Speech anonymization

Emmanuel Vincent
Inria Nancy – Grand Est
Acknowledgments

Joint work with

- **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
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?

- Embedded implementation
- Cryptography
  - Homomorphic encryption
  - Secure multiparty computation
  - Searchable encryption
- AI
  - Physical obfuscation / deletion
  - Decentralized learning
  - Speech/text anonymization
Speech anonymization — Overall principle

- 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 (≠ strict legal definition)

- Achieving this goal requires:
  - voice transformation or conversion (a.k.a. **voice anonymization**) preserving non-identifiable nonverbal attributes (ASR+TTS not enough),
  - verbal content anonymization,
  - possibly, hiding some identifiable nonverbal attributes.

- 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

From now on, focus on voice anonymization by voice transformation or conversion.
From now on, **focus on voice anonymization** by voice transformation or conversion.
- The success or failure of voice anonymization can be evaluated via **speaker verification**.

- In practice, speaker embeddings are x-vectors.

- Higher score $\Rightarrow$ greater chance of being from the same speaker.
Attacker’s knowledge

Original

Ignorant

Lazy-Informed

Semi-Informed (spk/utt level)

Informed

Privacy evaluation (EER, Linkability, ...)

Trial set

Enrollment set
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-2 of the VoicePrivacy 2020 and 2022 Challenges
  - VoiceMask
  - VTLN

![Diagram of speech anonymization process](image-url)
### Voice transformation — Results

**EER (Librispeech)**

<table>
<thead>
<tr>
<th>Attacker</th>
<th>Baseline-2</th>
<th>VoiceMask</th>
<th>VTLN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original speech</td>
<td></td>
<td>4.3%</td>
<td></td>
</tr>
<tr>
<td>Ignorant</td>
<td>26.2%</td>
<td>28.7%</td>
<td>27.4%</td>
</tr>
<tr>
<td>Semi-Informed (utt-level)</td>
<td>5.3%</td>
<td>5.0%</td>
<td>6.3%</td>
</tr>
</tbody>
</table>

Simple transformations *fail against non-ignorant attackers.*
Voice conversion — Approach

- Idea: replace user’s voice by that of a target speaker
- Baseline-1 of the VoicePrivacy 2020 Challenge

Phonetic features = bottleneck (BN)
Voice conversion — Approach

- Idea: replace user's voice by that of a target speaker
- Baseline-1 of the VoicePrivacy 2020 Challenge

Phonetic features = bottleneck (BN)
Voice conversion — Approach

- Idea: **replace user’s voice** by that of a target speaker
- Baseline-1 of the VoicePrivacy 2020 Challenge

Phonetic features = bottleneck (BN)

Input speech

Pitch extractor

Pitch

Speech recognizer

Phonetic features

Speaker embedding

x-vector

Anonymization

Anonymized x-vector

Pool of x-vectors

14 - Speech anonymization - E. Vincent
Voice conversion — Approach

- Idea: replace user’s voice by that of a target speaker
- Baseline-1 of the VoicePrivacy 2020 Challenge

Phonetic features = bottleneck (BN)
Voice conversion — Approach

- Idea: **replace user’s voice** by that of a target speaker
- Baseline-1 of the VoicePrivacy 2020 Challenge

Phonetic features = bottleneck (BN)
Voice conversion — Design choices

- Target selection procedure:

- Retained choice: random gender + dense

Original [Audio]  Modified [Audio]
Voice conversion — Privacy results

Top-20 PLDA-based identification accuracy (CommonVoice)

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.

Gender classification accuracy:
- (a) Original: 98.6%
- (b) Random (same): 100.0%
- (d) Dense (random): 53.3%
Voice conversion — Utility results

Speech recognition (LibriSpeech)

<table>
<thead>
<tr>
<th>Decoding scenario</th>
<th>O-O</th>
<th>A-O</th>
<th>O-A</th>
<th>A-A</th>
</tr>
</thead>
<tbody>
<tr>
<td>WER (%)</td>
<td>4.14</td>
<td>6.65</td>
<td>10.45</td>
<td>4.65</td>
</tr>
</tbody>
</table>

Emotion recognition (IEMOCAP)

<table>
<thead>
<tr>
<th>Decoding scenario</th>
<th>O-O</th>
<th>A-O</th>
<th>A-A</th>
</tr>
</thead>
<tbody>
<tr>
<td>UAR (%)</td>
<td>57.1</td>
<td>30.6</td>
<td>50.8</td>
</tr>
</tbody>
</table>

Small or negligible loss of utility after retraining on anonymized data (A-A).
Voice conversion — Subjective results

**Naturalness**

- Original
- B1

- Intelligibility

- Original
- B1

- Similarity

- Same-speaker
- Different-speaker

**Mean**

- 0.74 0.29 0.27 0.25 0.25 0.29 0.31

**Median**

- 0.23 0.38 0.09

- 0.69 0.34 0.30 0.28 0.30 0.34 0.36

- 0.42 0.45 0.46 0.38 0.44 0.48 0.50

- 0.33 0.39 0.19

- 0.81 0.42 0.45 0.46

- 0.37 0.38 0.38 0.32 0.39 0.36 0.37 0.35 0.32

- 0.39 0.37 0.38 0.32

19 - Speech anonymization - E. Vincent
Voice conversion — Limitation

• Key limitations:
  > insufficient protection when the attacker can narrow down the search to few speakers based on side information
  > pitch and phonetic features contain residual speaker information, which remains after resynthesis
  > it can be captured by a more powerful attacker

• Solutions explored:
  > adversarial representation learning
  > noise-based local differential privacy
  > slicing into shorter signals
Adversarial learning — Approach

Adversarial learning of phonetic features for speech recognition

\[ \frac{\partial L_p}{\partial \theta_{enc}} \]

\[ \beta \frac{\partial L_{ctc}}{\partial \theta_{ctc}} + (1 - \beta) \frac{\partial L_{att}}{\partial \theta_{att}} \]

\[ L_{asr} \]

\[ O_i \rightarrow \text{Encoder (} \theta_{enc} \text{)} \rightarrow B_\lambda \rightarrow CTC (\theta_{ctc}) \rightarrow Y_i \]

\[ \frac{\lambda}{\partial \theta_{enc}} \]

\[ GRL \rightarrow \text{Speaker adv. branch (} \theta_{spk} \text{)} \rightarrow z_i \]

\[ \frac{\partial L_{spk}}{\partial \theta_{spk}} \]

\[ L_{spk} \]
Adversarial learning — Results

Accuracy, EER and WER (Librispeech)

<table>
<thead>
<tr>
<th></th>
<th>Spec. feat.</th>
<th>$\lambda = 0$</th>
<th>$\lambda = 0.5$</th>
<th>$\lambda = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Speaker identification accuracy</strong></td>
<td>93.1%</td>
<td>46.3%</td>
<td>6.4%</td>
<td>2.5%</td>
</tr>
<tr>
<td><strong>Speaker verification EER</strong></td>
<td>5.7%</td>
<td>23.1%</td>
<td>22.0%</td>
<td>19.6%</td>
</tr>
<tr>
<td><strong>Speech recognition WER</strong></td>
<td>–</td>
<td>10.9%</td>
<td>12.5%</td>
<td>12.5%</td>
</tr>
</tbody>
</table>

Adversarial learning generalizes poorly to unseen speakers.
Local differential privacy (DP) principle:

- add **Laplacian noise** to pitch and phonetic features
- noise scale \( \propto \Delta / \epsilon \) with \( \Delta \) 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 anonymization — DP pitch

Loss = pitch sequence correlation
Original pitch sequence       DP pitch with $\varepsilon = 10$       DP pitch with $\varepsilon = 1$
DP anonymization — DP phonetic features

Loss = LF-MMI
Semi-Informed (utt-level) EER and WER (Librispeech)

<table>
<thead>
<tr>
<th>Phonetic $\epsilon$</th>
<th>Pitch $\epsilon$</th>
<th>EER</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\infty$</td>
<td>$\infty$</td>
<td>14.6%</td>
<td>5.4%</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>24.2%</td>
<td>6.0%</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>27.7%</td>
<td>7.0%</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>30.0%</td>
<td>7.8%</td>
</tr>
</tbody>
</table>

Laplacian noise improves privacy.

No formal guarantee though, because $\epsilon$ not small enough.

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

Semi-Informed (utt-level) linkability (Librispeech)

Slicing into 1 or 1.5 s segments improves privacy with no loss of utility.
Successive segment re-identification (Librispeech, 1.5 s segments)

<table>
<thead>
<tr>
<th></th>
<th>from text</th>
<th>from speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of segments</td>
<td>11,330</td>
<td>364</td>
</tr>
<tr>
<td>Average normalized rank (%)</td>
<td>28.3</td>
<td>43.5</td>
</tr>
<tr>
<td>Median normalized rank (%)</td>
<td>17.9</td>
<td>19.8</td>
</tr>
<tr>
<td>Precision at top-1 (%)</td>
<td>1.4</td>
<td>2.5</td>
</tr>
<tr>
<td>Precision at top-10% (%)</td>
<td>37.8</td>
<td>38.3</td>
</tr>
</tbody>
</table>

A Semi-Informed attacker **cannot reliably re-identify successive segments.** Reassembling entire utterances would be even harder.
Verbal content anonymization for AM training

- 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.

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.
Verbal content anonymization for text processing

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

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

- 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

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

- Full entity replacement preserves utility.

- However, it does not fully prevent speaker re-identification. Hiding age, gender, etc., 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 so far suggest that, if there are many speakers in the dataset, accurate text anonymization, no metadata, the answer is probably yes.

- This remains to be legally validated using, e.g., the three legally admitted criteria of the Article 29 Working Party (European Data Protection Board)
  - linkability: ability to link records related to an individual $\rightarrow$ we measured this
  - singling out: ability to single out an individual $\rightarrow$ TBD
  - inference: ability to re-identify an individual based on observed traits $\rightarrow$ TBD
Perspectives

- **Anonymization**:
  - Improved disentanglement
  - Improved feature decorrelation / non-i.i.d. noise for DP
  - 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 attackers, perhaps more realistic too (metadata, etc.)

- **Watermarking** to avoid anonymized voice sounding like another real speaker’s voice

- Efficient **embedded implementation**

- Combination with **encryption & decentralized learning**
• **Task**: develop a voice anonymization system.

• **Resources**:
  > Train, devel, test datasets
  > 3 baseline systems
  > Evaluation scripts

• **Updates w.r.t. VoicePrivacy 2020 Challenge**:
  > Stronger, Semi-Informed (utt-level) attack model
  > New ranking based on WER for multiple EER levels
  > Complementary pitch correlation and voice distinctiveness utility metrics

**Submission deadline**: July 31, 2022

**Workshop at Interspeech**: September 23–24, 2022