Speaker anonymization: current methods, challenges and perspectives

Michele Panariello
Audio, Security and Privacy group, EURECOM, France

SPSC Webinars, 06 June 2024

Outline

1. Intro to the task & VoicePrivacy Challenge 2024
2. Current directions in speaker anonymization
3. …and current challenges
4. Conclusions

Part 1

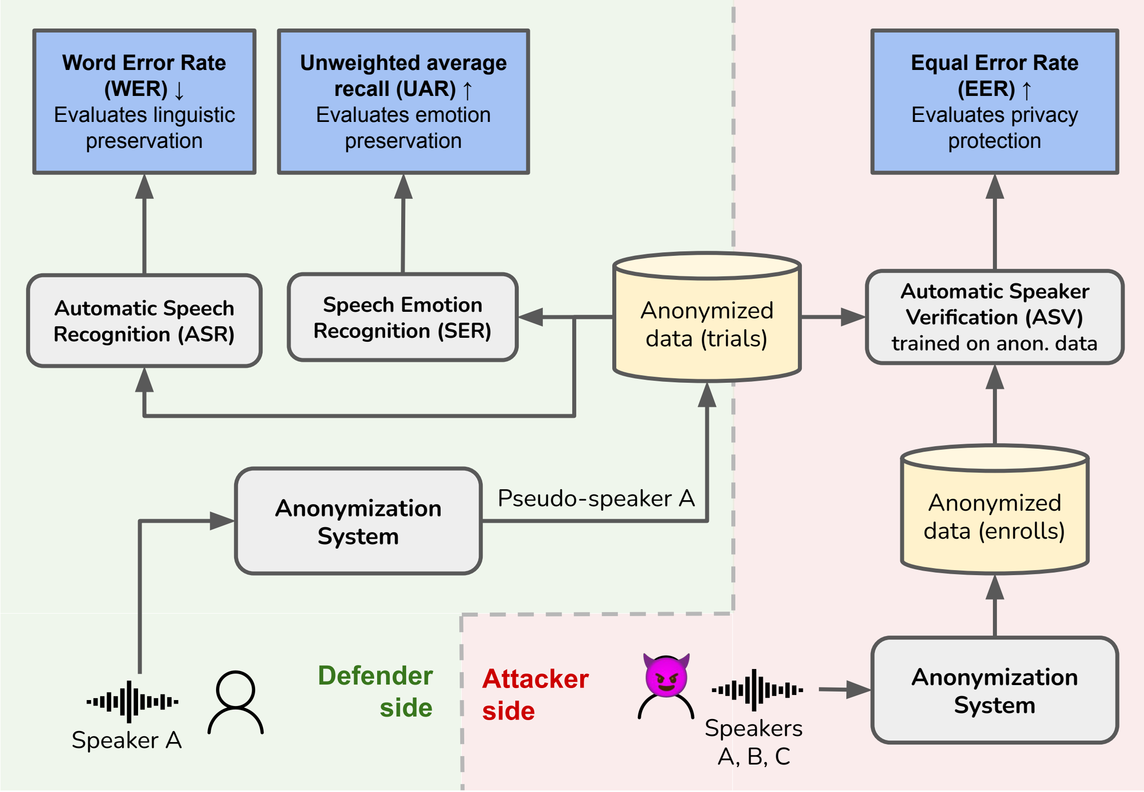
Speaker anonymization

Speaker anonymization in a nutshell

Process a waveform to:

- Conceal speaker identity
- Preserve linguistic content
- Preserve other paralinguistic aspects (e.g. "emotional" content)

Output should also be a waveform.



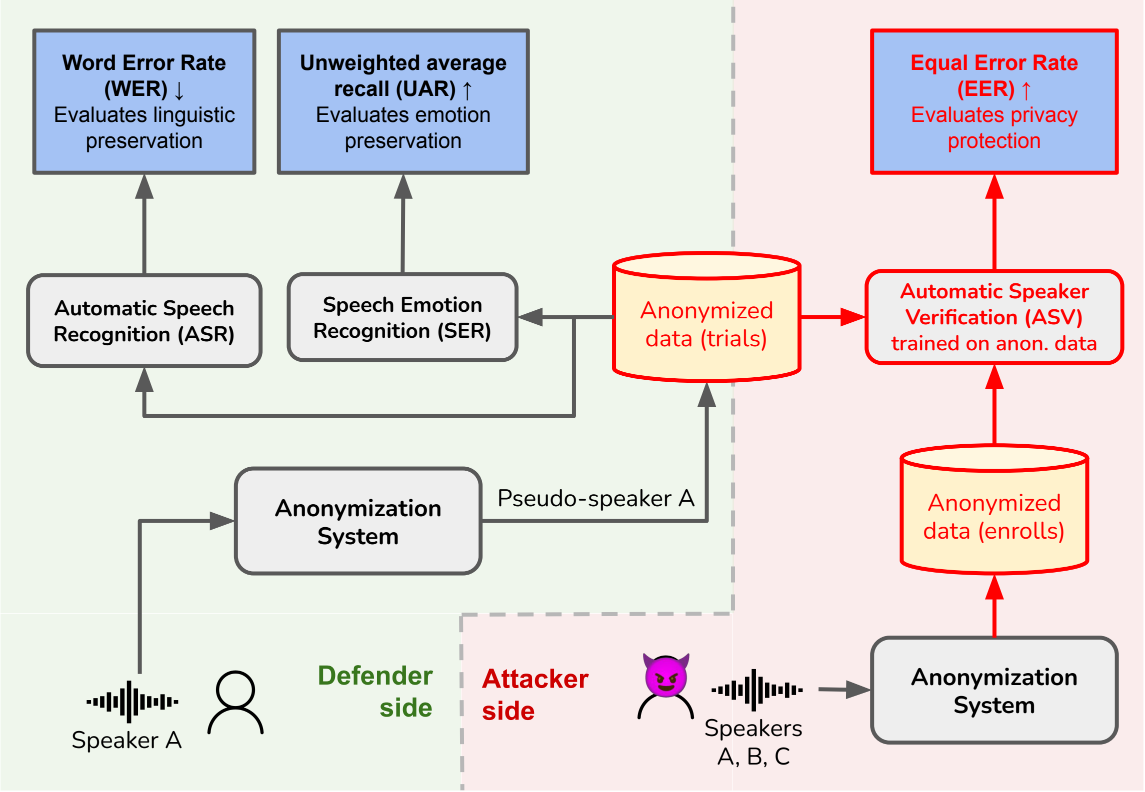
Speaker anonymization in a nutshell

Process a waveform to:

- Conceal speaker identity
- Preserve linguistic content
- Preserve other paralinguistic aspects (e.g. "emotional" content)

Output should also be a waveform.

Note: the attacker is "semi-informed"
(they know the anon. system and use it to re-train the ASV model)

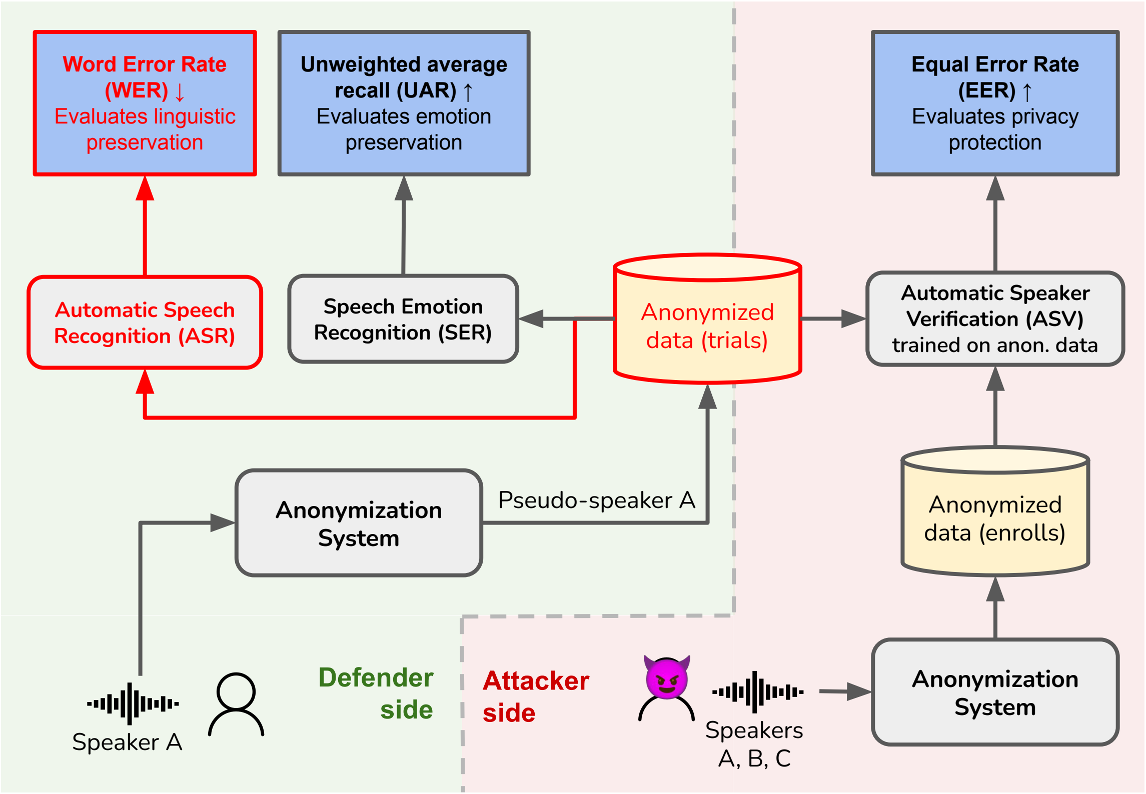


Speaker anonymization in a nutshell

Process a waveform to:

- Conceal speaker identity
- Preserve linguistic content
- Preserve other paralinguistic aspects (e.g. "emotional" content)

Output should also be a waveform.

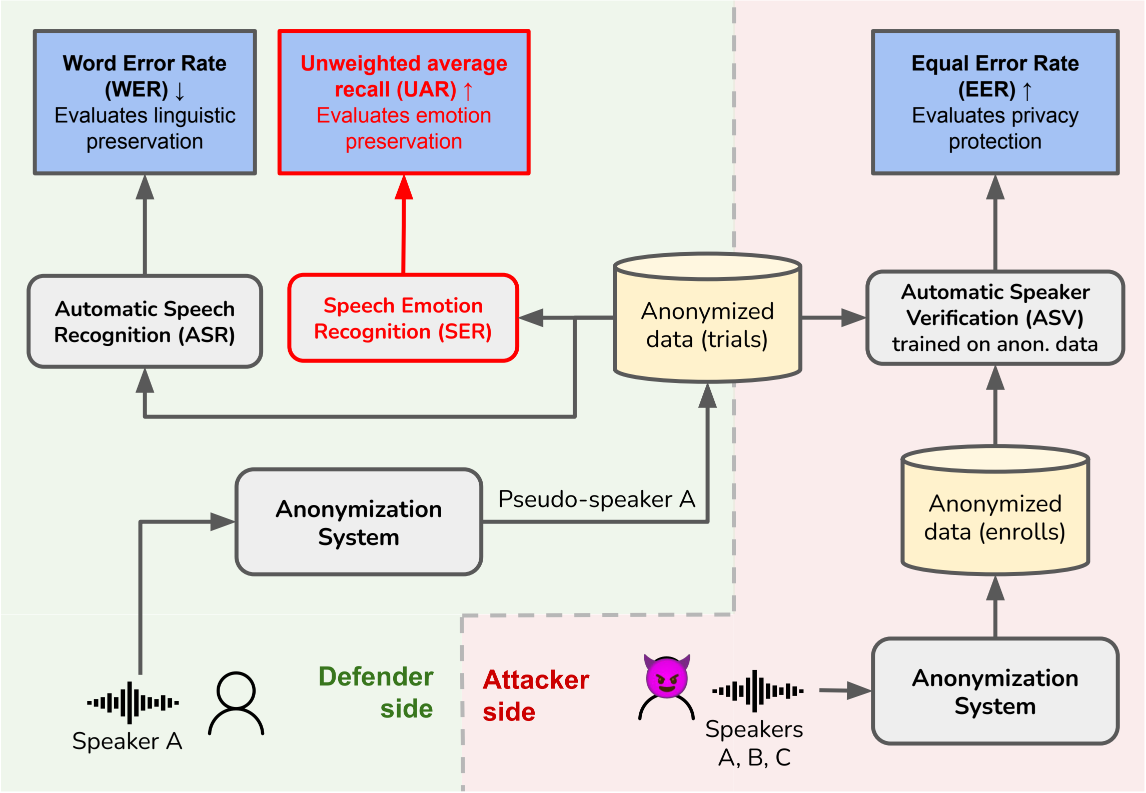


Speaker anonymization in a nutshell

Process a waveform to:

- Conceal speaker identity
- Preserve linguistic content
- Preserve other paralinguistic aspects (e.g. “emotional” content)

Output should also be a waveform.



VoicePrivacy Challenge (VPC) 2024

- Speaker anonymization competition
- Participants invited to design their own speaker anonymization system
- Ranked based on the presented metrics
- Notable changes w.r.t. 2022 edition:
  - Past para-linguistic preservation metrics: pitch correlation and voice distinctiveness
  - every utterance anonymized independently:
    no fixed speaker → pseudo-speaker link (“*utterance-level anon*”)
    - When the link is fixed (like in 2022): “*speaker-level anon*”

Part 2

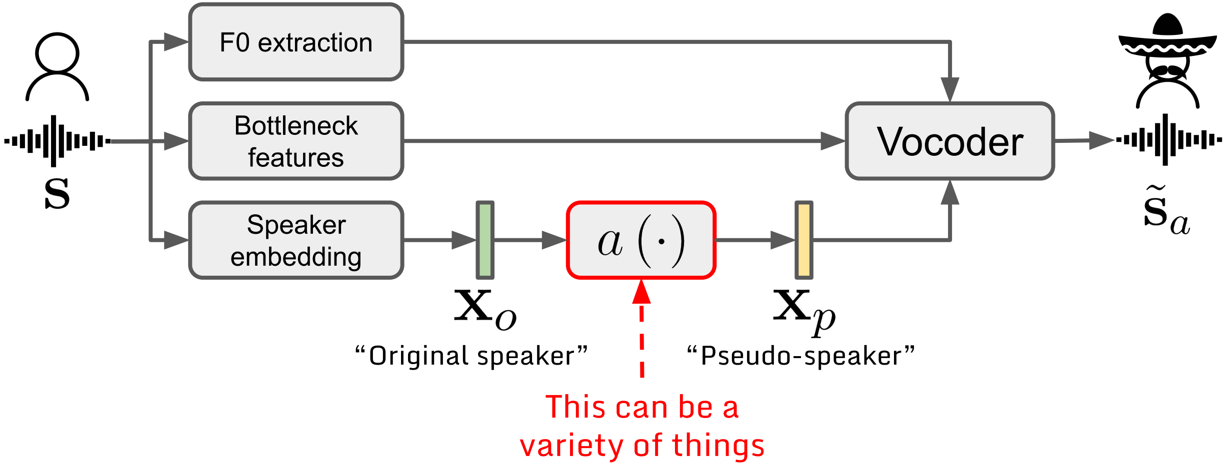
Current directions in speaker anonymization

Current directions

- Voice conversion via x-vector manipulation
- Transcription-based methods (aka. STTTS)
- Methods based on discrete audio units

Voice conversion via x-vector manipulation

- Extraction of
  - F0 curve (voice pitch per time frame)
  - “bottleneck”/“linguistic” features (encode spoken content: embeddings of ASR model)
  - deep speaker embedding vector (a.k.a. *“x-vector”*)
- “Anonymization function” perturbs the x-vector in some way
- Vocoder uses these concatenated features to synthesize a new voice



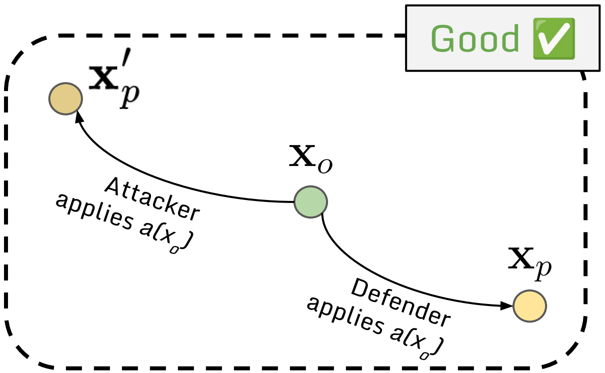
Voice conversion via x-vector manipulation

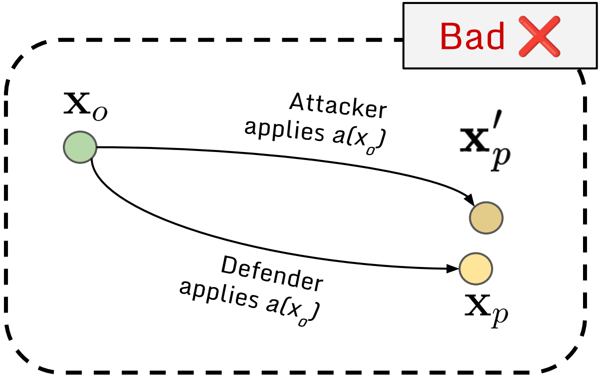
Two recent examples (seen at ICASSP 2024)

- *Language-independent speaker anonymization using orthogonal Householder neural network* (Miao et al.)
  - Learns a parametric function that maxizes distance between X_O and X_P, while preserving the overall distribution of x-vectors
- *Modeling pseudo-speaker uncertainty in voice anonymization* (Chen et. al)
  - Pseudo-speaker embedding is sampled from a gaussian distribution learned for each speaker

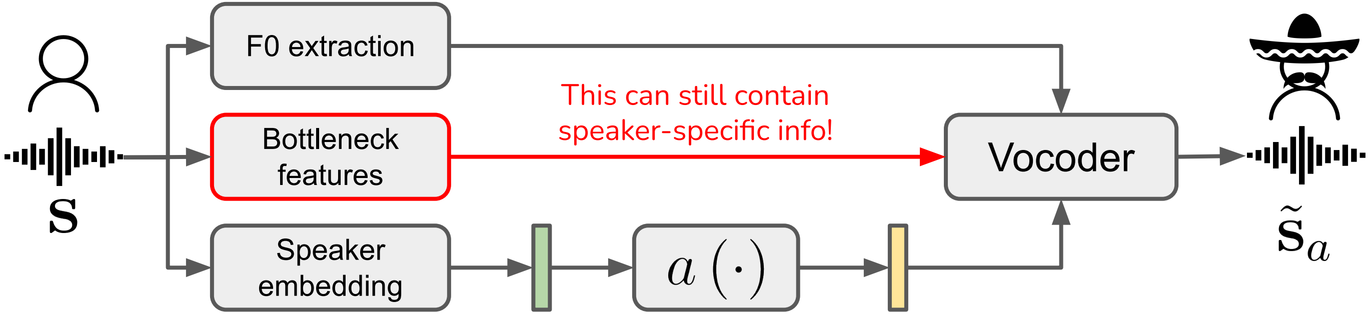
Voice conversion via x-vector manipulation

- "Vanilla" way
- Effective when the attacker is unable to reproduce the anonymization function
  - Makes it more difficult for attacker to train adversarial ASV system, resulting in increased privacy
- Conversely, a very "reproducible" function is bad





Transcription-based methods



- Erase speaker-specific info from bottleneck features by transcribing utterance (to the word or phoneme level)
- Waveform synthesis TTS-style
- “speech-to-text-to-speech” (STTTS)
- “Inject back” some information (e.g. F0 values after some random masking)

Transcription-based methods

Example: VPC baseline B3 from *Prosody Is Not Identity: A Speaker Anonymization Approach Using Prosody Cloning* (Meyer et al., ICASSP 2023)

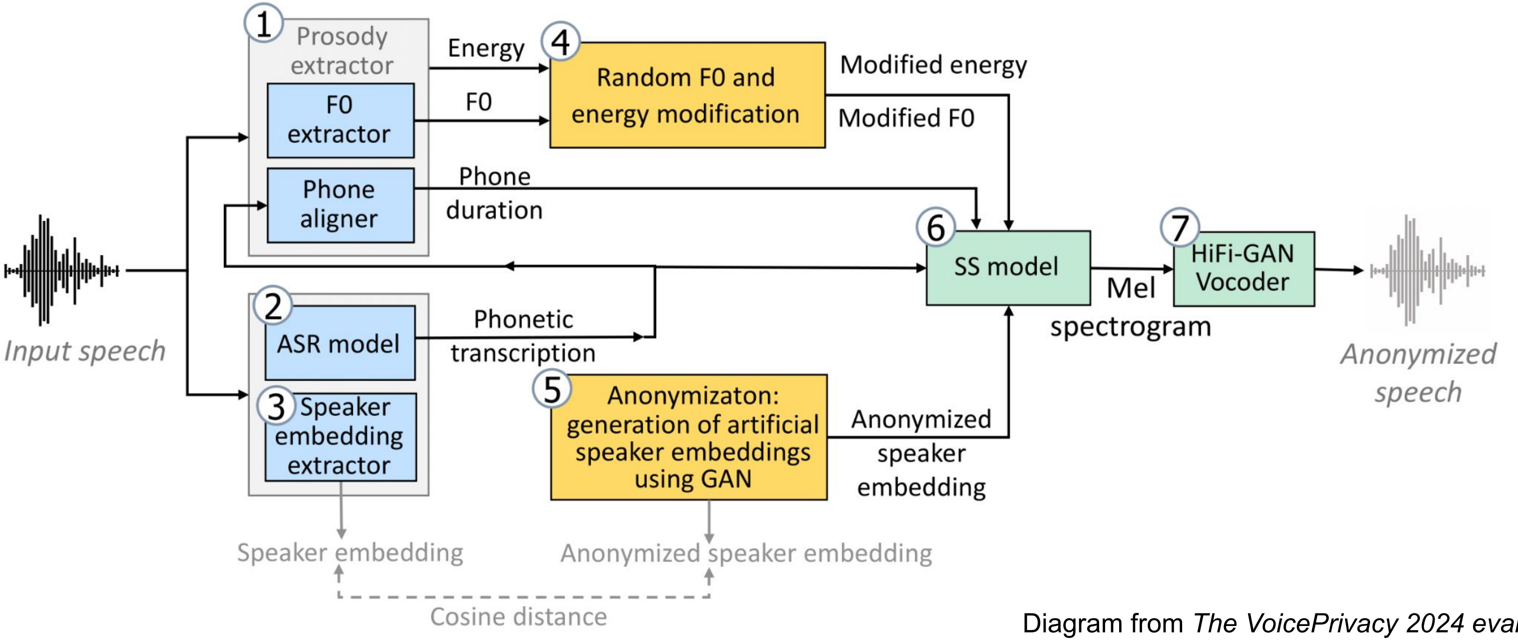
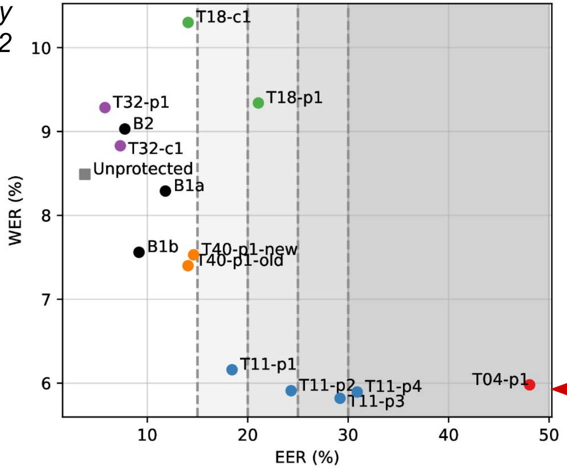


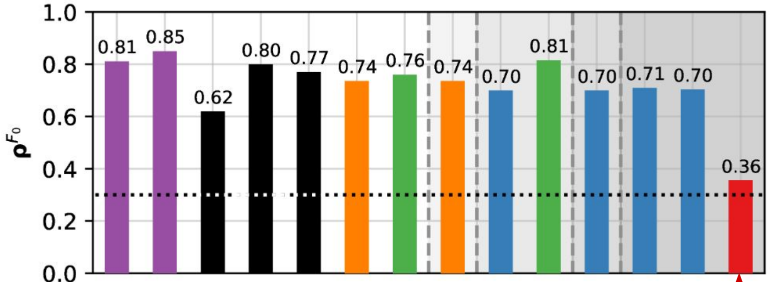
Diagram from *The VoicePrivacy 2024 evaluation plan*

Transcription-based methods

- Strong information bottleneck induced by the transcription: high privacy protection
  - But other desired information could be lost (intonation, emotion)
  - TTS module must be conditioned to preserve that information

Utility VS privacy scores in VPC 2022





Pitch correlation (ρ^{F0}) in VPC 2022

T04: transcription-based

Using discrete audio units

- Attempt to limit the amount of speaker information in linguistic features by quantizing them to discrete units
- Just another “information bottleneck”, not as extreme as STTTS
- Tradeoff between privacy and utility
  - Can depend on codebook size

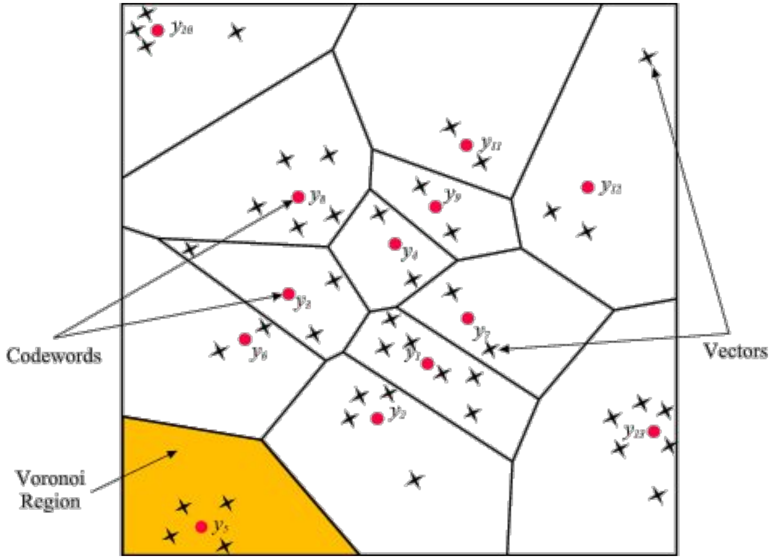


Diagram from *www.mqasem.net*

Using discrete audio units

Example 1: VPC 2024 baseline B5 from *Anonymizing Speech: Evaluating and Designing Speaker Anonymization Techniques* (Champion, PhD dissertation, 2023)

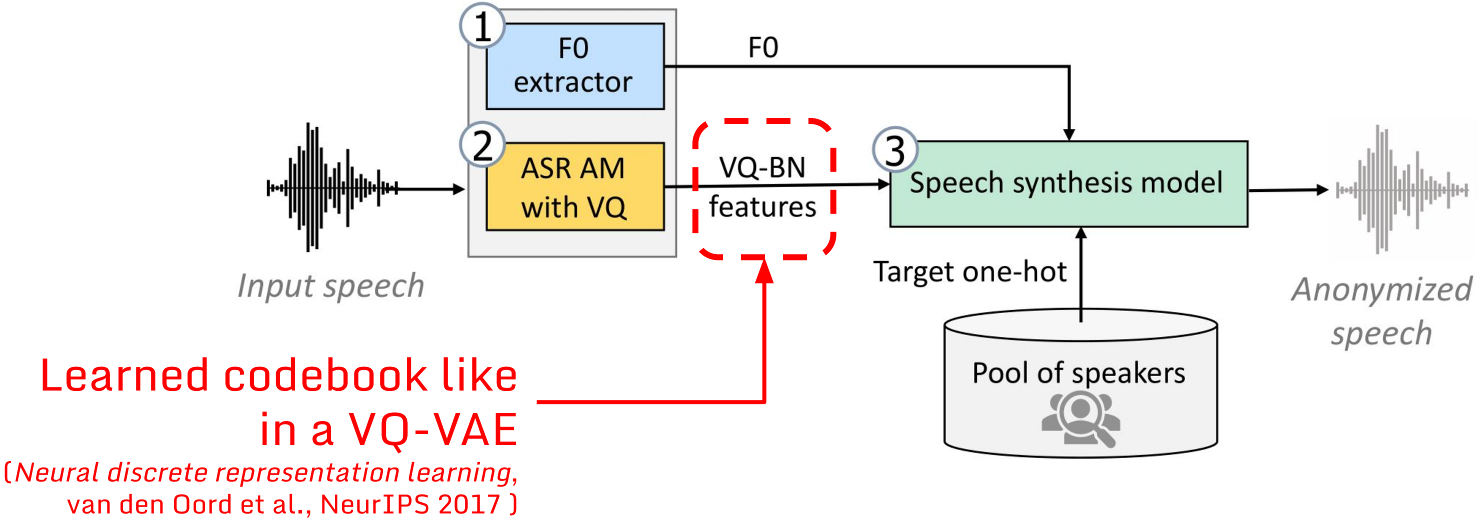
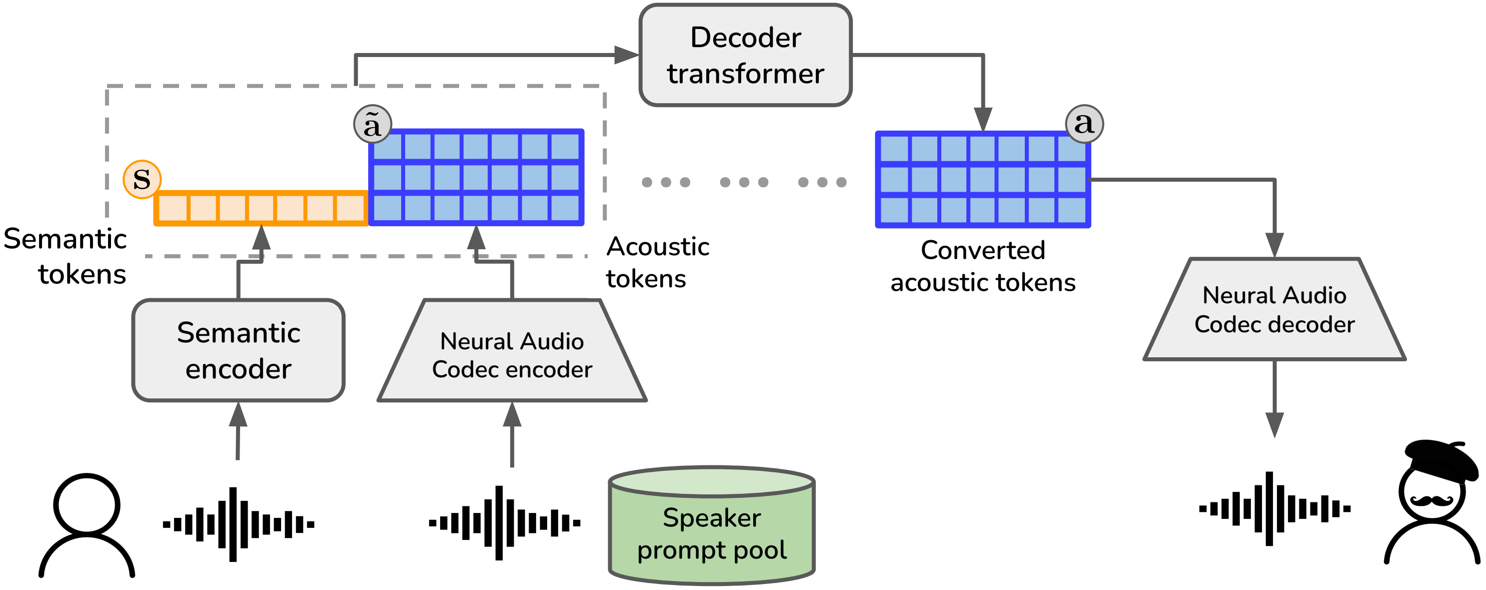


Diagram from *The VoicePrivacy 2024 evaluation plan*

Using discrete audio units

Example 2: VPC 2024 baseline B4 from *Speaker anonymization with neural audio codec language models* (Panariello et al., ICASSP 2024)



Part 3

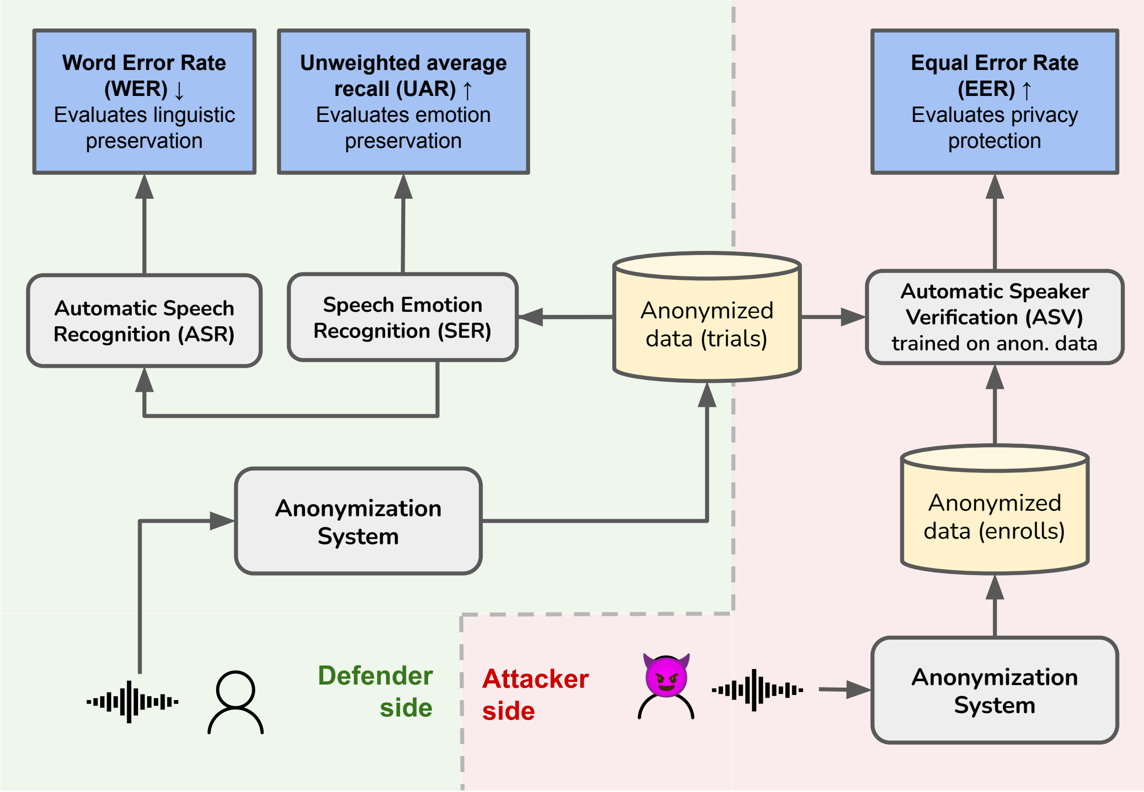
Current challenges in speaker anonymization

Evaluating anonymization

Evaluating spk anon is hard!

From a purely **technical** perspective:

- The task itself involves synthesis
- Several datasets to handle
- Several metrics to compute
- Privacy metric involves re-training a model: bugs/mistakes in doing that can result in overestimated privacy scores

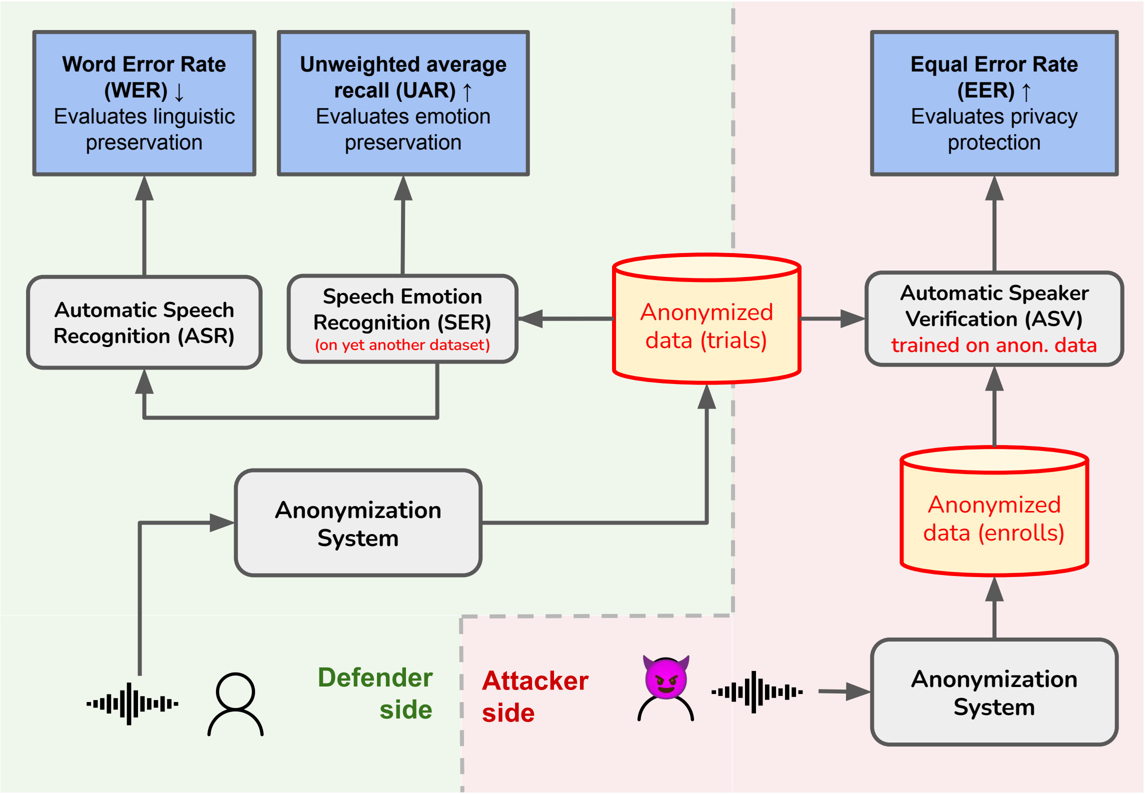


Evaluating anonymization

Evaluating spk anon is hard!

From a purely **technical** perspective:

- The task itself involves synthesis
- Several datasets to handle
- Several metrics to compute
- Privacy metric involves re-training a model: bugs/mistakes in doing that can result in overestimated privacy scores

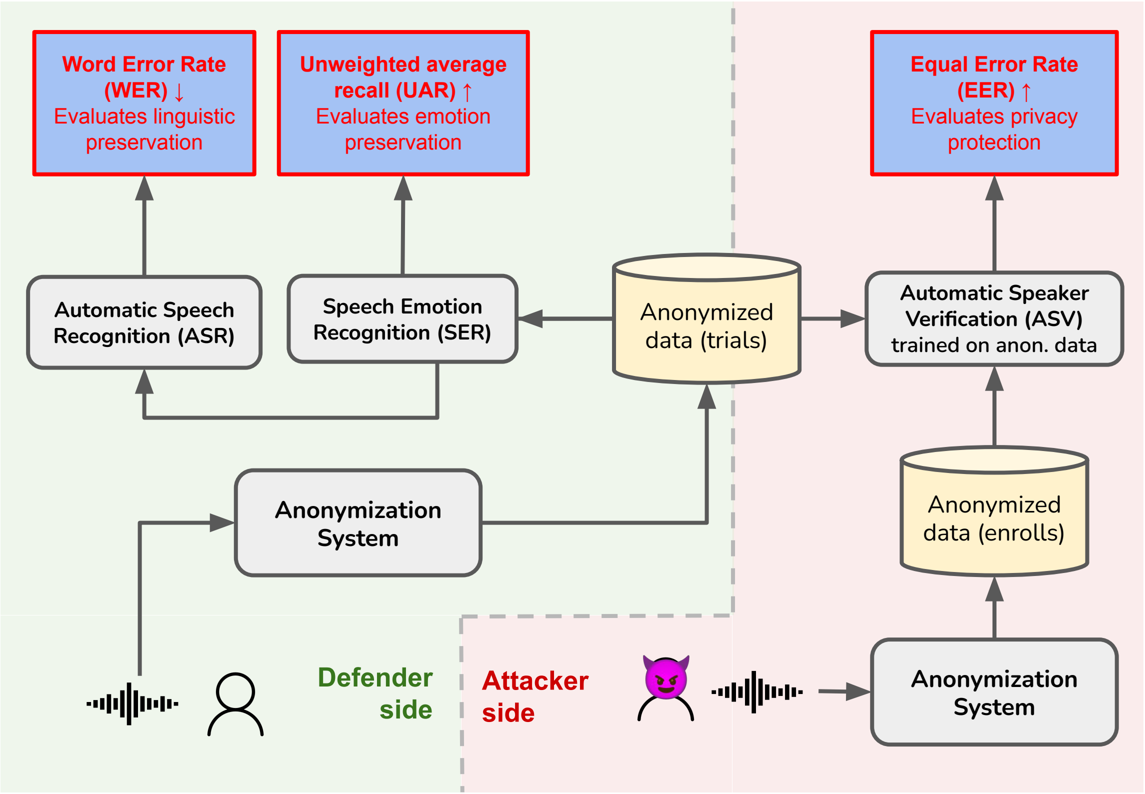


Evaluating anonymization

Evaluating spk anon is hard!

From a purely **technical** perspective:

- The task itself involves synthesis
- Several datasets to handle
- Several metrics to compute
- Privacy metric involves re-training a model: bugs/mistakes in doing that can result in overestimated privacy scores



INSTITUT CARNOT Télécom & Société numérique

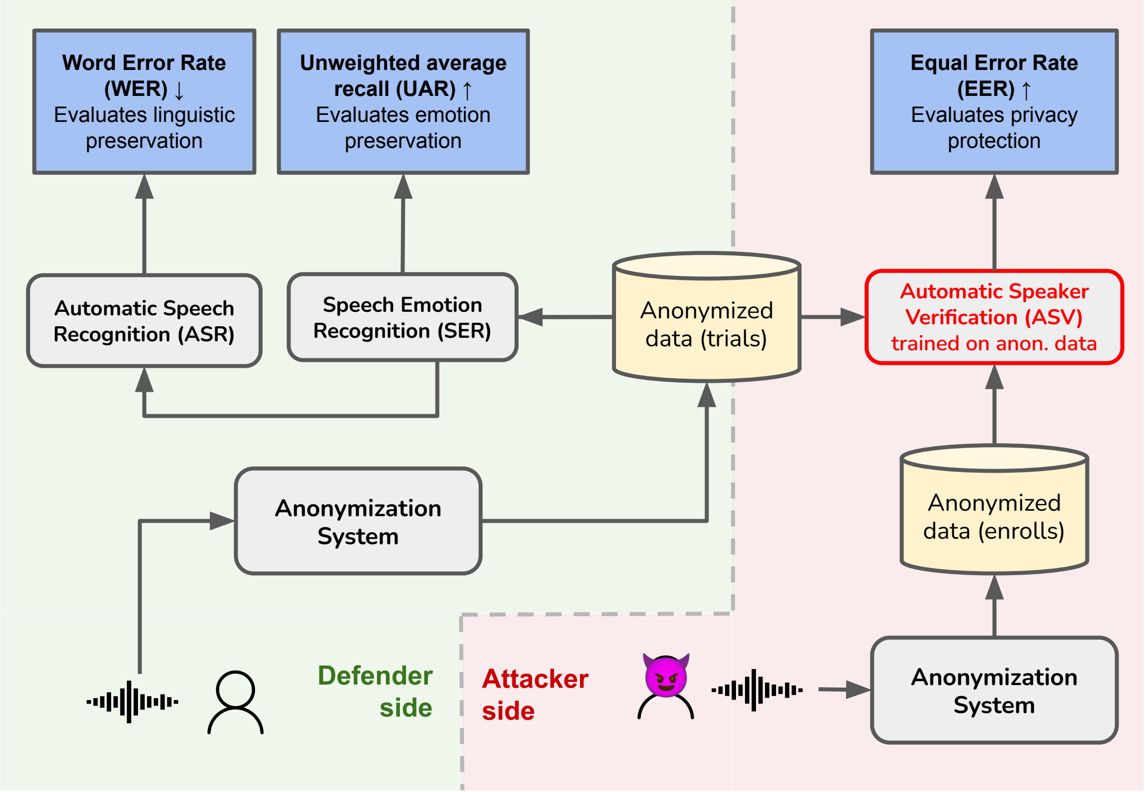
EURECOM Sophia Antipolis

Evaluating anonymization

Evaluating spk anon is hard!

From a purely **technical** perspective:

- The task itself involves synthesis
- Several datasets to handle
- Several metrics to compute
- Privacy metric involves re-training a model: bugs/mistakes in doing that can result in overestimated privacy scores
  (I speak out of experience...)



INSTITUT CARNOT Télécom & Société numérique

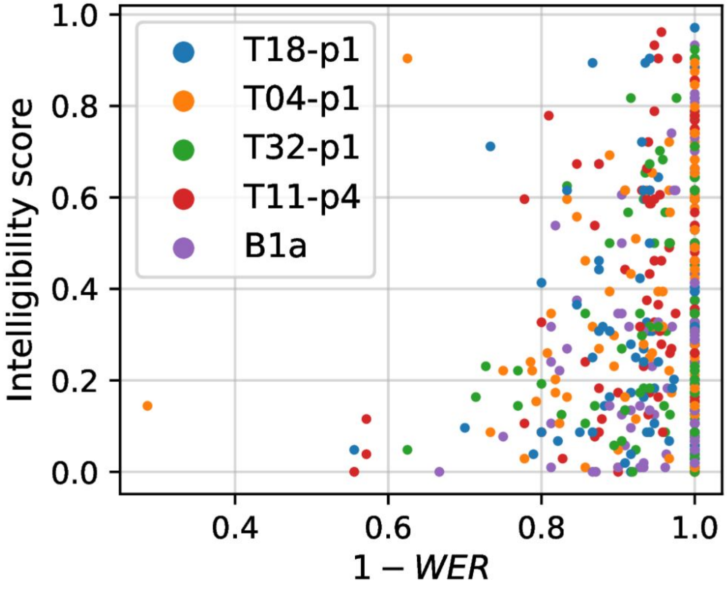
EURECOM Sophia Antipolis

Evaluating anonymization

And from a **conceptual** perspective:

- Do the metrics reflect real use cases?
  - E.g. subjective intelligibility and WER not strongly correlated (Pearson correlation: 0.14)
- Evaluating privacy protection requires impersonating the role of an attacker
  - But we do not know “the optimal attack”
  - …what do we actually know?

Subjective intelligibility rated by human listeners vs 1-WER score assigned by ASR system in VPC 2022

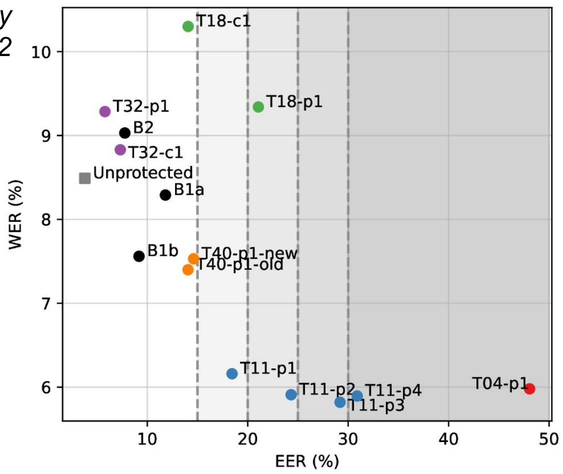


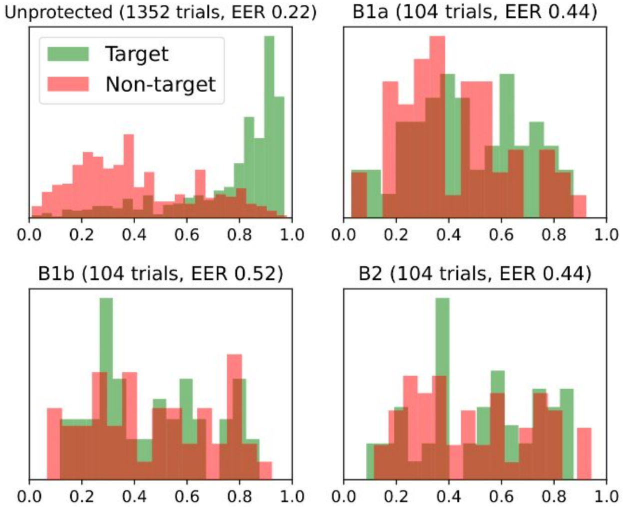
Evaluating anonymization

About the "attacker"

- Even simple algorithms (e.g. DSP-based ones) are effective against "uninformed" humans

Utility VS privacy scores in VPC 2022





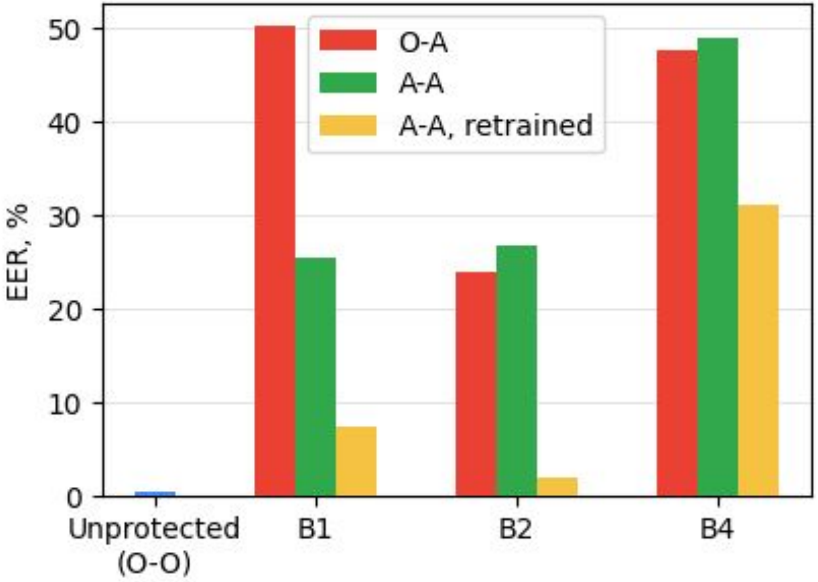
Score distribution of utterance similarity scores assigned by human listeners in VPC 2022

Evaluating anonymization

About the “attacker”

- Even with an ASV system, attacker has to have access to the anonymization system to be a real threat
    - Original enrollment VS anon. trials (O-A) close to 50% EER even for simpler systems
- Task “solved” for practical scenarios?

Privacy score (ASV EER, %) on Libri-dev Male of VPC24 baselines B1, B2, B4 under different attack scenarios



Evaluating anonymization

About the "attacker"

- Adversarial ASV must be retrained, but how?
  - More diversity in the training helps [1]: change *spk* → *pseudo-spk* mapping for every training sample (*utterance-level* anon)
    - But this depends on the anonymization function *a(·)*… different for every system, less comparable results
  - Using same pseudo-spk for all data ("*any-to-one*") would overcome this problem [2]
    - But quite unrealistic 🤷🏻‍♂️

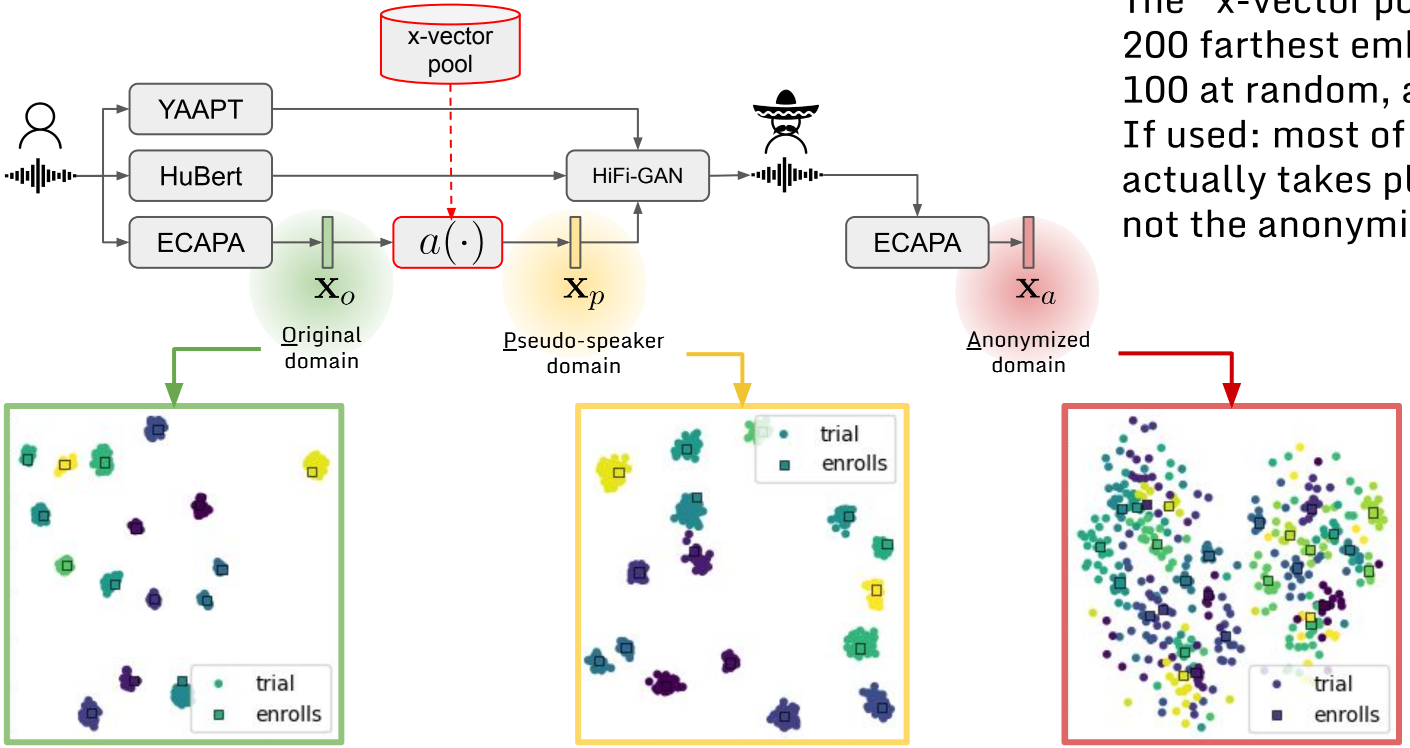
[1] A. S. Shamsabadi et al., "Differentially Private Speaker Anonymization," Proceedings on Privacy Enhancing Technologies, 2023.
[2] P. Champion, "Anonymizing Speech: Evaluating and Designing Speaker Anonymization Techniques." PhD dissertation, 2023.

Evaluating anonymization

… and about the “defender”!

- Speaker anonymization systems are complicated
  - Ablation studies require generating multiple anonymized datasets, can be costly
- How much personal information does each block of the system erase, exactly?

Evaluating anonymization



The "x-vector pool" anon. function: find 200 farthest embeddings from X_O, pick 100 at random, average into X_P.
If used: most of the anonymization actually takes place within the vocoder, not the anonymization module [3]...

[3] M. Panariello et al., "Vocoder drift in x-vector–based speaker anonymization," in INTERSPEECH 2023, ISCA, Aug. 2023, pp. 2863–2867.

Speaker anonymization VS voice conversion

“If we remove anon. module and do any-to-one pseudo-speaker, aren’t we just doing voice conversion?”

- Well… kind of
- A lot of ideas can be taken from the voice conversion community
  - We just have not done it that much… yet
- Overall, the goals differ:

	Objective	Metrics
Voice Conversion	Recording of source speaker should sound like specific target speaker	• “Speaker similarity” • MOS or other subjective metrics • WER/CER
Speaker Anonymization	Recording of source speaker should **NOT** sound like source speaker	• Specifically trained adversarial ASV model • WER • Some utility metric…

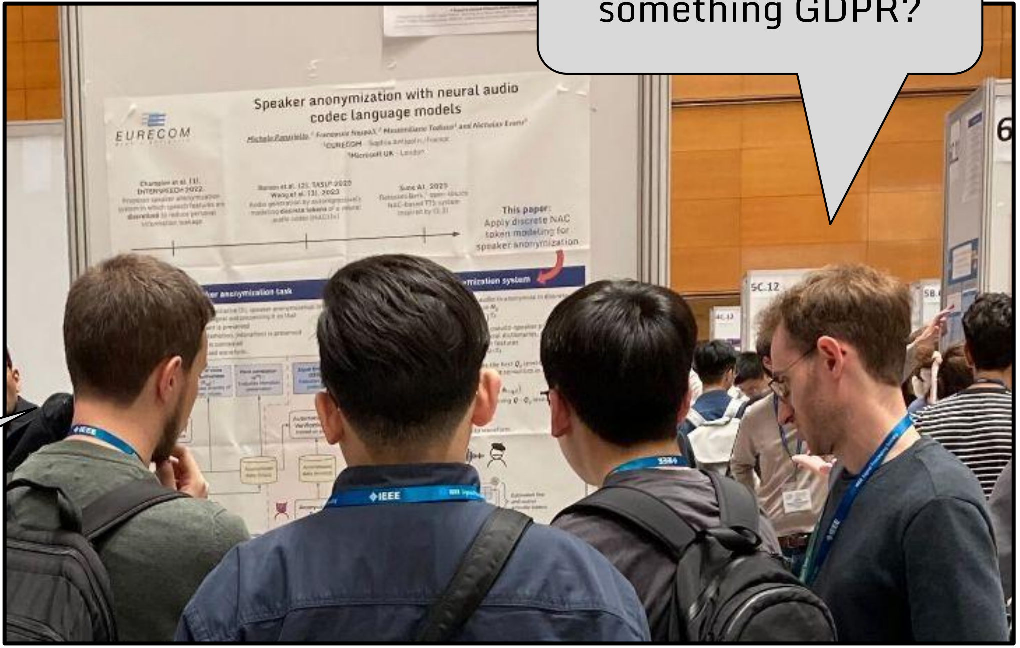
Which utiliy metric? The use case matters

- Aside from WER, the actual utility metric depends on the task
- VPC rules attempt a general “one-size-fits-all” approach to utility:
  - 2022: WER + F0 curve preservation + variety of pseudo-spk voices
    (plus the subjective evaluation)
  - 2024: WER + emotion preservation
- Specific use cases might have different requirements
  - Downstream task fixed → No need to go back to waveform?
  - Anonymization needs to be evident → Better if speech does NOT sound natural?
  - What matters is only the spoken content → ...just transcribe it?
- **VoicePrivacy proposes a general protocol, but it can be adapted!**

How do we find practical use cases though?

- More dialogue with the legal community would be beneficial
  - Find out if, when and how anonymization actually matters from a legal standpoint
  - So that you don't end up like me at ICASSP (or in many other situations):

…something something GDPR?



This anonymization thing sounds cool, but why do we need it?

Part 4

Conclusion

To recap…

- Introduced speaker anonymization
  - Take a speech waveform
  - Mask the speaker identity
  - Preserve the rest
- Presented VoicePrivacy Challenge 2024 *(deadline: 15th of June)*
- Main research directions
  - Voice conversion based on x-vector manipulation
  - Transcription-based (STTTS)
  - Quantized speech units
- Current challenges
  - Both privacy and utility difficult to evaluate
  - Deal with an intrinsically “vague” task

Thank you!