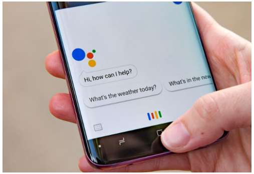
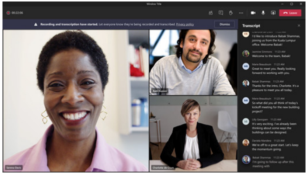
Speech privacy

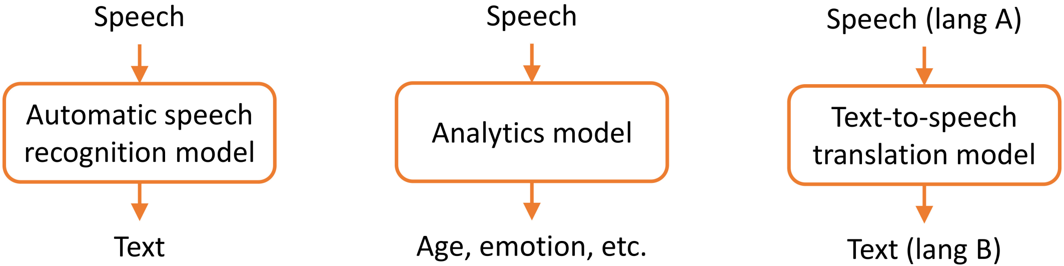
Emmanuel Vincent
Inria Nancy – Grand Est

Speech technologies



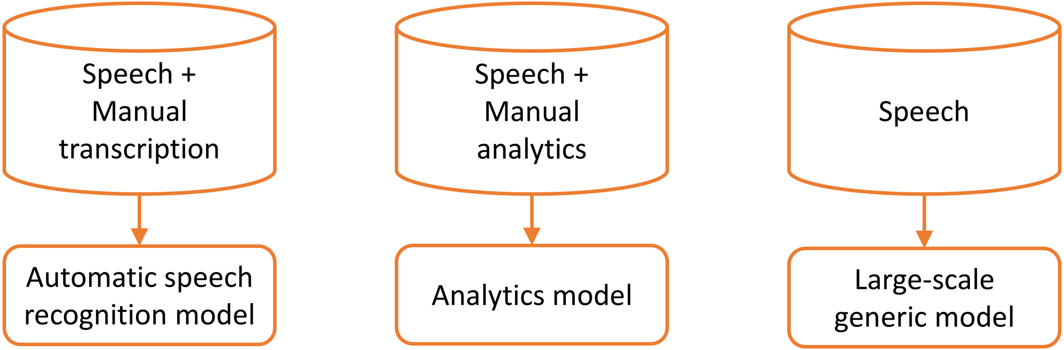


Inference = process one user utterance to answer their request.



Model = computation performed. Usually specified by a set of numerical values (e.g., neural network).

Training = (annotate and) process utterances from many users to improve the model.



Which information is conveyed?

Speech signals convey personal 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

Models and model outputs may also convey the same information.

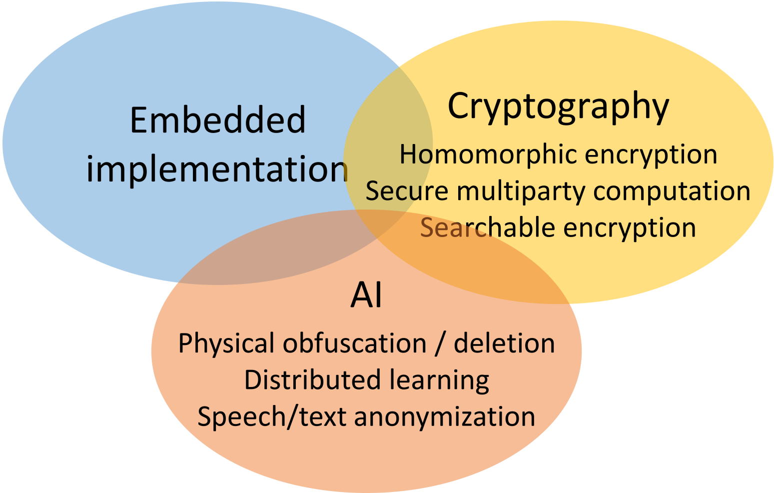
Data often complemented by **metadata**, e.g., user identifier.

What are the risks?

Additional risks w.r.t. text input include

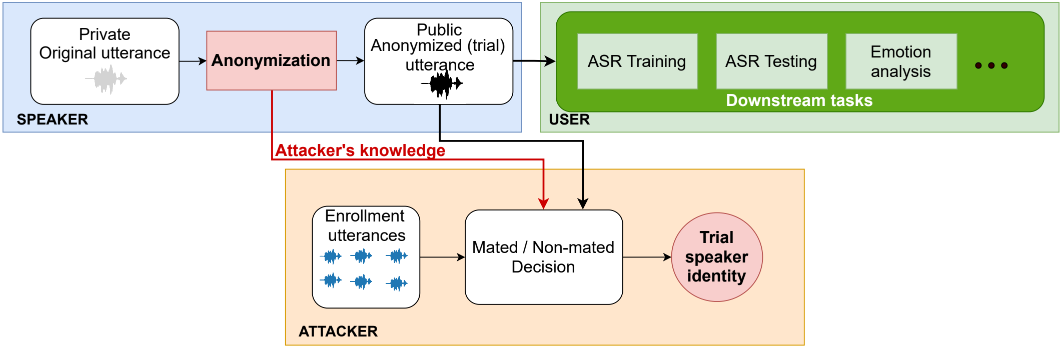
- user profiling
- user identification
- voice cloning (a.k.a. spoofing)

Goal: **protect users while allowing inference and training** with no loss of accuracy.

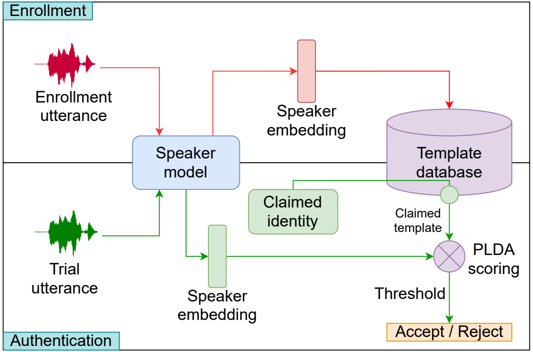


- Speech anonymization:
  - Transform speech to **hide speaker identity**
  - **Leave other information unchanged**, so that it's useful for downstream tasks
- Defines the goal, even when it's not achieved ($\neq$ legal definition)
- Achieving this goal requires:
  - **voice anonymization** via voice transformation/conversion,
  - **verbal content anonymization**,
  - possibly, hiding some identifiable nonverbal attributes.

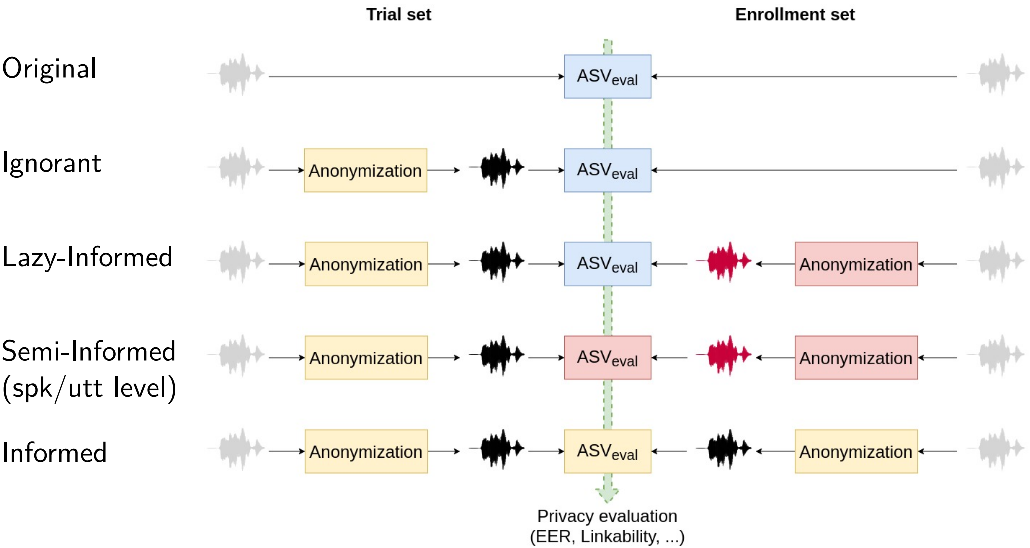
Voice anonymization — Threat model



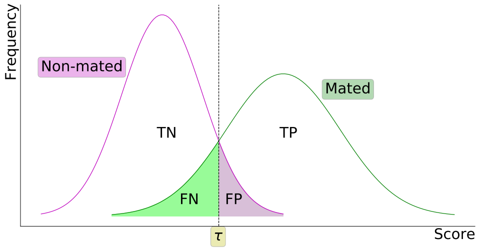
- The success or failure of voice anonymization can be evaluated via **speaker verification**.
- Higher score ⇒ greater chance of being from the same speaker



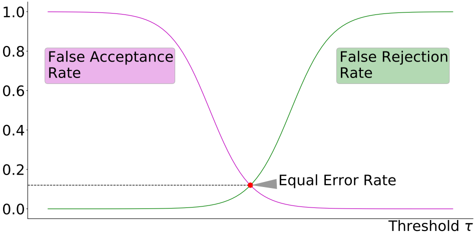
Voice anonymization — Attacker's knowledge



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.

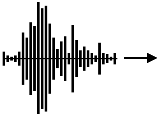
Simple transformations such as **pitch shifting** (often used on TV/radio) do not work!

Original   -3 tone shift   Multiple shifts

EER (Librispeech)

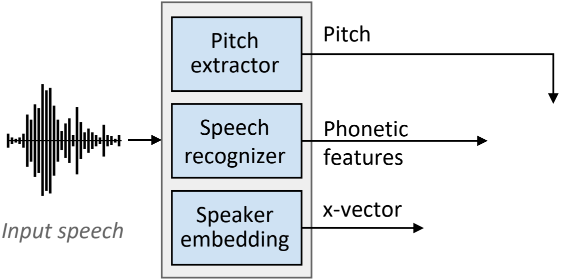
Attacker	VoiceMask
Original speech	4.3%
Ignorant	28.7%
Semi-Informed (utt-level)	5.0%

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

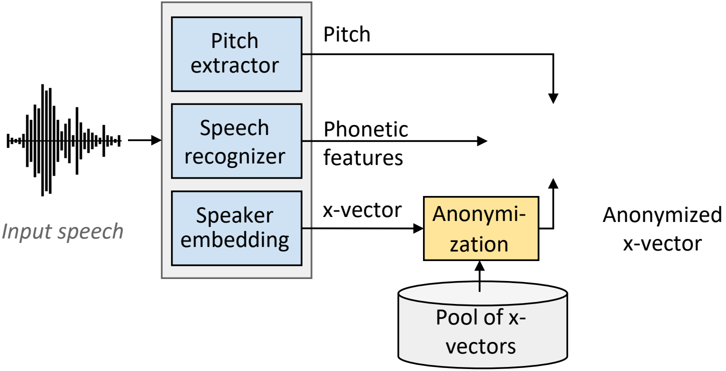


Input speech

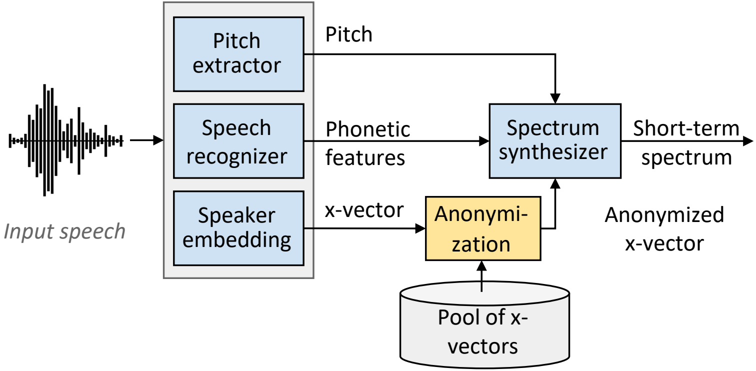
- Idea: **replace user's voice** by that of a target speaker
- Baseline-1 of the VoicePrivacy 2020 Challenge



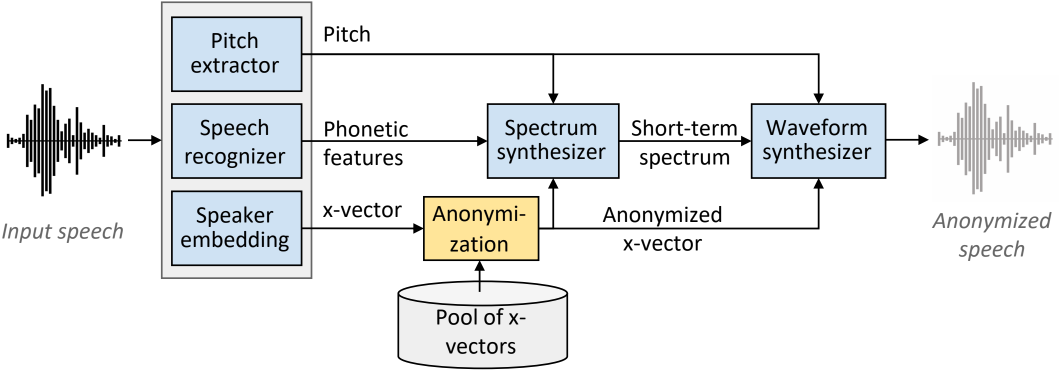
- Idea: **replace user's voice** by that of a target speaker
- Baseline-1 of the VoicePrivacy 2020 Challenge



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



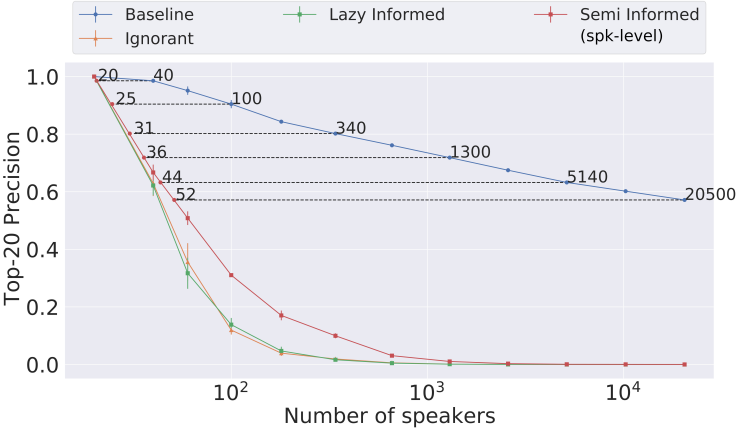
- Idea: **replace user's voice** by that of a target speaker
- Baseline-1 of the VoicePrivacy 2020 Challenge



Original

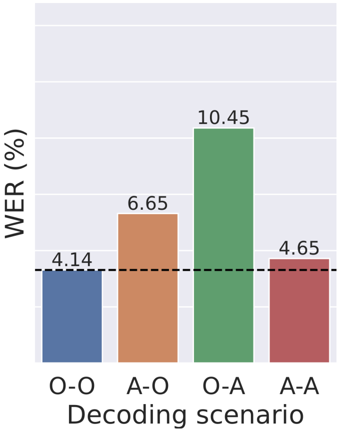
Modified

Top-20 PLDA-based identification accuracy (CommonVoice)

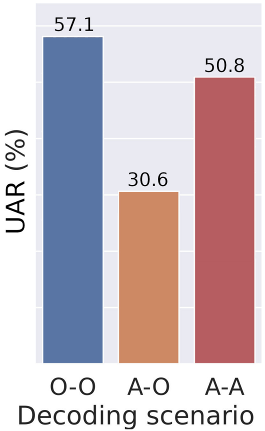


Re-identification risk → 0 with **2,000+ speakers** with best (Semi-Informed) attack.

Speech recognition (LibriSpeech)

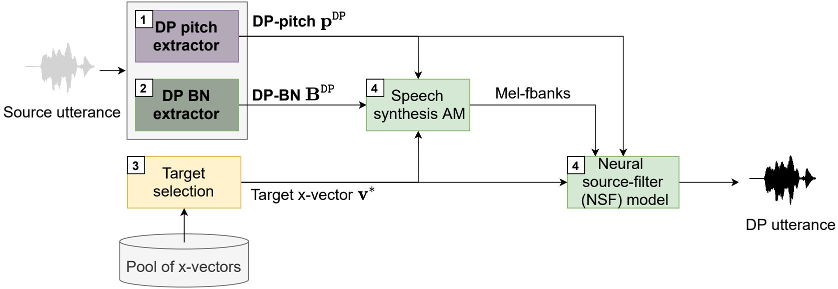


Emotion recognition (IEMOCAP)



Small or negligible loss of utility after **retraining on anonymized data** (A-A).

- 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:
  - using a representation trained on more data, e.g., wav2vec2.0 (works but privacy?)
  - adversarial representation learning (fails)
  - slicing into shorter signals (works but makes human annotation harder)
  - adding noise



Local differential privacy (DP) principle:

- add **random noise** to pitch and phonetic features with scale $\propto 1/\epsilon$
- if $\epsilon \ll 1$, formal privacy guarantees against any attack
- popular for tabular data (e.g., Apple uses $2 \leq \epsilon \leq 8$)

Semi-Informed (utt-level) EER and WER (Librispeech)

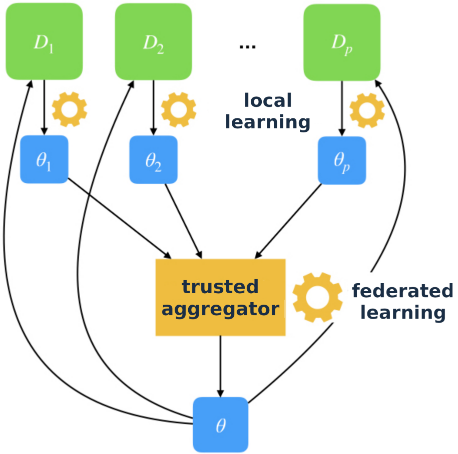
Phonetic ϵ (frame)	**Pitch ϵ (utterance)**	**EER**	**WER**
∞	∞	14.6%	5.4%
100	100	24.2%	6.0%
10	10	27.7%	7.0%
1	1	30.0%	7.8%

Adding noise to the features improves privacy.

Gap between empirical and formal privacy guarantees.

Federated learning

- Presented as an alternative solution for training large-scale generic models which do not require human annotation. . .
- . . . but recent studies reveal that model updates do reveal speaker information.



Perspectives

- **Anonymization**:
  - Reduce residual speaker information
  - Verbal content anonymization
  - Useful formal guarantees?
  - Watermarking to avoid avoid anonymized voice sounding like another real speaker
- **Federated learning**
  - Solutions needed (anonymization?)
- **Evaluation**
  - Link with legal criteria (linkability, singling out, inference)
  - Stronger attackers, perhaps more realistic too (metadata, etc.)
  - Explore attacks on (big) models (membership inference, model inversion, etc.)
- **Give control to users**:
  - Privacy and utility w.r.t. other attributes (e.g., age, accent, medical)
  - User-friendly interface
- Efficient **embedded implementation**