utility for other tasks than ASR, e.g., medical
  - User-friendly interface
- **Evaluation**
  - Stronger, more realistic attackers (metadata, etc.)
  - Quantify re-identification risk based on nonverbal attributes
- **Watermarking** to avoid anonymized voice sounding like another real speaker
- Efficient **embedded, real-time implementation**
- Combination with **encryption** & **decentralized learning**