Soutenance de thèse

Speaker Anonymization: Representation, Evaluation and Formal Guarantees

December 2nd, 2021
Brij Mohan Lal Srivastava

Supervisors:
Dr. Aurélien Bellet (Magnet)
Dr. Emmanuel Vincent (Multispeech)
Prof. Marc Tommasi (Université de Lille)
Context

Widespread usage of voice interfaces. Relies on:
- Massive centralized storage of data
- Advances in speech processing
- Enormous computing capabilities

Raises privacy threats beyond the spoken message alone.
Sensitivity of speech data

A voice technology company or a third-party attacker may be interested in finding out

- the speaker’s identity
- speaker attributes (age, gender, accent, etc.)
- the emotions expressed in the utterance
- personality traits
- health status
- etc.
Relevant legal constraints

Voice data can produce distinguishing and repeatable biometric features.

1. Right to privacy — a fundamental right
2. General Data Protection Regulation (GDPR, 2016) – requires compliance by May 2018
3. Exploring the ethical, technical and legal issues of voice assistants (2020) – white paper by CNIL
4. EDPB Guidelines 02/2021 on virtual voice assistants
We aim to answer the following central question in this thesis:

*How to remove the biometric identity of the speaker from any speech utterance, while maintaining its usefulness for Automatic Speech Recognition (ASR)?
Summary of contributions

1. Definition of a threat model for speaker anonymization, along with strong attacks that leverage auxiliary knowledge

2. Privacy-preserving adversarial learning method for end-to-end ASR

3. Optimization of the privacy-utility trade-off in x-vector-based anonymization

4. Demonstration of the viability of anonymized speech to train an ASR system

5. Differentially-private speaker anonymization
Outline

1. Background on speech processing tasks
2. Threat Model and Privacy Evaluation using Informed Attackers
3. X-vector based Anonymization
4. Removing Residual Speaker Information — Towards Provable Guarantees
5. Conclusion and Perspectives
Automatic Speech Recognition (ASR)

Evaluation metric: Word Error Rate (WER)
- Edit distance between the reference and the estimated transcription
Automatic Speaker Identification (ASI)

Test utterance

Classification

Most probable identity among a closed set of speakers

- **Evaluation metric**: Accuracy
- **Setting**: Closed set of speakers
Speaker Anonymization | Background

Automatic Speaker Verification (ASV)

- Evaluation metric: Equal Error Rate (EER)
- Setting: Open set of speakers
Outline

1. Background on speech processing tasks

2. Threat Model and Privacy Evaluation using Informed Attackers

3. X-vector based Anonymization

4. Removing Residual Speaker Information — Towards Provable Guarantees

5. Conclusion and Perspectives
Subsequently adopted for the first VoicePrivacy challenge
Attacker’s knowledge

- **Ignorant**: Unaware of anonymization
- **Lazy-Informed**: Aware of anonymization but partial exploitation
- **Semi-Informed**: Full exploitation but with different parameters
- **Informed**: Complete knowledge and exploitation
Using voice conversion (VC) for anonymization

**Goal:**
To convert a given source speaker’s voice into a target speaker’s voice without changing the content.
Speaker Anonymization | Threat Model and Privacy Evaluation

Voice conversion methods

Considered three representative transformation methods (sample original “stuff it into you, his belly counseled him”)

- **Voicemask:**
  - Time-invariant spectral envelope warping + linear pitch transformation

- **Vocal tract length normalization (VTLN):**
  - Phonetic class-wise spectral envelope warping + linear pitch transformation

- **Disentangled speech representation (DSR):**
  - End-to-end encoder-decoder based speaker information removal
Target selection strategies

(a) const

(b) perm

(c) random
Experimental setup

- **Data set:** LibriSpeech, a 960-hour English read speech corpus derived from audiobooks containing 1,283 male and 1,201 female speakers
Experimental setup

- **Data set**: LibriSpeech, a 960-hour English read speech corpus derived from audiobooks containing 1,283 male and 1,201 female speakers

- **Privacy metrics**: Linkability ($D_{\leftrightarrow}^{sys}$)
  - $D_{\leftrightarrow}^{sys} \in [0, 1]$
  - $0 \Rightarrow$ full protection, $1 \Rightarrow$ no protection

Gomez-Barrero et al., “General framework to evaluate unlinkability in biometric template protection systems”. 
Experimental setup

- **Data set:** LibriSpeech, a 960-hour English read speech corpus derived from audiobooks containing 1,283 male and 1,201 female speakers.

- **Privacy metrics:** Linkability ($D_{\leftrightarrow}^{sys}$)
  - $D_{\leftrightarrow}^{sys} \in [0, 1]$
  - $0 \Rightarrow$ full protection, $1 \Rightarrow$ no protection

- **Utility metric:** Word Error Rate (WER)

Gomez-Barrero et al., “General framework to evaluate unlinkability in biometric template protection systems”.
Privacy evaluation (core contribution)

- **Trial set**
  - Original
  - Ignorant
  - Lazy-Informed
  - Semi-Informed
  - Informed

- **Enrollment set**
  - ASV\textsubscript{eval}
  - ASV\textsuperscript{anon}eval

- **Linkability** $D_{\leftrightarrow}^{sys}$
Comparison of different attackers (privacy)

- **VoiceMask**
  - Ign: 0.35
  - Inf: 0.82

- **VTLN**
  - Ign: 0.27
  - Semi: 0.34
  - Inf: 0.89

- **DSR**
  - Ign: 0.20
  - Semi: 0.69
  - Inf: 0.69

- Linkability increases as the attacker’s knowledge increases

---

Speaker Anonymization | Threat Model and Privacy Evaluation
Comparison of different attackers (utility)

- WER (%) of the anonymized speech as compared to the baseline

<table>
<thead>
<tr>
<th>Original data – Original model</th>
<th>Anonymized data – Retrained model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VoiceMask</td>
</tr>
<tr>
<td></td>
<td>random</td>
</tr>
<tr>
<td>9.4</td>
<td>18.1</td>
</tr>
</tbody>
</table>

- VoiceMask and VTLN show similar degradation in terms of WER, while DSR degrades the quality significantly
Summary of this part

- Identified actors and proposed a threat model for speech anonymization
- Defined several attackers with increasing knowledge
- Evaluated three voice conversion strategies against these attackers
- Established that auxiliary knowledge strengthens the attack
Summary of this part

- Identified actors and proposed a threat model for speech anonymization
- Defined several attackers with increasing knowledge
- Evaluated three voice conversion strategies against these attackers
- Established that auxiliary knowledge strengthens the attack
- **Limitations**: Fixed set of “real” target speakers and significant degradation of quality
Outline

1. Background on speech processing tasks
2. Threat Model and Privacy Evaluation using Informed Attackers
3. X-vector based Anonymization
4. Removing Residual Speaker Information — Towards Provable Guarantees
5. Conclusion and Perspectives
X-vector based anonymization

1. F0 extractor
2. Anonymization
3. Pitch conversion
4. Speech synthesis

Input speech → Anonymized speech

Pool of x-vectors


- Mixed-target pseudo-speaker and flexible scaling of target pool
### Design choices in x-vector space

**Question by speakers and users:**
- How to choose the target pseudo-speaker for an optimal privacy-utility trade-off?

<table>
<thead>
<tr>
<th>Assignment</th>
<th>Speaker / Utterance</th>
<th>Source</th>
<th>x-vector</th>
<th>Same / Other / Random</th>
<th>Dense / Sparse</th>
<th>Gender Selection</th>
<th>Near / Far</th>
<th>Proximity</th>
<th>Candidate x-vectors</th>
<th>Target x-vector</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

24/45
Speaker Anonymization | X-vector based Anonymization

Design choices in x-vector space

Question by speakers and users:
- How to choose the target pseudo-speaker for an optimal privacy-utility trade-off?
Comparison under different attack scenarios

Recommended anonymization scheme:
Distance PLDA, Proximity dense, Gender random, Assignment speaker-level
Large-scale speaker study

- Realistically, without auxiliary information, the attacker may need to search the true identity among several speakers
- **Goal**: Attacker’s performance as a function of the number of enrollment speakers
Realistically, without auxiliary information, the attacker may need to search the true identity among several speakers.

**Goal:** Attacker’s performance as a function of the number of enrollment speakers.

**Data set:** Mozilla Common Voice (English), a speech data set collected by crowdsourcing.
  - Used 24,610 speakers out of 52,000, with total 320,000 utterances.
  - 20 speakers under re-identification attack.
Realistically, without auxiliary information, the attacker may need to search the true identity among several speakers.

**Goal:** Attacker’s performance as a function of the number of enrollment speakers.

**Data set:** Mozilla Common Voice (English), a speech data set collected by crowdsourcing.

- Used 24,610 speakers out of 52,000, with total 320,000 utterances.
- 20 speakers under re-identification attack.

**Privacy metrics:** top-\(k\) speaker identification precision.
Better protection after anonymization (Top-$k$)

- Top-20 precision for different attackers as a function of the number of speakers in the population.
- After anonymization, a crowd of 52 speakers provides as good protection as 20,500 speakers before anonymization.
Utility evaluation

Original data – Original model

Anonymized data – Original model

Anonymized data – Retrained model

Original data – Retrained model
Utility of anonymized speech

- Re-training ASR system with anonymized speech
- Close to baseline performance over anonymized data
Summary of this part

- Actively participated in the design and organization of the VoicePrivacy Challenge
- Compared and recommended the best combination of the four design choices for x-vector based anonymization scheme
- Established the utility of anonymized speech for both ASR training and decoding
- Large-scale speaker study showed that the speakers are much better protected after anonymization
Summary of this part

▶ Actively participated in the design and organization of the VoicePrivacy Challenge
▶ Compared and recommended the best combination of the four design choices for x-vector based anonymization scheme
▶ Established the utility of anonymized speech for both ASR training and decoding
▶ Large-scale speaker study showed that the speakers are much better protected after anonymization
▶ Limitation 1: disentanglement of speaker information not perfect
▶ Limitation 2: only empirical evaluation of privacy using ASI and ASV
Outline

1. Background on speech processing tasks
2. Threat Model and Privacy Evaluation using Informed Attackers
3. X-vector based Anonymization
4. Removing Residual Speaker Information — Towards Provable Guarantees
5. Conclusion and Perspectives
Differential privacy (1/2)

Definition (Local differential privacy)
Let \( A \) be a randomized algorithm taking as input a data point in some space \( \mathcal{X} \), and let \( \epsilon > 0 \). We say that \( A \) is \( \epsilon \)-local differentially private (\( \epsilon \)-LDP) if for any \( x, x' \in \mathcal{X} \) and any \( S \subseteq \text{range}(A) \):

\[
\Pr[A(x) \in S] \leq e^\epsilon \Pr[A(x') \in S],
\]

where the probabilities are taken over the randomness of \( A \).
Differential privacy (2/2)

Definition (Laplace mechanism)
Let $f : \mathcal{X} \to \mathbb{R}^d$ and let the $\ell_1$-sensitivity of $f$ be defined as

$$\Delta_1(f) = \max_{x, x' \in \mathcal{X}} |f(x) - f(x')|_1.$$ 

Let $\eta = [\eta_1, \ldots, \eta_d] \in \mathbb{R}^d$ be a vector where each $\eta_i \sim \text{Lap}(\Delta_1(f)/\epsilon)$ is drawn from the centered Laplace distribution with scale $\Delta_1(f)/\epsilon$. The algorithm $A(\cdot) = f(\cdot) + \eta$ is $\epsilon$-local DP.
Replaced the F0 extractor and ASR AM with their DP versions — trained with the noise layer 🎧
Differentially-private pitch extractor

\[ h \in [0, 1]^{C \times T} \quad \Delta_1(\mathcal{E}) = C \times T \times 1 \quad h^{DP} = N_p(h) = h + \text{Lap}(\Delta_1(\mathcal{E})/\epsilon) \]

\[ L_{\text{pitch}} = 1 - \sum_{i=1}^{N} \text{corr}(p_i, p_i^{DP}) \]
Effect of DP on pitch sequence

- Original (non-private) and noisy pitch for $\epsilon = 10$ and $\epsilon = 1$
- DP-Pitch preserves the intonation reasonably well
Privacy and utility of DP-Pitch features

- DP-Pitch significantly reduces the speaker identification accuracy
- Pearson correlation is preserved for $\epsilon > 1$
Speaker Anonymization | Removing Residual Speaker Information

**Differentially-private BN extractor**

\[
B^{DP} = \mathcal{N}_B(B) = \left[ \begin{array}{c} \mathcal{N}_b(b_1) \\ \vdots \\ \mathcal{N}_b(b_T) \end{array} \right] \quad \quad \mathcal{N}_B(b) = \frac{b}{\|b\|_1} + \text{Lap}(2/\epsilon)
\]
Privacy and utility of DP-BN features

- DP-BN significantly reduces speaker identification accuracy
- Gradual decline of utility as $\epsilon$ increases
## Combination of DP-BN and DP-Pitch features

<table>
<thead>
<tr>
<th>Method</th>
<th>Privacy</th>
<th>Utility</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Local $\epsilon$</td>
<td>Practical $D_{\leftrightarrow}$</td>
<td>Practical WER</td>
</tr>
<tr>
<td>Without DP (part 2)</td>
<td>$\infty$</td>
<td>$\infty$</td>
<td>0.14</td>
</tr>
<tr>
<td>With DP</td>
<td>100</td>
<td>1.0</td>
<td>0.11</td>
</tr>
<tr>
<td>With DP</td>
<td>100</td>
<td>0.1</td>
<td>0.10</td>
</tr>
<tr>
<td>With DP</td>
<td>10</td>
<td>1.0</td>
<td>0.13</td>
</tr>
<tr>
<td>With DP</td>
<td>10</td>
<td>0.1</td>
<td>0.13</td>
</tr>
<tr>
<td>With DP</td>
<td>1</td>
<td>1.0</td>
<td>0.12</td>
</tr>
<tr>
<td>With DP</td>
<td>1</td>
<td>0.1</td>
<td>0.10</td>
</tr>
</tbody>
</table>

- Rise in privacy protection after pluggin-in DP feature extractors
- Marginal rise in utility with DP-BN $\epsilon = 100$ and $\epsilon = 10$
Summary of this part

- Challenged the disentanglement assumption made in the previous part
- Formulated methods for obtaining differentially-private BN and Pitch features
- The utterance-level privacy budget for DP-Pitch is $\epsilon$, while for DP-BN it is $\epsilon \times T$
- Although the overall privacy budget is too large, DP noise addition translates into clear gain in privacy, and sometimes in utility
Outline

1. Background on speech processing tasks
2. Threat Model and Privacy Evaluation using Informed Attackers
3. X-vector based Anonymization
4. Removing Residual Speaker Information — Towards Provable Guarantees
5. Conclusion and Perspectives
Global summary

- Identified the actors and defined a threat model for speech anonymization, which was adopted by the VoicePrivacy challenge
- Proposed strict evaluation protocol using a continuum of attackers
- Proposed design choices and pitch conversion methods for x-vector based anonymization
- Proposed differentially-private scheme
- Conducted large-scale speaker study to realistically measure the strength of anonymization
- Established the utility of anonymized speech for ASR training and decoding
- The proposed solution provides a high degree of protection against the strongest attack
Extensions and open problems

- Use of adversarially-learned bottleneck features in x-vector based anonymization
- More design choices, such as the selection of different speaker pools
- Stronger attackers built using utterance-level assignment
- Assessment of usability in a wider context, such as remote health monitoring, emotion preservation, etc.
- Extension to other languages
Thank you for your attention!