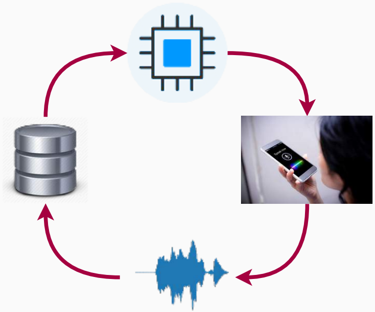
DIFFERENTIALLY PRIVATE SPEAKER ANONYMIZATION

Aurélien Bellet (Inria)

Joint work with **A. Shahin Shamsabadi**, **B. Srivastava**, **N. Vauquier**, **E. Vincent**, **M. Maouche**, **M. Tommasi** and **N. Papernot**

Privaski
March 8, 2022

CONTEXT: VOICES INTERFACES



- Massive collection of speech by service providers and third-party contractors[1] to:
  - Process user queries (*inference*)
  - Train Automatic Speech Recognition (ASR) systems (*training*)

[1] https://www.bbc.com/news/technology-31296188

SPEECH DATA IS VERY SENSITIVE

Speech data contains a wealth of personal information:

- Linguistic content (*what is being said*)
- Speaker information (*who is saying it*)
    - **Identity**: voice is a biometric modality. In [Srivastava et al., 2021] we show that a standard speaker recognition system reaches top-1 precision above 50% in a crowd of 10k speakers
    - Other paralinguistic and extra-linguistic speaker information [Schuller and Batliner, 2013] such as age, gender, accent, emotional state, personality traits, health status...

PRIVACY IN SPEECH: A PRESSING ISSUE

- Recent guidelines on voice assistants emphasize importance of privacy and security
  - 2020: CNIL white paper on ethical, technical and legal issues of voice assistants
  - 2021: EDPB guidelines on virtual voice assistants
- Several initiatives in the speech processing community in the last 2 years:
  - Special interest group of the International Speech Communication Association[2]
  - VoicePrivacy initiative [Tomashenko et al., 2020]
  - Ongoing efforts to understand the requirements of effective privacy preservation for speech [Nautsch et al., 2019b] in light of recent regulation [Nautsch et al., 2019a]

[2] https://www.spsc-sig.org

TODAY'S FOCUS: SPEAKER ANONYMIZATION

Speaker anonymization[3] aims to transform speech so as to
conceal the speaker's identity while
preserving the linguistic and prosodic content and diversity of speech

- This was the focus of the recent VoicePrivacy Challenge [Tomashenko et al., 2022]
- A successful speaker anonymization scheme enables people to freely share their speech data for both inference and training purposes, while concealing their identity
- It does **not** address the complementary objective of protecting personally identifiable information in the linguistic content (see e.g., [Ahmed et al., 2020])

[3]Note: the term "anonymization" refers to the ideal objective

PROBLEM STATEMENT

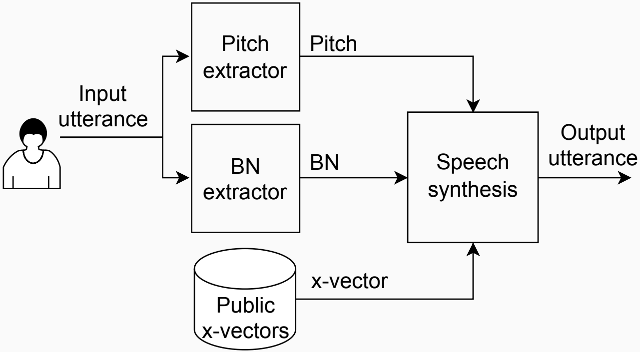
A speaker anonymization scheme

- outputs an intelligible speech waveform (so it can be annotated by humans)
- preserves as well as possible phonetic and prosodic content (*utility*)
- conceals as well as possible the identity of the speaker (*privacy*)

Threat model [Srivastava et al., 2020b]

- The adversary wants to know if a given speaker spoke a target anonymized utterance
- The adversary has access to raw speech utterances from the hypothesized speaker as well as to a large public speech corpus with speaker labels
- The speaker anonymization scheme is public (but not its internal randomness)

STATE-OF-THE-ART ARCHITECTURE [Fang et al., 2019, Srivastava et al., 2020a]



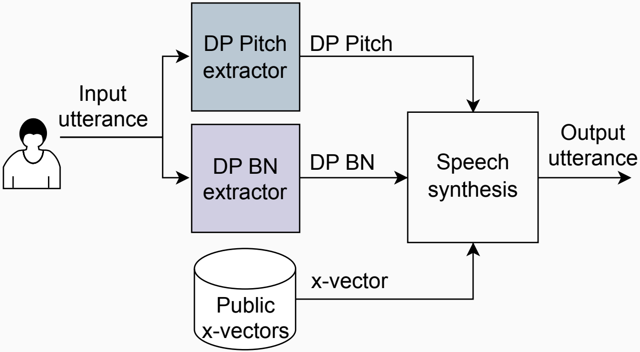
1. Extract prosodic (pitch) and linguistic (BN) feature sequences from input utterance
2. Re-synthesize speech from pitch, BN and a public speaker embedding (x-vector)

$\rightarrow$ best method in the VoicePrivacy Challenge

LIMITATIONS OF STATE-OF-THE-ART SCHEME

1. There is still a lot of room for improvement in protecting against concrete attacks [Maouche et al., 2021]
2. Disentanglement is not perfect: pitch and BN features contain speaker information
   - We design a re-identification attack to predict speaker identity from these features
   - The accuracy of this attack is 37% with pitch and 97% with BN (among 900+ speakers)!
3. No formal privacy guarantees

PROPOSED APPROACH [Shahin Shamsabadi et al., 2022]



- Use Differential Privacy (DP) to bound the risk of the speaker identity leaking through pitch and BN features
- Choose target x-vector independently of input utterance
- Then the complete pipeline satisfies DP (by composition + post-processing)

Definition (Differential Privacy)

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

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

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

- Key properties of DP:
  - Robustness to postprocessing: if $\mathcal{A}$ is ϵ-DP, then any $g \circ \mathcal{A}$ is also ϵ-DP
  - Composition: if $\mathcal{A}_1$ is ϵ_1-DP and $\mathcal{A}_2$ is ϵ_2-DP, then $\mathcal{A} = (\mathcal{A}_1, \mathcal{A}_2)$ is $(\epsilon_1 + \epsilon_2)$-DP
- In our setting, $\mathbf{x}$ will be a speech utterance and $\mathcal{A}$ will be the speaker anonymization scheme that produces an anonymized utterance
- Note that DP is stronger than what we need: it entails hiding the speaker identity but may also suppress other information that we wish to preserve

Definition (Laplace mechanism)

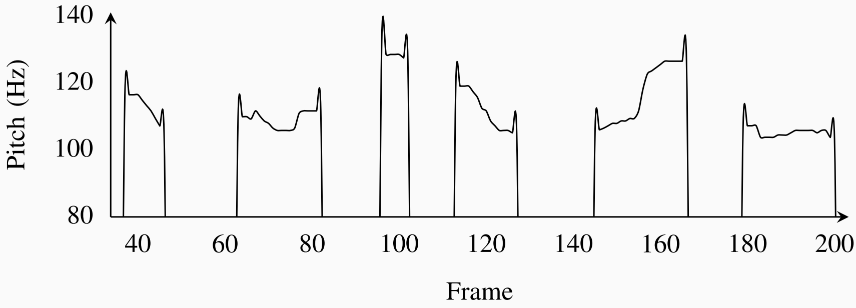
Let $f : \mathcal{X} \to \mathbb{R}^d$ and let the ℓ_1-sensitivity of f be defined as

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

Let $\eta = [\eta_1, \ldots, \eta_d] \in \mathbb{R}^d$ be a vector where each $\eta_i \sim \text{Lap}(\Delta_1(f)/\varepsilon)$ is drawn from the centered Laplace distribution with scale $\Delta_1(f)/\varepsilon$. Then, $\mathcal{A}(\cdot) = f(\cdot) + \eta$ is ε-DP.

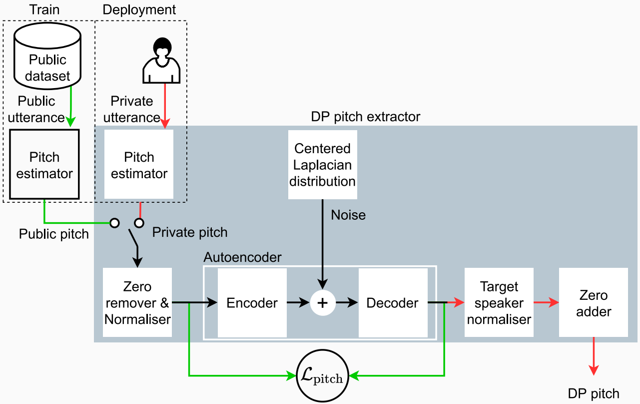
- The sensitivity $\Delta_1(f)$ measures how much changing the input can affect the value of f
- To satisfy ϵ-DP, the Laplace noise is calibrated to $\Delta_1(f)$ and ϵ

PITCH SEQUENCE



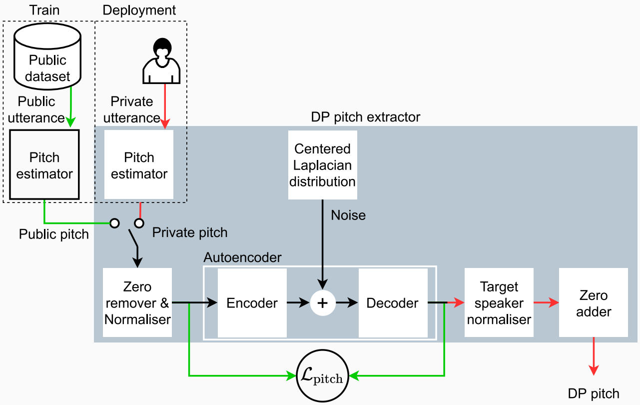
- Global dynamics are related to sentence prosody
- Local variations are known to be more speaker-specific (see e.g., [Dehak et al., 2007, Mary and Yegnanarayana, 2008])

DP PITCH EXTRACTOR

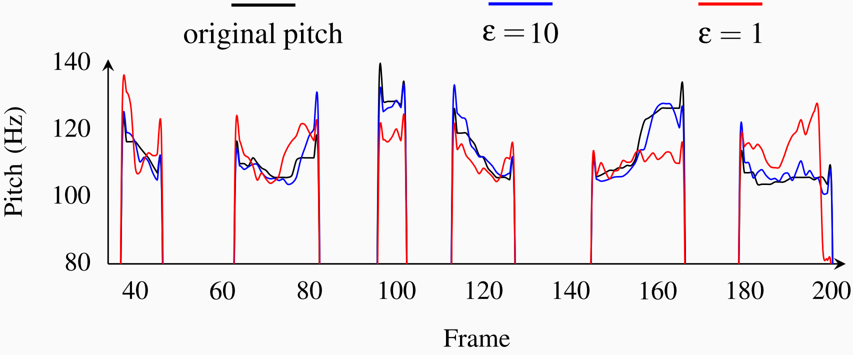


- Our fully convolutional autoencoder $\mathcal{A} = \mathcal{D} \circ \mathcal{N}_p \circ \mathcal{E}$ takes input pitch $\mathbf{p} \in \mathbb{R}^K$ and:
    1. Maps it to a latent representation $\mathbf{h} = \mathcal{E}(\mathbf{p}) \in [0, 1]^{C \times K}$ using convolutional layers
    2. Generates a perturbed $\mathbf{h}^{\text{DP}} = \mathcal{N}_p(\mathbf{h}) = \mathbf{h} + \text{Lap}(CK/\varepsilon)$
    3. Decodes it into a perturbed pitch sequence $\mathbf{p}^{\text{DP}} = D(\mathbf{h}^{\text{DP}}) \in \mathbb{R}^K$ using convolutional layers

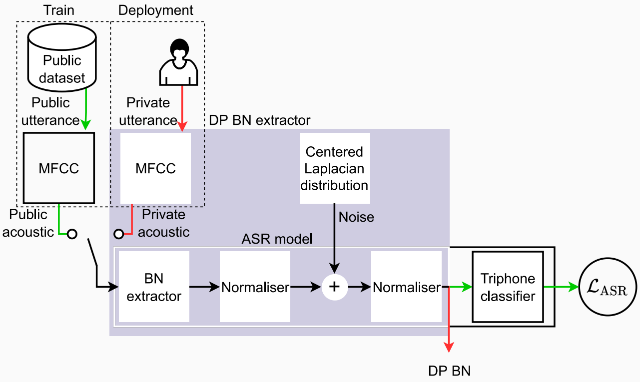
DP PITCH EXTRACTOR



- Training phase on public speech: train autoencoder to maximize correlation between input and reconstructed pitch
- Deployment phase: generate perturbed pitch and normalize it to target speaker



- By maximizing correlation, the autoencoder learns to preserve global dynamics as much as possible while sacrificing local variations, as desired
- By the Laplace mechanism, $\mathcal{N}_p \circ \mathcal{E}$ satisfies ε-DP, and so does the autoencoder $\mathcal{A} = \mathcal{D} \circ \mathcal{N}_p \circ \mathcal{E}$ by the post-processing property of DP

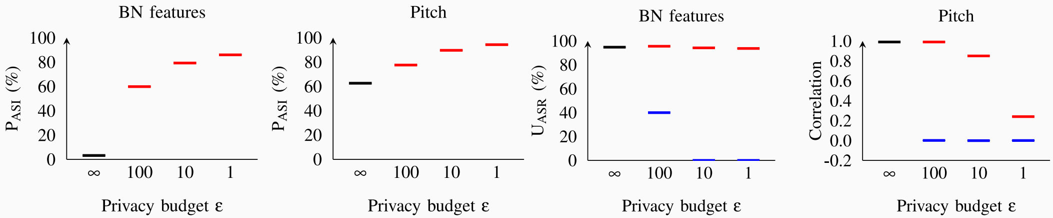


- BN features are typically obtained as an intermediate layer of an ASR acoustic model
- We add a noise layer and train on public speech to maximize ASR performance
- We used the same architecture and training objective as in VPC baseline

EXPERIMENTAL SETUP

- Librispeech dataset, essentially follow VPC setup
- X-vector selection: utterance-level, variant of dense strategy [Srivastava et al., 2020a]
- Informed attackers
    - Re-identification attacks: follows standard ASI system but trained on BN and pitch instead of MFCCs
    - Speaker linkage attacks: follows standard ASV system, but trained on utterance-level assignment which gives a stronger attack (see also [Maouche et al., 2021])

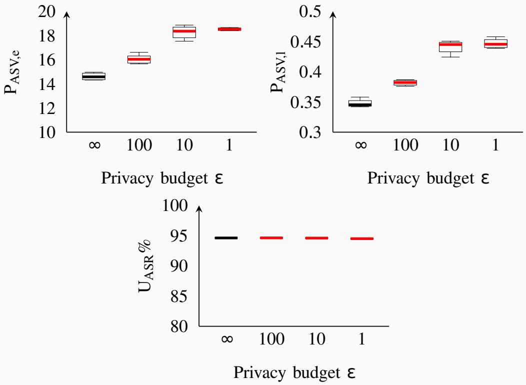
RESULTS — PRIVACY AND UTILITY OF PITCH AND BN

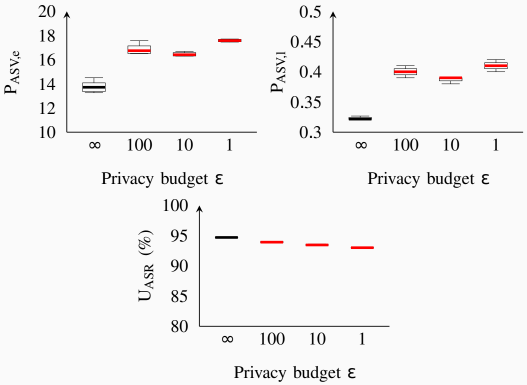


- Our DP extractors largely improve the protection against re-identification attacks from pitch and BN features (P_{ASI} : error of attack)
- Our DP extractors preserve utility (U_{ASR}: ASR performance), unlike naive DP baselines

Method	Privacy				Utility
	Analytical (ε)		Empirical		Empirical
	BN	Pitch	Equal Error Rate	Unlinkability	U_{ASR} (%)
Anon (state-of-the-art)	∞	∞	$14.62 \pm .25$	$.35 \pm .01$	$94.64 \pm .06$
Anon+DP (ours)	100	100	$24.22 \pm .44$	$.57 \pm .01$	$94.00 \pm .10$
Anon+DP (ours)	10	10	$27.68 \pm .25$	$.65 \pm .01$	$93.01 \pm .07$
Anon+DP (ours)	1	1	$29.98 \pm .76$	$.70 \pm .01$	$92.16 \pm .05$

- Empirical privacy is evaluated by the performance of a speaker verification attack trained on anonymized speech
- Utility is evaluated by the performance of ASR system trained and tested on anonymized speech
- Our approach provides twice better empirical privacy at a negligible cost in utility





- Left: `Anon+DP_Pitch` vs. `Anon+PC`; Right: `Anon+DP_BN` vs. `Anon`
- Reducing speaker information in *both* pitch and BN features provides a large gain

DISCUSSION

- Large gap between analytical and empirical privacy guarantees
  - Reported ε is frame-level for BN features $\rightarrow$ weak sequence-level guarantee
  - This gap is expected and in line with other findings on learning with DP [Nasr et al., 2021]
  - Could bound the analytical privacy more tightly
  - Design appropriate relaxations of DP for speaker anonymization?
- Better utility measures
  - Human intelligibility, naturalness and diversity of anonymized utterances
  - Correlation is merely a proxy for the utility of pitch $\rightarrow$ prediction of prosodic attributes?
- Concealing other speaker information with DP
  - Gender, age, emotions, etc...
  - Tools that let the user choose what to protect depending on the context?

[Ahmed et al., 2020] Ahmed, S., Chowdhury, A. R., Fawaz, K., and Ramanathan, P. (2020).
Preech: A system for privacy-preserving speech transcription.
In *Proceedings of the USENIX Security Symposium*, Virtual Event.

[Dehak et al., 2007] Dehak, N., Dumouchel, P., and Kenny, P. (2007).
Modeling prosodic features with joint factor analysis for speaker verification.
IEEE Transactions on Audio, Speech, and Language Processing, 15(7):2095–2103.

[Fang et al., 2019] Fang, F., Wang, X., Yamagishi, J., Echizen, I., Todisco, M., Evans, N., and Bonastre, J.-F. (2019).
Speaker anonymization using x-vector and neural waveform models.
In *Proceedings of the ISCA Speech Synthesis Workshop*, Vienna, Austria.

[Maouche et al., 2021] Maouche, M., Srivastava, B. M. L., Vauquier, N., Bellet, A., Tommasi, M., and Vincent, E. (2021).
Enhancing Speech Privacy with Slicing.
preprint.

[Mary and Yegnanarayana, 2008] Mary, L. and Yegnanarayana, B. (2008).
Extraction and representation of prosodic features for language and speaker recognition.
Speech Communication, 50(10):782–796.

[Nasr et al., 2021] Nasr, M., Song, S., Thakurta, A., Papernot, N., and Carlini, N. (2021).
Adversary instantiation: Lower bounds for differentially private machine learning.
In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, San Francisco, CA, USA.

[Nautsch et al., 2019a] Nautsch, A., Jasserand, C., Kindt, E., Todisco, M., Trancoso, I., and Evans, N. W. D. (2019a).
The GDPR & speech data: Reflections of legal and technology communities, first steps towards a common understanding.
In *Proceedings of the conference of the International Speech Communication Association (Interspeech)*, Graz, Austria.

[Nautsch et al., 2019b] Nautsch, A., Jiménez, A., Treiber, A., Kolberg, J., Jasserand, C., Kindt, E., Delgado, H., Todisco, M., Hmani, M. A., Mtibaa, A., Abdelraheem, M. A., Abad, A., Teixeira, F., Matrouf, D., Gomez-Barrero, M., Petrovska-Delacrétaz, D., Chollet, G., Evans, N. W. D., and Busch, C. (2019b).
Preserving privacy in speaker and speech characterisation.
Computer Speech and Language, 58:441–480.

[Schuller and Batliner, 2013] Schuller, B. and Batliner, A. (2013).
Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing.
Wiley.

[Shahin Shamsabadi et al., 2022] Shahin Shamsabadi, A., Mohan Lal Srivastava, B., Bellet, A., Vauquier, N., Vincent, E., Maouche, M., Tommasi, M., and Papernot, N. (2022).
Differentially Private Speaker Anonymization.
Technical report, arXiv:2202.11823.

[Srivastava et al., 2021] Srivastava, B. M. L., Maouche, M., Sahidullah, M., Vincent, E., Bellet, A., Tommasi, M., Tomashenko, N., Wang, X., and Yamagishi, J. (2021).
Privacy and utility of x-vector based speaker anonymization.
preprint.

[Srivastava et al., 2020a] Srivastava, B. M. L., Tomashenko, N. A., Wang, X., Vincent, E., Yamagishi, J., Maouche, M., Bellet, A., and Tommasi, M. (2020a).
Design choices for x-vector based speaker anonymization.
In *Proceedings of the conference of the International Speech Communication Association (Interspeech)*, Virtual Event.

[Srivastava et al., 2020b] Srivastava, B. M. L., Vauquier, N., Sahidullah, M., Bellet, A., Tommasi, M., and Vincent, E. (2020b).
Evaluating voice conversion-based privacy protection against informed attackers.
In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain.

[Tomashenko et al., 2020] Tomashenko, N. A., Srivastava, B. M. L., Wang, X., Vincent, E., Nautsch, A., Yamagishi, J., Evans, N. W. D., Patino, J., Bonastre, J., Noé, P., and Todisco, M. (2020).
Introducing the VoicePrivacy initiative.
In *Proceedings of the conference of the International Speech Communication Association (Interspeech)*, Virtual Event.

[Tomashenko et al., 2022] Tomashenko, N. A., Wang, X., Vincent, E., Patino, J., Srivastava, B. M. L., Noé, P., Nautsch, A., Evans, N. W. D., Yamagishi, J., O'Brien, B., Chanclu, A., Bonastre, J., Todisco, M., and Maouche, M. (2022).
The VoicePrivacy 2020 Challenge: Results and findings.
Computer Speech & Language.