Privacy in Speech Processing

Brij Mohan Lal Srivastava
PhD research scholar
Inria, Nancy

Laboratoire d'Informatique de l'Université du Mans (LIUM), Le Mans
24 January, 2020

Collaborators

Supervisors:

Aurélien Bellet (Magnet, Inria Lille)

Marc Tommasi (Magnet, Université de Lille)

Emmanuel Vincent (Multispeech, Inria Nancy)

Other collaborators:

Nathalie Vauquier (Magnet, Inria Lille)

Md Sahidullah (Multispeech, Inria Nancy)

Overview

- Privacy - background
- Objectives of anonymization
- Some previous approaches
- Our approaches
    - Adversarial training
    - Voice conversion
- Attacker
- Voice Privacy Challenge
- Conclusion

Privacy

There is not a single or universal legal definition of “privacy” [1].

First legal definition by Warren and Brandeis, “the right to be let alone or free from intrusion”.

HARVARD

LAW REVIEW.

VOL. IV. DECEMBER 15, 1890. NO. 5.

THE RIGHT TO PRIVACY.

[1] Computer Speech & Language (Jun 2019), *Preserving Privacy in Speaker and Speech Characterisation*, Nautsch et al.

Four types of privacy

US Constitution (incl. the Fourth Amendment) defines 4 distinct types of privacy [2]

1. Physical/Accessibility : *non-intrusion involving one’s physical space*
2. Decisional : *non-interference involving one’s choices*
3. Psychological/Mental : *non-intrusion/interference involving one’s thoughts or identity*
4. Informational : *limiting access to one’s personal information (data privacy)*

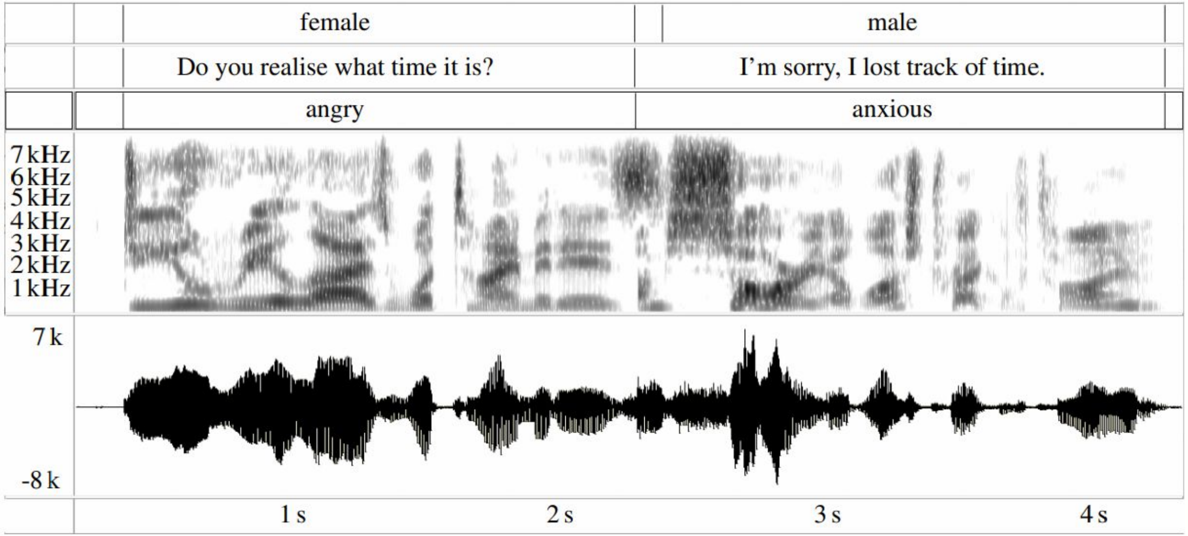
[2] The Handbook of Information and Computer Ethics (2008), *Informational Privacy: Concepts, Theories, and Controversies,* Herman T. Tavani.

GDPR

At the EU level:

- General Data Protection Regulation (Regulation 2016/679)
- ‘Police’ directive (Directive 2016/680)
- Defines “biometric data” as data which allows or confirms the unique identification of that natural person.

Why privacy in speech processing?



Rich in information: speaker's identity, gender, emotional state, pathological conditions, intention, personality, race and culture.

[3] The GDPR & Speech Data: Reflections of Legal and Technology Communities, First Steps towards a Common Understanding; *Nautsch et al.* Proc Interspeech 2019

Previous approaches (limitations)

- Voice conversion and cryptographic approaches were conventionally investigated.
- “Found data” must be rendered neutral due to advances in voice cloning.
- De-identification vs Anonymization
- Strict evaluation criteria must be enforced not “security by obscurity”

CYBERTRUST

“Alexa, Can I Trust You?”

Hyunji Chung, Michaela Iorga, and Jeffrey Voas, NIST
Sangjin Lee, Korea University

Consumer Attitudes Towards Privacy and Security in Home Assistants

Can we steal your vocal identity from the Internet?: Initial investigation of cloning Obama’s voice using GAN, WaveNet and low-quality found data

Jaime Lorenzo-Trueba[1], Fuming Fang[1], Xin Wang[1], Isao Echizen[1], Junichi Yamagishi[1,2], Tomi Kinnunen[3]

[1] National Institute of Informatics, Tokyo, Japan [2] University of Edinburgh, Edinburgh, U
[3] University of Eastern Finland, Joensuu, Finland
{jaime, fang, wangxin, iechizen, jyamagis}@nii.ac.jp, tkinnu@cs.uef.fi

Manas A. Pathak, Bhiksha Raj, Shantanu Rane, and Paris Smaragdis

Privacy-Preserving Speech Processing

PRIVACY PRESERVING ENCRYPTED PHONETIC SEARCH OF SPEECH DATA

Cornelius Glackin[1], Gerard Chollet[1], Nazim Dugan[1], Nigel Cannings[1], Julie Wall[2], Shahzaib Tahir[3], Indranil Ghosh Ray[3], and Muttukrishnan Rajarajan[3]*

[1] Intelligent Voice Ltd., London, UK [2] University of East London, London, UK [3] City University London, London, UK
Email: neil.glackin@intelligentvoice.com

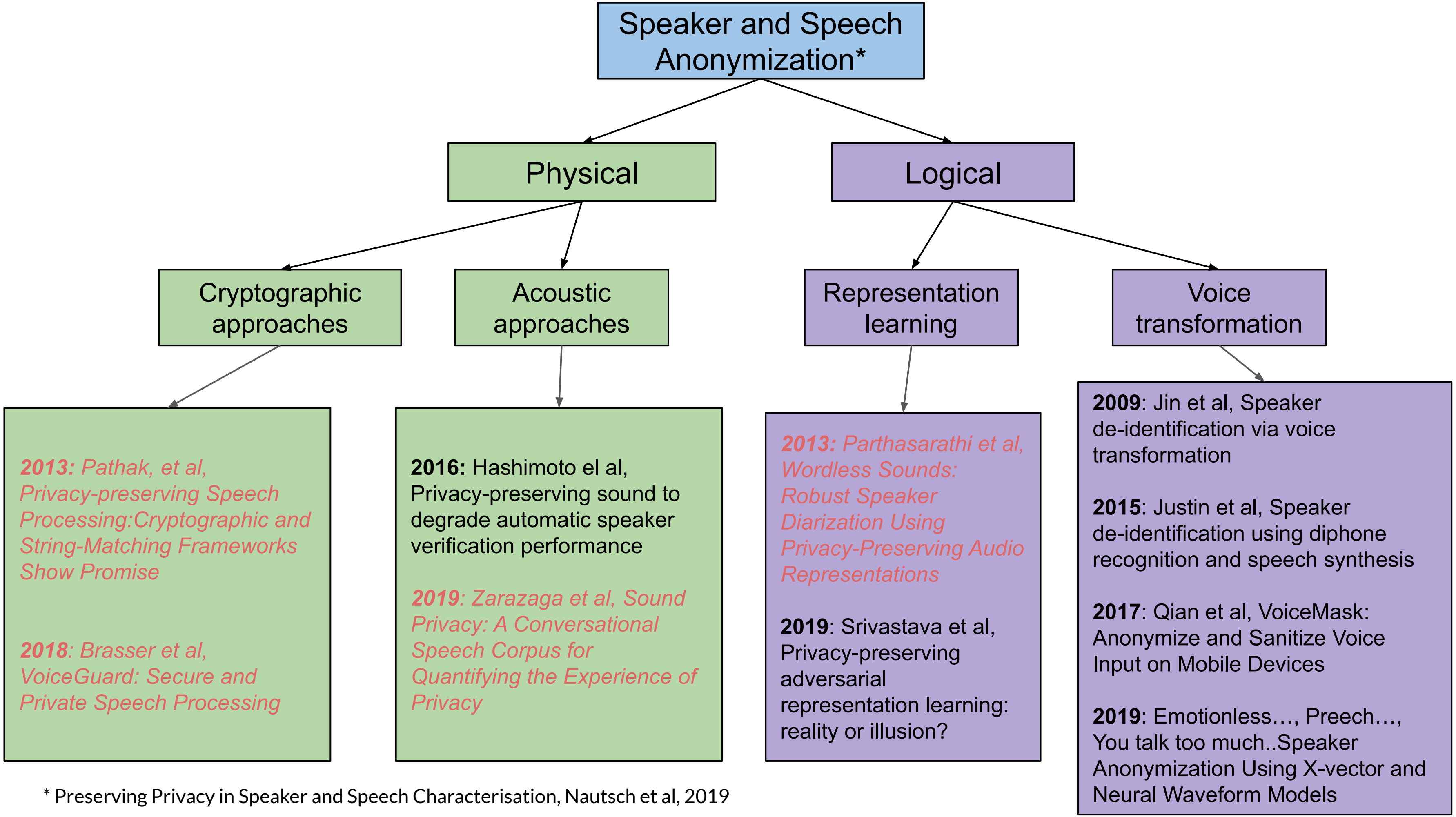
Two objectives of anonymization

- User must have complete control over the sharing of sensitive attributes of speech with the service provider.
  - Application level permission must be granted
  - Disentanglement of attributes must be done
- Anonymization should not affect the utility of speech, e.g. linguistic variability and content.
  - Output must be usable for further processing, e.g. pitch extraction, phonetic analysis, etc.
  - Output must be intelligible and suitable for annotation and training of automatic speech recognition (ASR) systems.

Speech vs speaker anonymization

Speech anonymization deals with non-biometric yet sensitive attributes, for instance: bank details in the spoken text.

Speaker anonymization deals with biometric attributes, such as speaker’s identity, personality traits, gender, race, etc.



* Preserving Privacy in Speaker and Speech Characterisation, Nautsch et al, 2019

Our approach to anonymize speaker’s identity

1. Representation learning:
   a. Removing speaker-specific features from bottleneck representation of ASR through adversarial training.
   b. Noisy representation for ASR to hide speaker information using differentially private noise
2. Voice conversion: Anonymize identity by transforming into random pseudo-speakers

Motivation: Adversarial approach

Shown to learn a representation which:

1. is speaker-invariant.
2. performs well for ASR task.
3. allows ASR by a third party.

Following the literature of **speaker invariance** in different context (bottleneck features, traditional models, ...): ICASSP 2018.

SPEAKER INVARIANT FEATURE EXTRACTION FOR ZERO-RESOURCE LANGUAGES WITH ADVERSARIAL LEARNING

Taira Tsuchiya, Naohiro Tawara, Testuji Ogawa and Tetsunori Kobayashi

Department of Communications and Computer Engineering, Waseda University, Tokyo, Japan

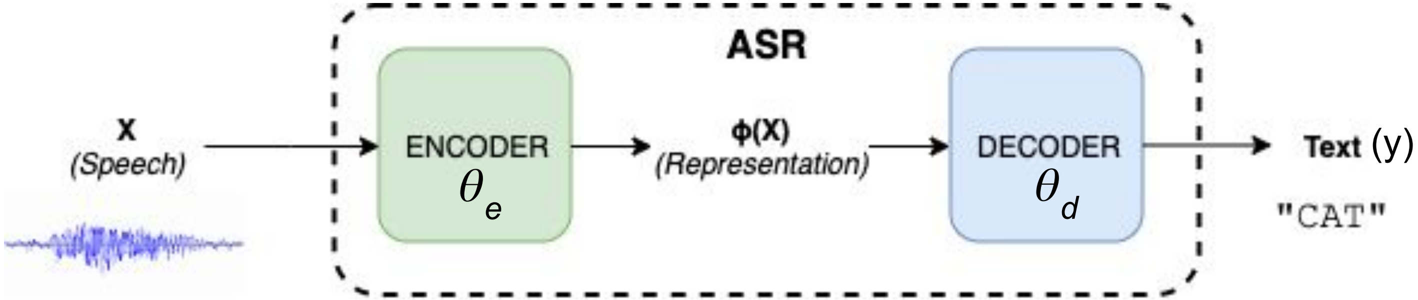
SPEAKER-INVARIANT TRAINING VIA ADVERSARIAL LEARNING

Zhong Meng[1,2], Jinyu Li[1], Zhuo Chen[1], Yong Zhao[1], Vadim Mazalov[1], Yifan Gong[1], Biing-Hwang (Fred) Juang[2]*

[1] Microsoft AI and Research, Redmond, WA, USA
[2] Georgia Institute of Technology, Atlanta, GA, USA

Adversarial approach

Conventional end-to-end speech recognition



$$P(y|x) = \prod_t P(y_t|x, y_{1:t-1})$$

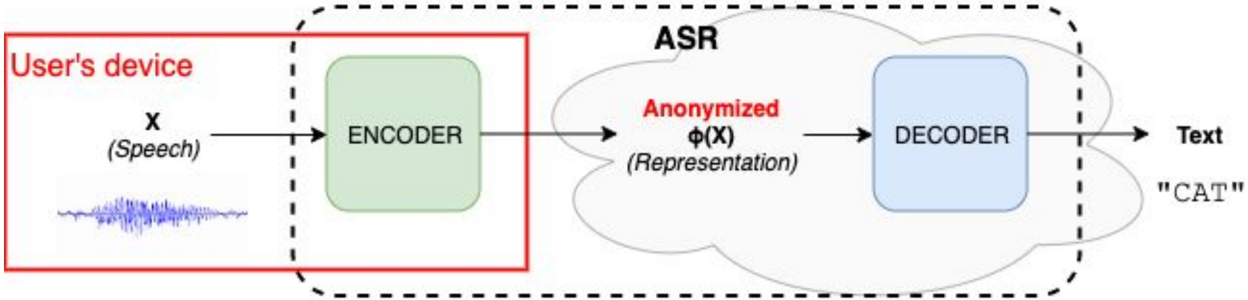
$$\phi = Encoder(x)$$

$$y_t = Decoder(\phi, y_{1:t-1})$$

$$L_{asr}(\theta_e, \theta_d) = -\sum_t \ln P(y_t^*|x, y_{1:t-1}^*)$$

Third party ASR decoding

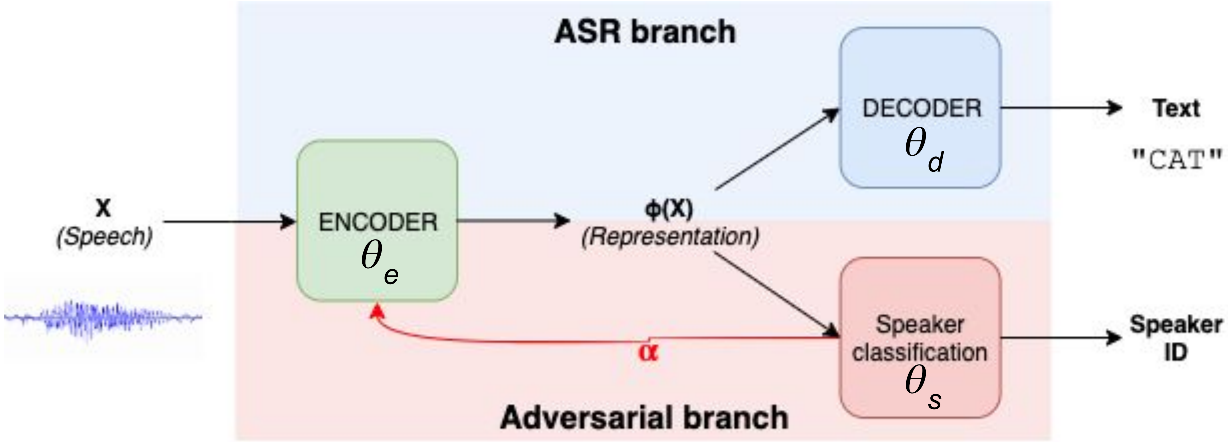
- Speaker anonymization will be performed on device
- Anonymized representation would be sent to the server for decoding



Adversarial anonymization...

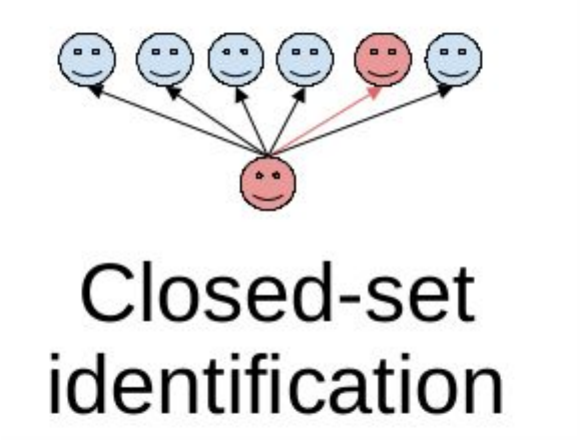
Gradients from adversarial branch are reversed and scaled by α.

Scheduling: α starts from a small value and slowly grows to a constant value.

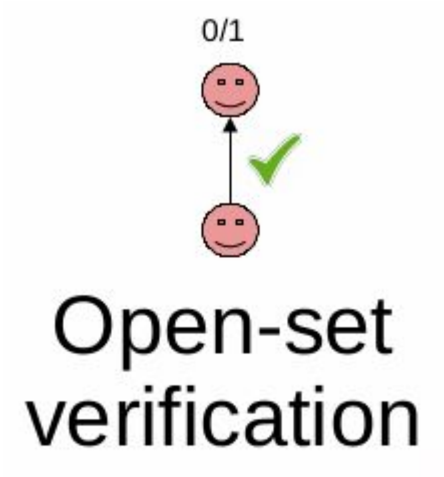


$$\min_{\theta_e,\theta_d}\max_{\theta_s} L_{asr}(\theta_e,\theta_d) - \alpha L_{spk}(\theta_e,\theta_s)$$

Attacker scenarios - evaluation schemes



Inside the adversarial ASR

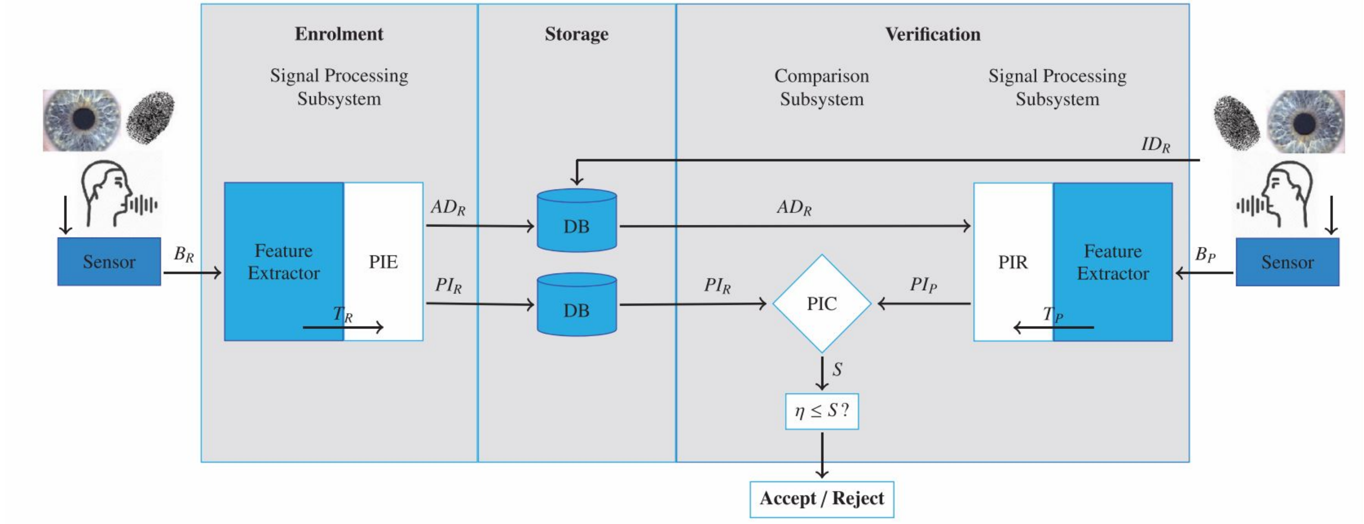


X-Vector based Speaker Verification

Open-set evaluation based on ISO standard

ISO/IEC 24745 prescribes a "biometric information protection" scheme, which involves

- Enrollment of biometric identity,
- Storage, and
- Verification using relevant scoring mechanism.



* Preserving Privacy in Speaker and Speech Characterisation, Nautsch et al, 2019

Results (open vs closed set)

	Spectral features	α = 0	α = 10
WER (ASR)		9.40	11.30 ⬆
Accuracy (closed)	97.22	48.63 ⬇	5.60 ⬇
EER (open)	4.31	24.77 ⬆	25.97 ⬆

- We first computed WER at **α = 0** to get a fair baseline, then trained over this network with **α = 10.**
- Adversary architecture is similar to open-set architecture.
- WER increases slightly indicating bearable utility loss.
- The speaker recognition accuracy (closed-set) decreases significantly.
- The speaker verification error (informed attacker) only increases slightly indicating that adversarial training does not immediately generalize over unseen speakers.

Lessons learnt and future direction

- Significant privacy gain in closed-set with little loss of utility.
- Unstable and require careful hyperparameter tuning.
- A single adversary may not be enough for adequate generalization, multiple adversaries with complexities should be investigated.
- Different scheduling strategies, eg: per-batch gradient application, hypervolume maximization.
- Establish correlation between dataset and appropriate value of $\boldsymbol{\alpha}$.
- Instance normalization for removing speaker information.
- Experiments with siamese and variational setting.

Motivation: Voice conversion approach

- Adequate literature and previous studies
- Allows publication of anonymized speech corpus
- Intuitive anonymization framework
  - Diffuse speaker’s identity among randomly selected pseudo-speakers
  - Spectrogram warping using functions with random parameters
- Requirements
  - Non-parallel
  - Many-to-many

Hidebehind: Enjoy Voice Input with Voiceprint Unclonability and Anonymity

Jianwei Qian
Illinois Institute of Technology
Chicago, IL
jqian15@hawk.iit.edu

Haohua Du
Illinois Institute of Technology
Chicago, IL
hdu4@hawk.iit.edu

Jiahui Hou
Illinois Institute of Technology
Chicago, IL
jhou11@hawk.iit.edu

Linlin Chen
Illinois Institute of Technology
Chicago, IL
lchen96@hawk.iit.edu

Taeho Jung
University of Notre Dame
Notre Dame, IN
tjung@nd.edu

Xiang-Yang Li
University of Science and Technology of China
Hefei, Anhui

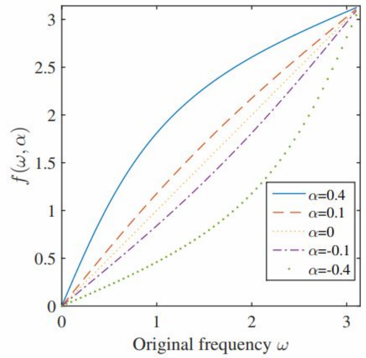
Speaker Anonymization Using X-vector and Neural Waveform Models

Fuming Fang[1], Xin Wang[1], Junichi Yamagishi[1], Isao Echizen[1], Massimiliano Todisco[2], Nicholas Evans[2], Jean-François Bonastre[3]

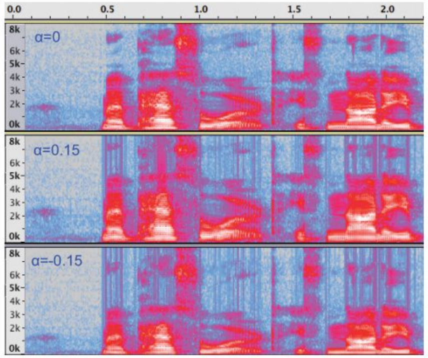
[1]National Institute of Informatics, Tokyo, Japan
[2]EURECOM, France [3]University of Avignon, France
{fang,wangxin,jyamagis,iechizen}@nii.ac.jp, {todisco,evans}@eurecom.fr, jean-francois.bonastre@univ-avignon.fr

VoiceMask

Frequency warping based on composition of **quadratic and bilinear function** using two different parameters.



(a) Bilinear functions



(b) Example of effect

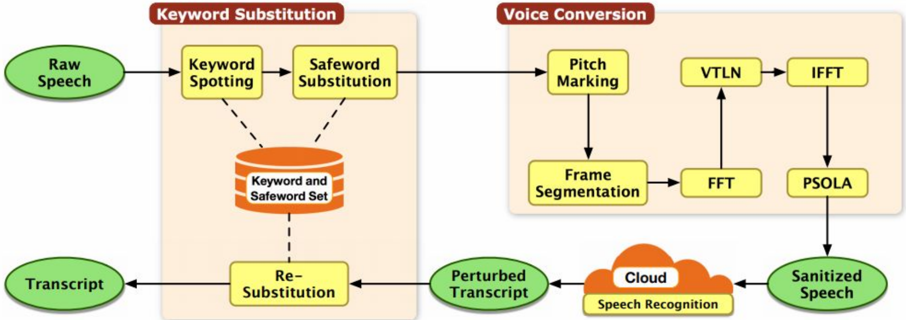


Fig. 2: **The internal architecture of VoiceMask.**

Vocal Tract Length Normalization (VTLN)

- K phonetic classes, learnt in unsupervised fashion using GMMs
- Transformation parameters are found by minimizing the distance between target class spectra and transformed source class spectra.
- K is a hyperparameter

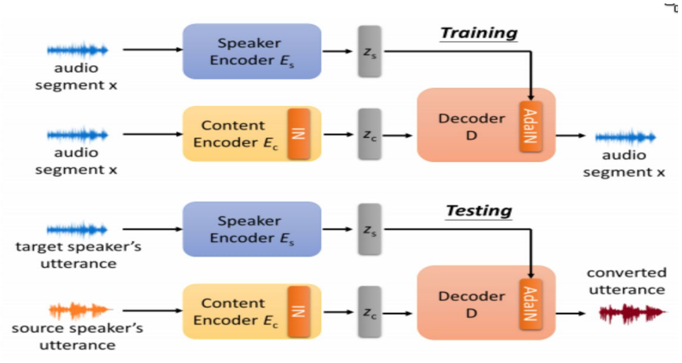
VTLN-BASED VOICE CONVERSION

David Sündermann and Hermann Ney

RWTH Aachen – University of Technology
Computer Science Department
Ahornstr. 55, 52056 Aachen, Germany
{suendermann,ney}@cs.rwth-aachen.de

Disentangled speech representations (DSR)

- Speaker information is static throughout the utterance, while content is dynamic
- Application of instance normalization in the content encoder, removes speaker information
- With a single utterance of source and target speakers, voice conversion can be performed with reasonable quality



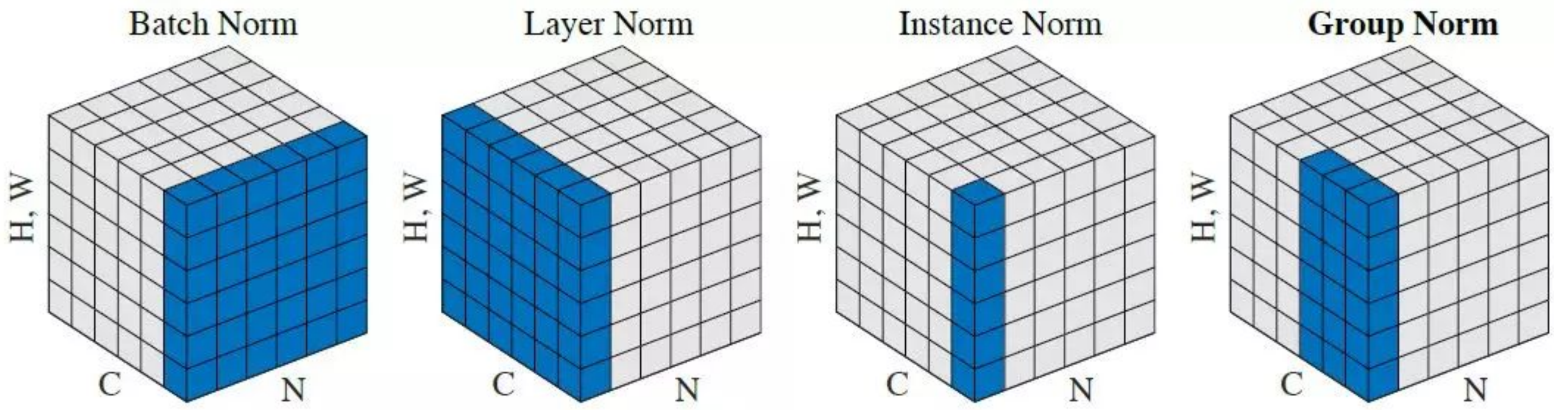
One-shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization

Ju-chieh Chou, Cheng-chieh Yeh, Hung-yi Lee

College of Electrical Engineering and Computer Science, National Taiwan University
{r06922020,r06942067,hungyilee}@ntu.edu.tw

$$M_c'[w] = \frac{M_c[w] - \mu_c}{\sigma_c}$$

Instance normalization



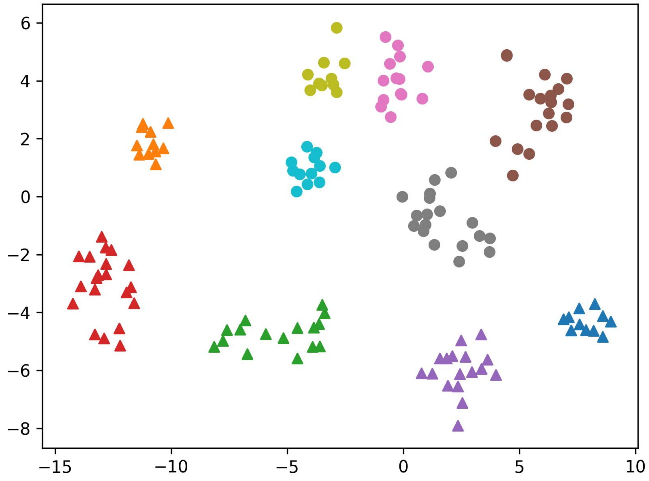
Source:

https://medium.com/syncedreview/facebook-ai-proposes-group-normalization-alternative-to-batch-normalization-fb0699bffae7

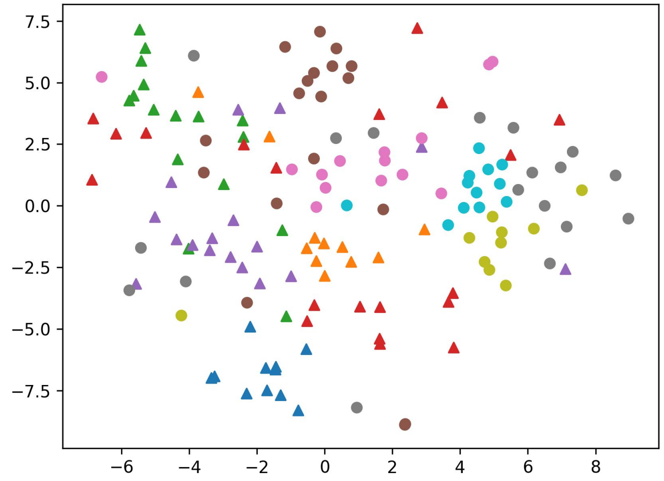
One-shot embeddings over unseen corpus

t-SNE embeddings where each speaker is represented by a unique color

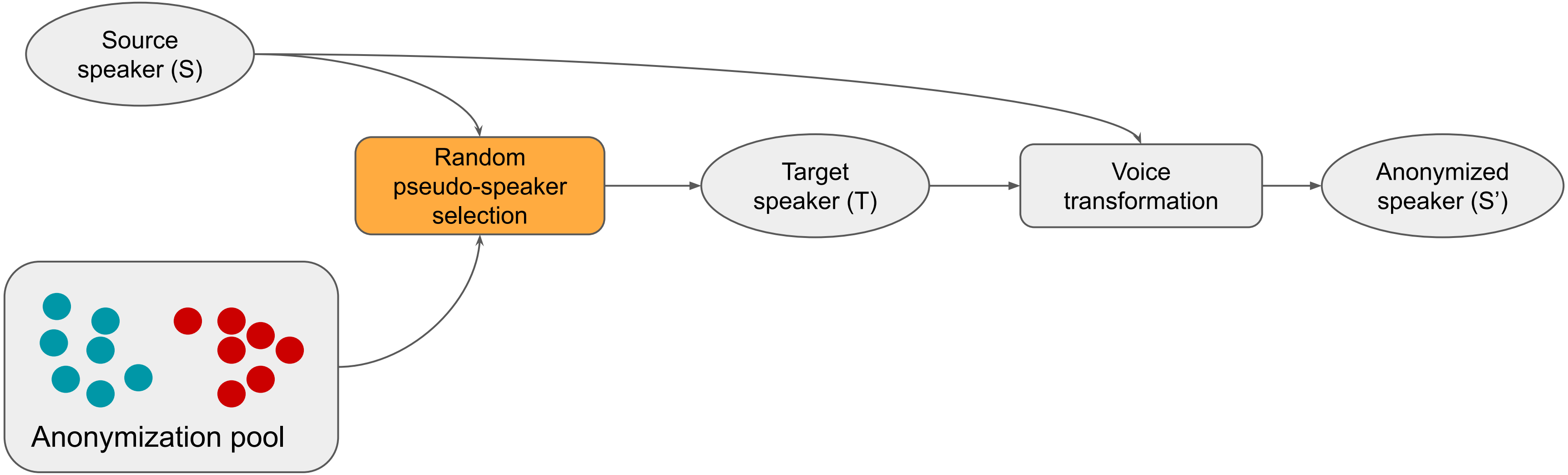
Speaker



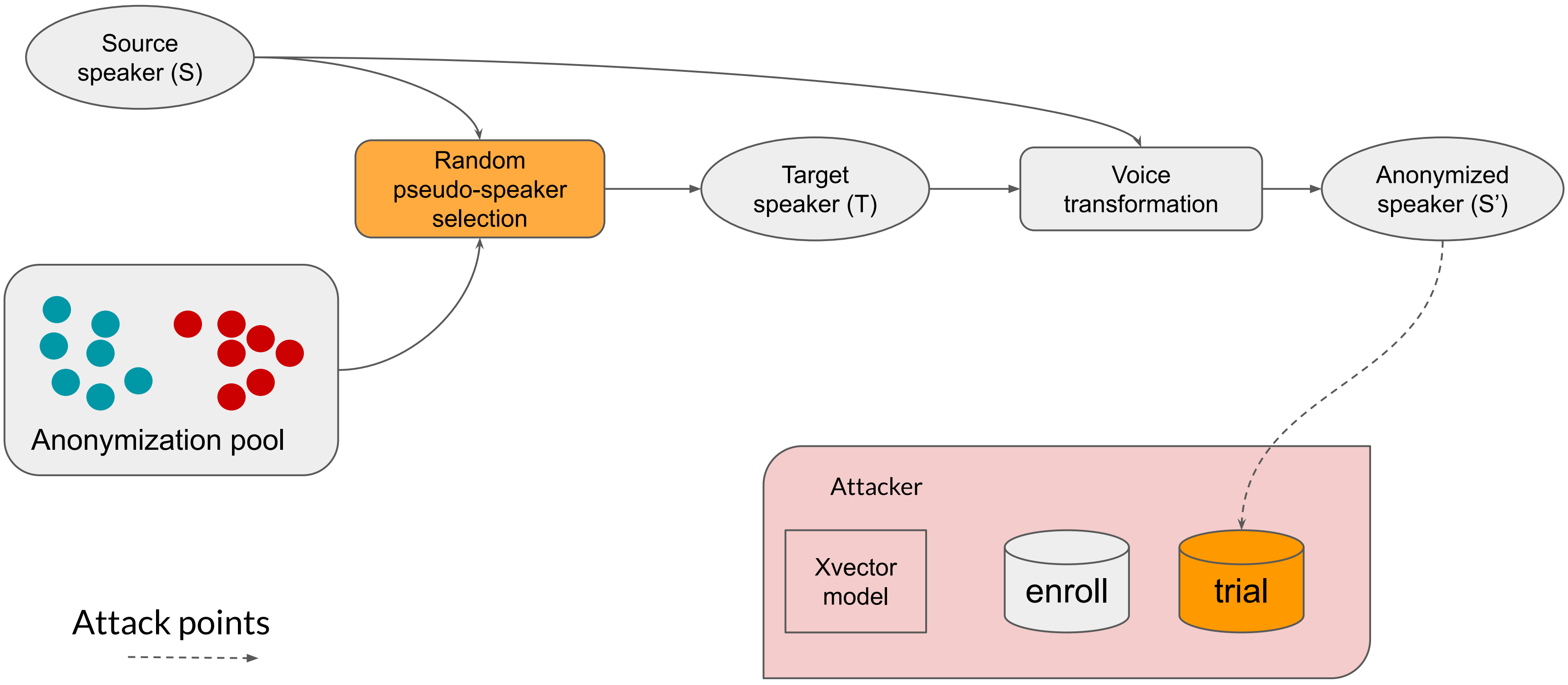
Content



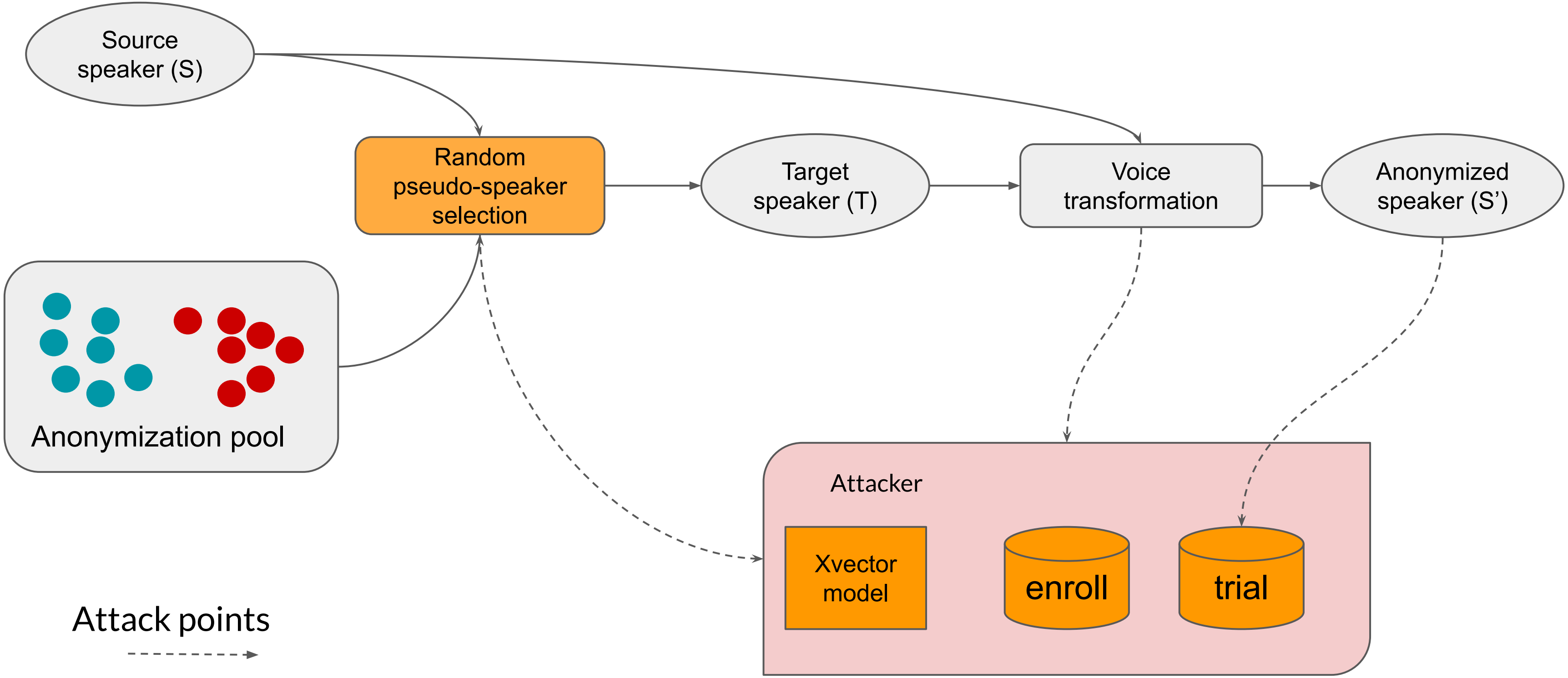
Privacy scheme



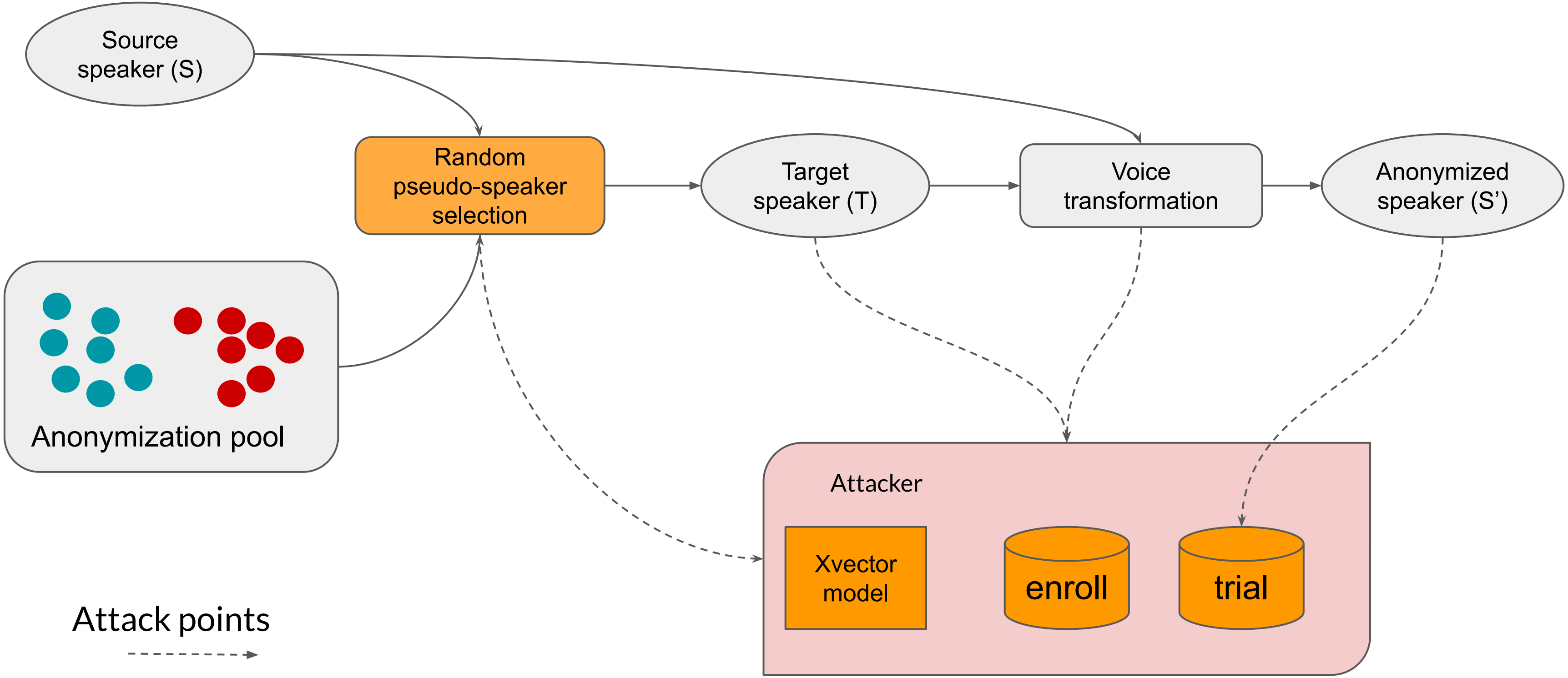
Ignorant attacker (previous studies)



Semi-informed attacker

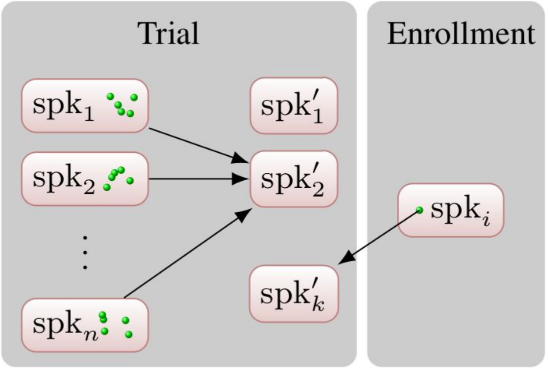


Informed attacker

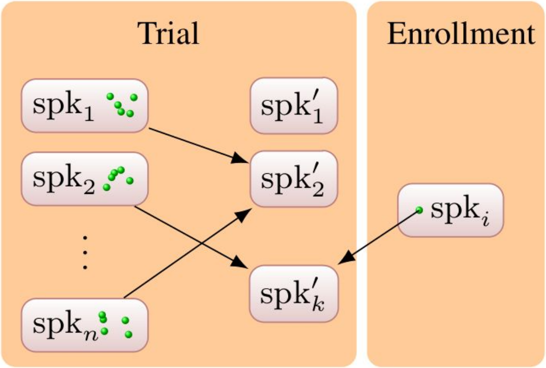


Strategies of defence...

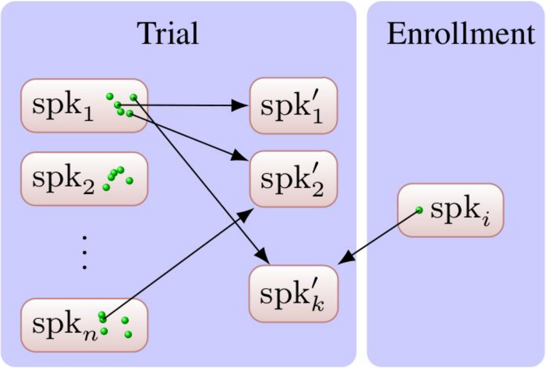
const



perm

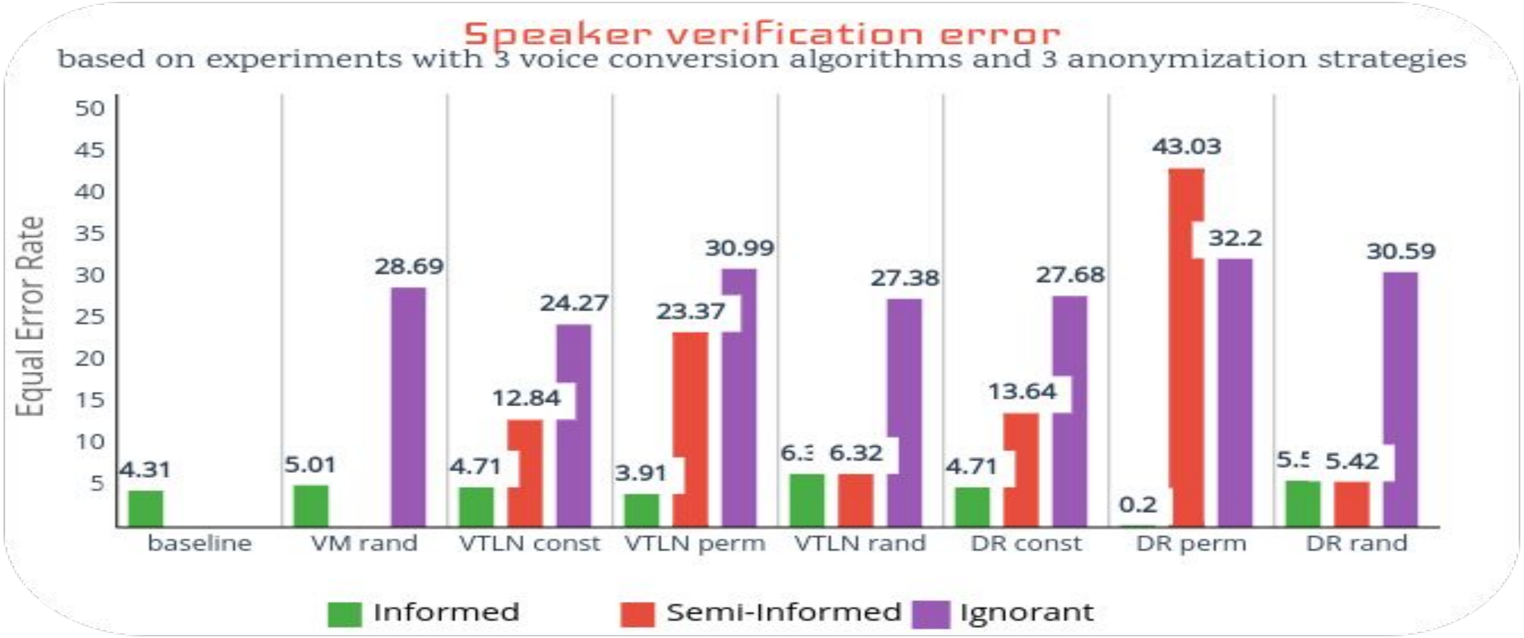


rand



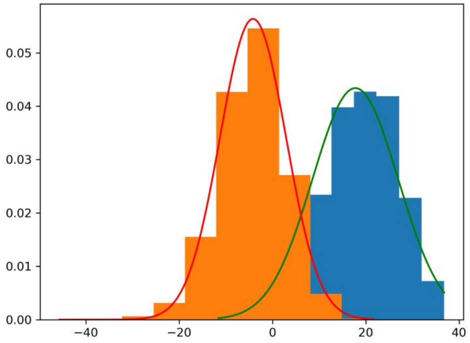
Results

Higher Equal Error Rate (EER) indicates higher privacy gain.

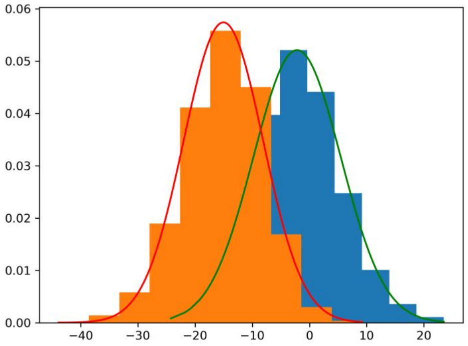


Score distribution

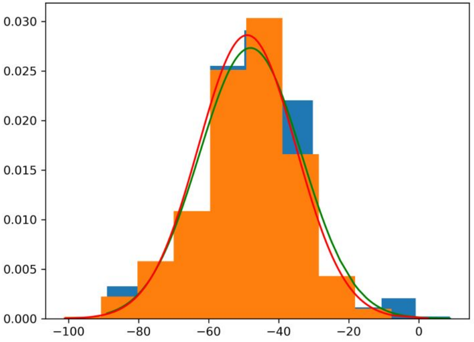
- Impostor (orange) and genuine (blue) trial scores overlap indicates higher confusion during authentication
- Informed attacker is able to authenticate speakers even after anonymization.



(a) *Informed*



(b) *Semi-Informed*



(c) *Ignorant*

Conclusion and future directions

- Authentic measure of privacy can be achieved through “informed” attacker model.
- Several attackers can be simulated based on real-world application.
- Random pseudo-speaker selection can be performed based on:
  - Gender
  - Distance metric
  - Speaker distribution
- Investigate if the anonymization can scale to multiple languages.

Summary

- There is little or no synchronization between legal and technical experts of privacy, at least in the domain of speech processing.
- Reviewed some previous studies related to speaker anonymization
- Anonymization must empower the user to take control over sensitive attributes and allow corporations to publish data safely.
- Adversarial representation learning is promising for a distributed ASR setup.
- Voice conversion based anonymization allows private data publishing to some extent.
- Strict evaluation protocols must be enforced to authentically measure the privacy gain.

Voice Privacy Challenge

The challenge is to develop anonymization solutions which suppress personally identifiable information contained within speech signals.

Using freely available datasets.

https://www.voiceprivacychallenge.org/

Baseline recipe available at:

https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2020

Organized by:

Thanks for your attention!

More details on :

https://brijmohan.github.io/

Email : brij.srivastava@inria.fr