Soutenance de thèse

Speaker Anonymization: Representation, Evaluation and Formal Guarantees

December 2nd, 2021

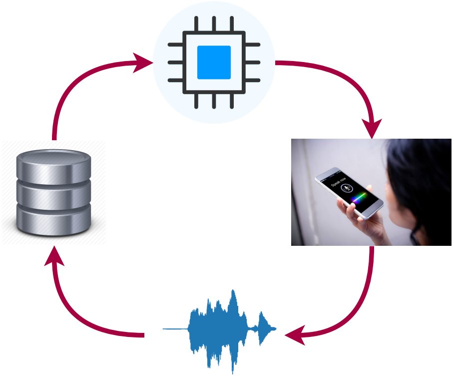
Brij Mohan Lal Srivastava

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

Problem

We aim to answer the following central question in this thesis:

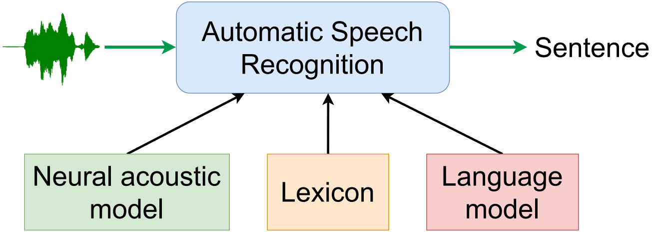
> *How to remove the biometric identity of the speaker from any speech utterance, while maintaining its usefulness for Automatic Speech Recognition (ASR)?*

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

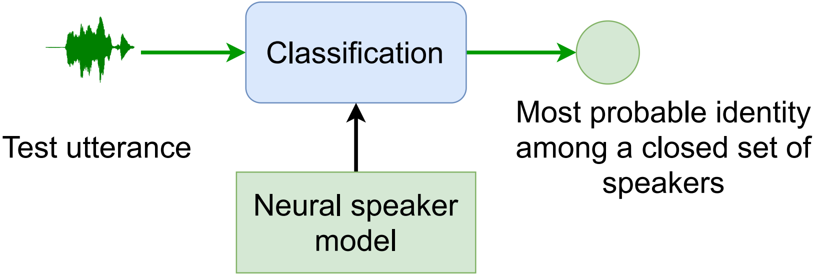
Outline

Automatic Speech Recognition (ASR)



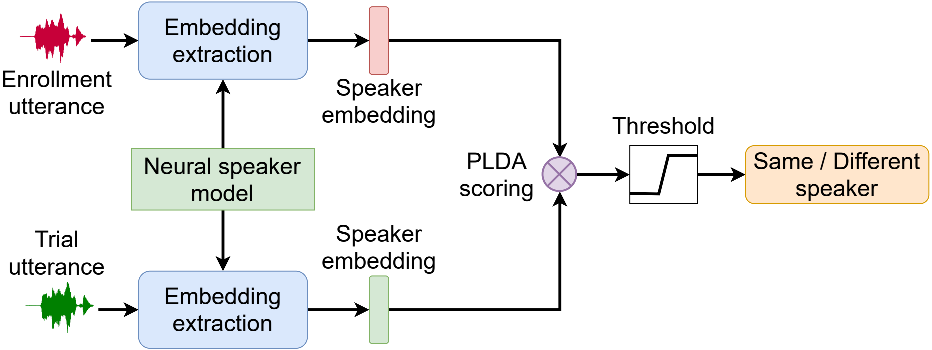
- Evaluation metric: Word Error Rate (WER)
  - Edit distance between the reference and the estimated transcription

Automatic Speaker Identification (ASI)



- Evaluation metric: Accuracy
- Setting: Closed set of speakers

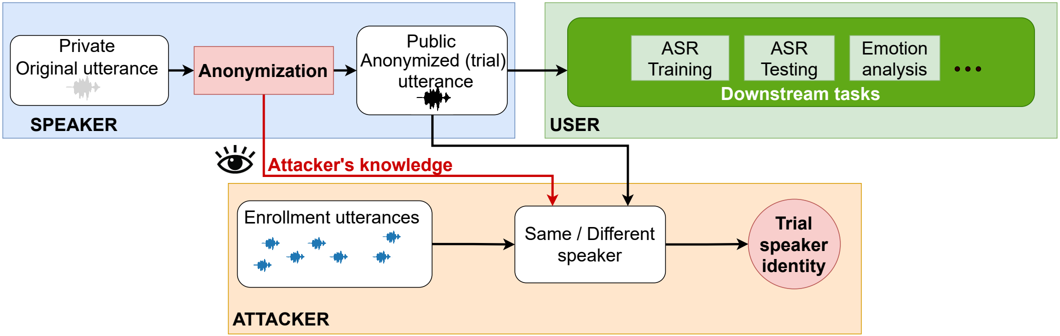
Automatic Speaker Verification (ASV)



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

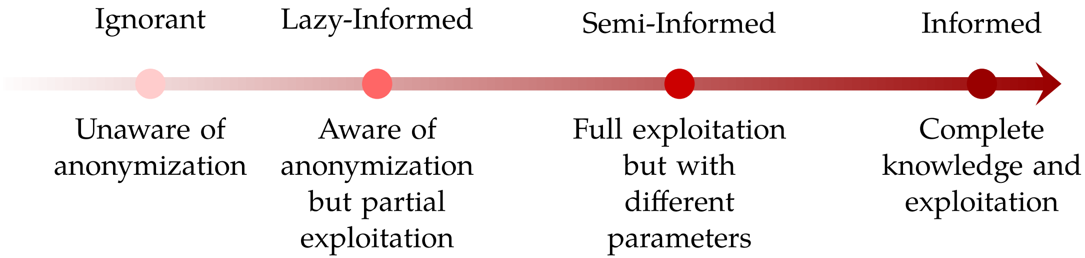
Outline

Proposed threat model



- Subsequently adopted for the first VoicePrivacy challenge

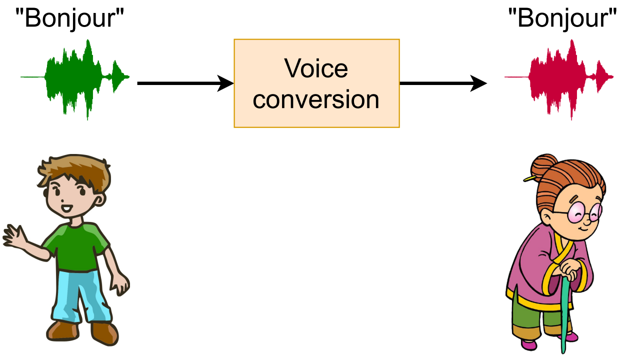
Attacker’s knowledge



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.

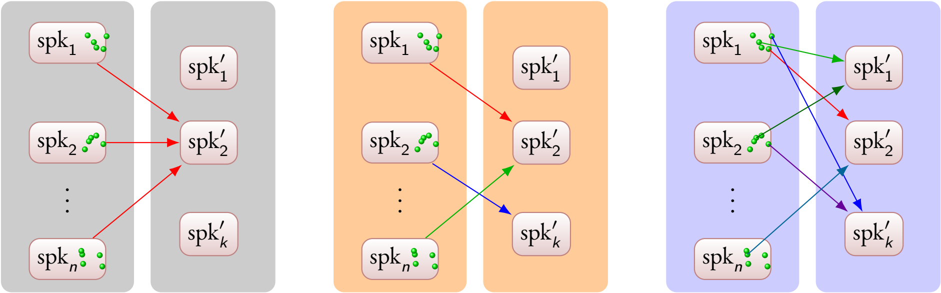


Voice conversion methods

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

Marta Gomez-Barrero et al. "General framework to evaluate unlinkability in biometric template protection systems". In: *IEEE Transactions on Information Forensics and Security* 13.6 (2017), pp. 1406–1420.

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

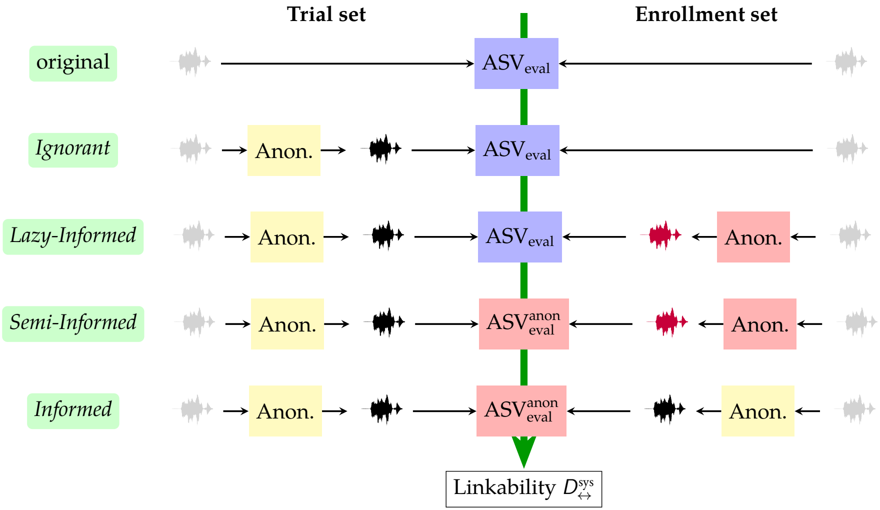
Gomez-Barrero et al., "General framework to evaluate unlinkability in biometric template protection systems".

Experimental setup

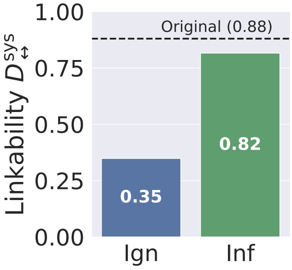
- Data set: LibriSpeech, a 960-hour English read speech corpus derived from audiobooks containing 1,283 male and 1,201 female speakers
- Privacy metrics: Linkability ($D_{\leftrightarrow}^{\text{sys}}$)
  - $D_{\leftrightarrow}^{\text{sys}} \in [0, 1]$
  - **0** $\Rightarrow$ **full** protection, **1** $\Rightarrow$ **no** protection
- Utility metric: Word Error Rate (WER)

Gomez-Barrero et al., "General framework to evaluate unlinkability in biometric template protection systems".

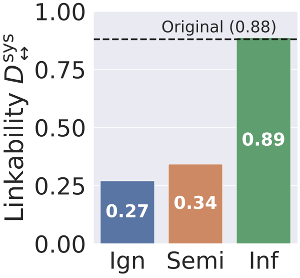
Privacy evaluation (core contribution)



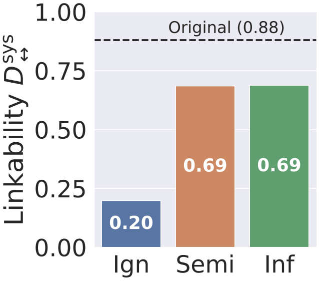
Comparison of different attackers (privacy)



(a) VoiceMask



(b) VTLN



(c) DSR

- Linkability increases as the attacker's knowledge increases

Comparison of different attackers (utility)

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

Original data – Original model	**Anonymized data – Retrained model**						
	VoiceMask	**VTLN**			**DSR**		
	random	*const*	*perm*	*random*	*const*	*perm*	*random*
9.4	18.1	19.8	18.4	15.9	41.5	23.7	115.1

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

Summary of this part

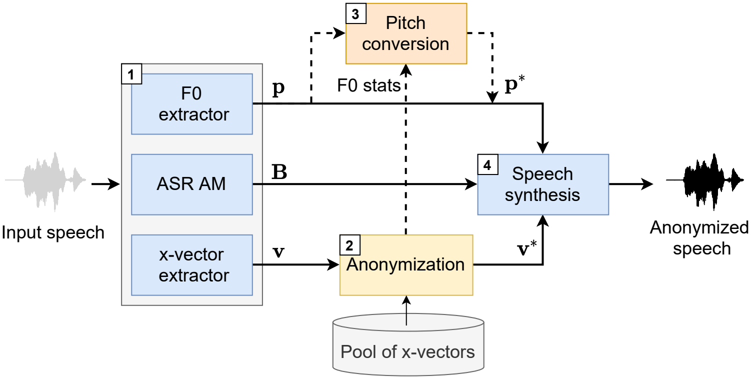
- Identified actors and proposed a threat model for speech anonymization
- Defined several attackers with increasing knowledge
- Evaluated three voice conversion strategies against these attackers
- Established that auxiliary knowledge strengthens the attack

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

X-vector based anonymization



- Mixed-target *pseudo-speaker* and flexible scaling of target pool

Fuming Fang et al. “Speaker Anonymization Using x-vector and Neural Waveform Models”. In: *10th ISCA Speech Synthesis Workshop*. 2019.

Design choices in x-vector space

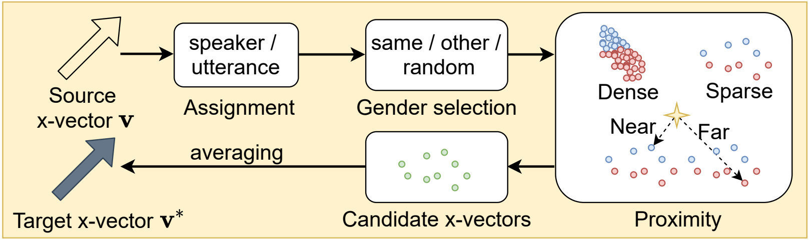
Question by speakers and users:

- How to choose the target pseudo-speaker for an optimal privacy-utility trade-off?

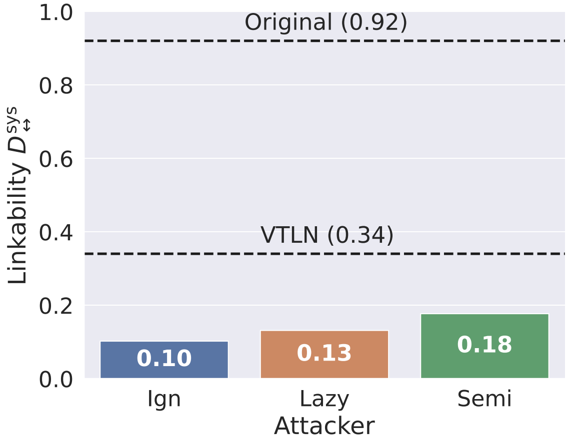
Design choices in x-vector space

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

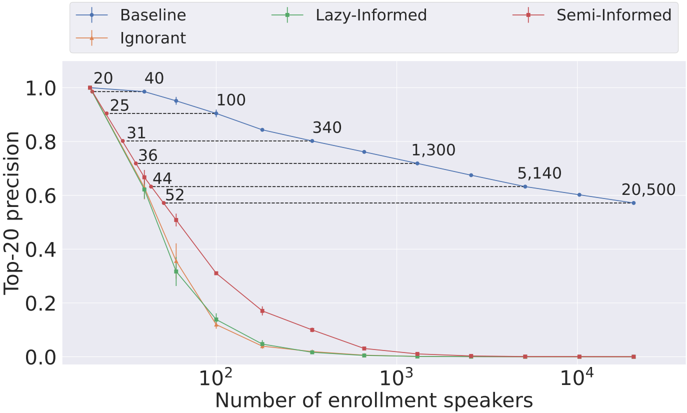
Large-scale speaker study

- 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
- 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

Large-scale speaker study

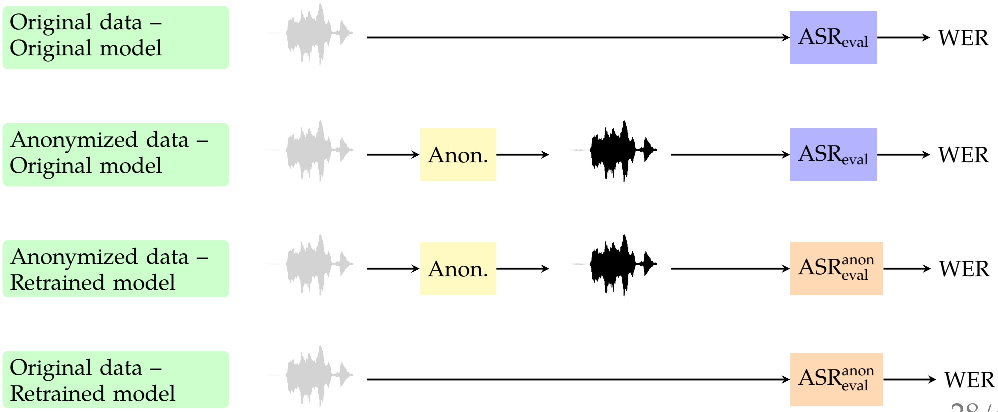
- Realistically, without auxiliary information, the attacker may need to search the true identity among several speakers
- Goal: Attacker's performance as a function of the number of enrollment speakers
- Data set: Mozilla Common Voice (English), a speech data set collected by crowdsourcing
  - Used 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)

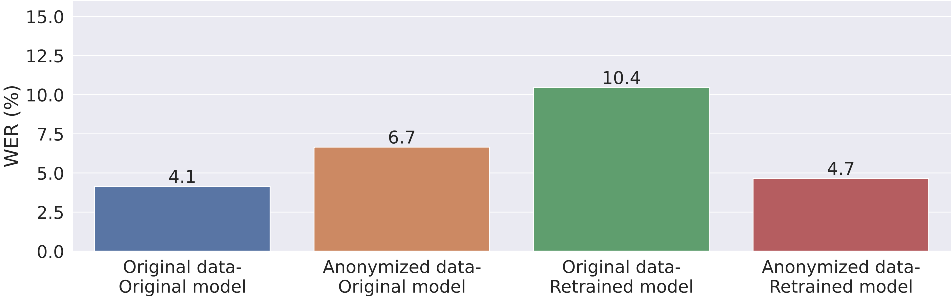


- Top-20 precision for different attackers as a function of the number of speakers in the population
- After anonymization, a crowd of **52** speakers provides as good protection as **20,500** speakers before anonymization

Utility evaluation



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

Differential privacy (1/2)

Definition (Local differential privacy)

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

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

where the probabilities are taken over the randomness of $\mathcal{A}$.

John C Duchi, Michael I Jordan, and Martin J Wainwright. "Local privacy and statistical minimax rates". In: *54th IEEE Annual Symposium on Foundations of Computer Science*. 2013.

Differential privacy (2/2)

Definition (Laplace mechanism)

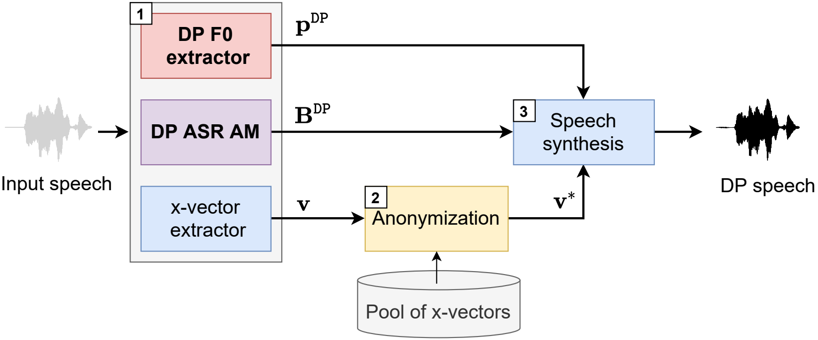
Let $f : \mathcal{X} \rightarrow \mathbb{R}^d$ and let the ℓ_1-sensitivity of f be defined as

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

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 $\mathcal{A}(\cdot) = f(\cdot) + \eta$ is ϵ-local DP.

Cynthia Dwork et al. "Calibrating noise to sensitivity in private data analysis". In: *3rd Theory of Cryptography Conference*. 2006.

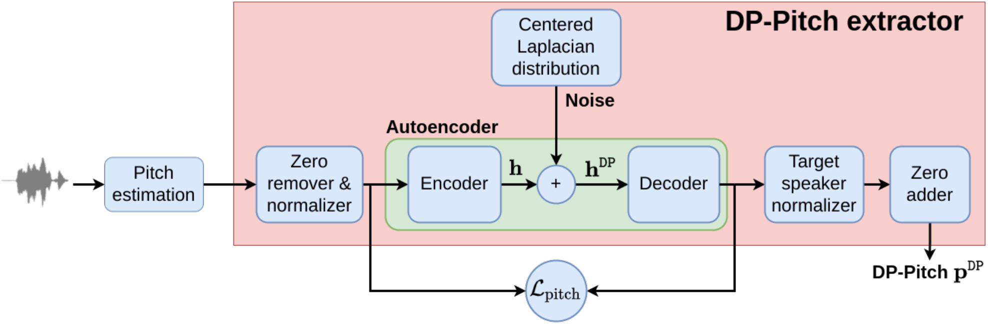
Overview of approach



- Replaced the F0 extractor and ASR AM with their DP versions — trained with the noise layer

Differentially-private pitch extractor

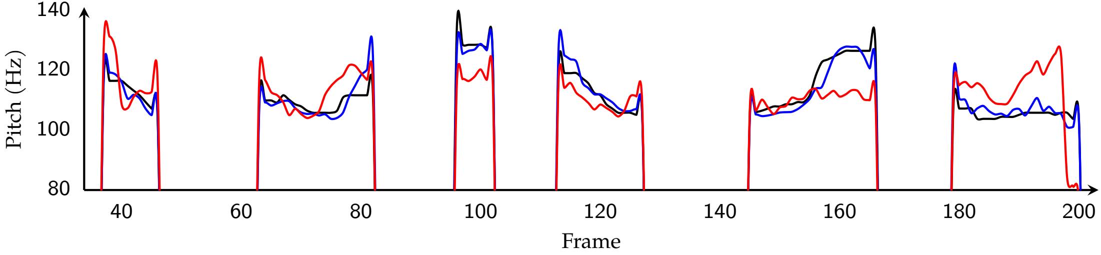
$\mathbf{h} \in [0,1]^{C \times T}$	$\Delta_1(\mathcal{E}) = C \times T \times 1$	$\mathbf{h}^{\text{DP}} = \mathcal{N}_p(\mathbf{h}) = \mathbf{h} + \text{Lap}(\Delta_1(\mathcal{E})/\epsilon)$



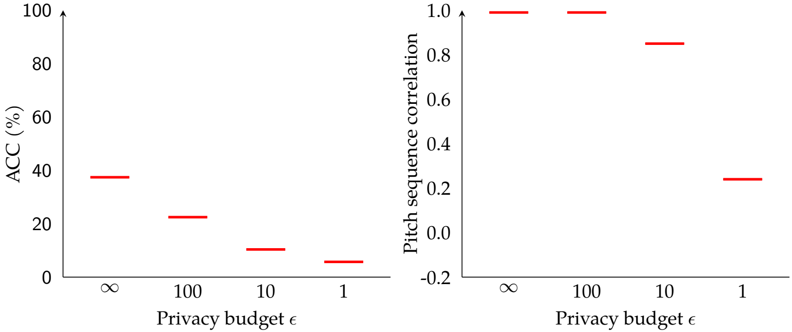
$$\mathcal{L}_{\text{pitch}} = 1 - \sum_{i=1}^{N} \texttt{Corr}(\mathbf{p}_i, \mathbf{p}_i^{\text{DP}})$$

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



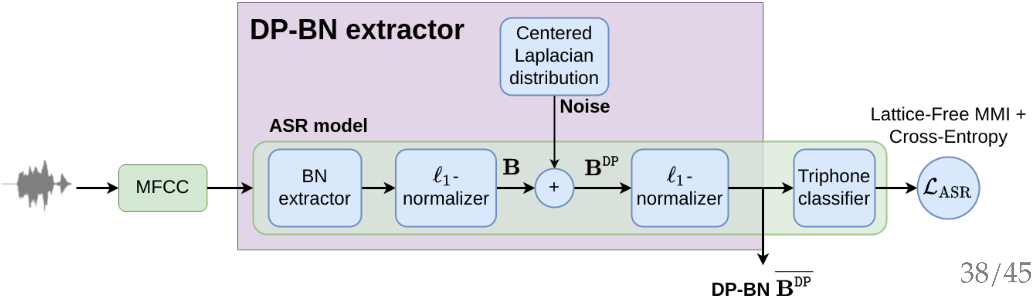
Privacy and utility of DP-Pitch features



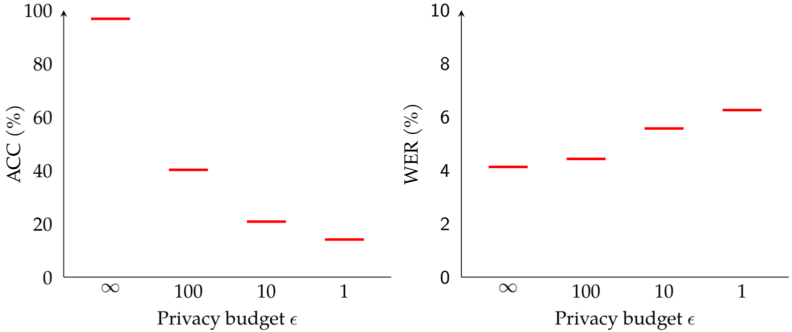
- DP-Pitch significantly reduces the speaker identification accuracy
- Pearson correlation is preserved for $\epsilon > 1$

Differentially-private BN extractor

$$\mathbf{B}^{\mathrm{DP}} = \mathcal{N}_B(\mathbf{B}) = \begin{bmatrix} \mathcal{N}_b(\mathbf{b}_1) \\ \vdots \\ \mathcal{N}_b(\mathbf{b}_T) \end{bmatrix} \qquad \mathcal{N}_B(\mathbf{b}) = \frac{\mathbf{b}}{\|\mathbf{b}\|_1} + \mathrm{Lap}(2/\epsilon)$$



Privacy and utility of DP-BN features



- DP-BN significantly reduces speaker identification accuracy
- Gradual decline of utility as ϵ increases

Combination of DP-BN and DP-Pitch features

Method	**Privacy**			**Utility**
	Local ϵ		*Practical*	*Practical*
	BN	Pitch	$D_{\leftrightarrow}^{sys}$	WER
Without DP (part 2)	∞	∞	0.14	6.8%
With DP	100	1.0	0.11	5.8%
With DP	100	0.1	0.10	5.6%
With DP	10	1.0	0.13	6.5%
With DP	10	0.1	0.13	6.4%
With DP	1	1.0	0.12	7.0%
With DP	1	0.1	0.10	6.7%

- 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 ϵ, 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

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!