Legally validated evaluation framework for voice anonymization

Nathalie Vauquier[1], *Brij Mohan Lal Srivastava*[1], *Seyed Ahmad Hosseini*[1], *Emmanuel Vincent*[1,2]

[1]Nijta SAS, France
[2]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

brij@nijta.com

Abstract

Classical speaker verification metrics used to evaluate voice anonymization systems, such as the equal error rate (EER), fail to properly quantify the residual re-identification risk. This paper introduces a new evaluation framework based on two metrics, Linkability and Singling Out, derived from the legal definitions in the Article 29 Working Party's Opinion 05/2014 on Anonymization Techniques endorsed by the European Data Protection Board (EDPB). Our framework translates these legal concepts into quantitative metrics for speech data. The proposed framework has been legally validated by the French Data Protection Authority. Experiments across various attack scenarios reveal that, while the EER remains stable, Linkability and Singling Out vary much more. This demonstrates that the residual privacy risk after anonymization is far more variable than indicated by the EER, underscoring the need for evaluation metrics that align with legal criteria.

Index Terms: voice anonymization, legal compliance, evaluation, privacy.

1. Introduction

Large amounts of speech data are being collected and analyzed to build modern healthcare systems [1, 2]. These voice-based systems are capable of handling emergency calls 24/7, scheduling visits, following up with patients, making early diagnoses, redirecting calls to physicians, consulting insurers, and interfacing with pharmacies. The analysis of speech data is essential to improve the quality of these systems. However, alongside these benefits, come obvious risks to data privacy [3, 4]. Due to these risks, many innovative projects face strict regulations that may delay or even block their implementation. Voice anonymization [5–7] aims to provide a fair trade-off by removing sensitive patient-related information from speech data while retaining the information required for healthcare-related analyses. In the past five years, many systems have been proposed towards this goal, with application to healthcare or other sectors [8–15].

Recital 26 of the EU General Data Protection Regulation (GDPR) [16] defines anonymous information as "information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable". The Article 29 Working Party's Opinion 05/2014 on Anonymization Techniques [17] offers a legally valid interpretation of this definition endorsed by the European Data Protection Board (EDPB). It stipulates that data can be considered anonymous if it is not possible to 1) single out an individual record (*singling out* attack), 2) link records that pertain to the same data subject (*linkability* attack), or 3) infer additional information about the data subject by combining the anonymized data with other available information (*inference* attack). This offers a pathway for innovators in healthcare and other sectors as data can be claimed anonymous (thereby ensuring legal protection) when re-identification is practically infeasible according to these three criteria. A robust anonymization process, therefore, aims to minimize the re-identification risk below a threshold depending on the use case, and to assess and control any residual risk over time.

Despite their legal importance, these criteria have not been translated into quantitative metrics by the speech processing community so far. Instead, the evaluation of voice anonymization has relied on the classical evaluation framework for speaker verification and associated metrics such as the equal error rate (EER) [18–22]. In this paper, we bridge the gap between technical evaluation and legal requirements by introducing a new evaluation framework based on two quantitative metrics: **Singling Out** and **Linkability**. Our Singling Out metric computes the probability that an attacker defined by a speaker embedding and an optimal cosine similarity threshold can *isolate* a single speaker from a set of anonymized speech samples based on the notion of predicate singling out (PSO) [23]. Concurrently, our Linkability metric measures the probability that anonymized speech samples can be correctly *linked* to the corresponding enrollment speaker by analyzing cosine similarity scores between speaker embeddings.[1] To ensure that these metrics reliably capture the re-identification risk, we develop an evaluation framework consisting of random sampling, threshold calibration, and statistical averaging over multiple runs and folds.[2] The proposed framework and metrics have been formally validated by *Commission Nationale de l'Informatique et des Libertés* (CNIL), the French Data Protection Authority. Experiments on the Mozilla Common Voice dataset reveal that these metrics capture significant variations in the re-identification risk under different attack scenarios — variations that remain hidden when relying solely on the EER.

Section 2 introduces the two legally validated metrics. Section 3 details the example anonymization system under study, the various attack models, the data and the evaluation framework. Section 4 illustrates the results of the three metrics under these attack models. Finally, Section 5 concludes the paper with a summary of our contributions and future work.

2. Legally validated evaluation metrics

We introduce two complementary metrics, Singling Out and Linkability, to quantify the re-identification risk of anonymized speech data. Together, they ensure that both the isolation and linkage aspects of the re-identification risk are addressed, as per the above legal criteria. Singling Out assesses whether an indi-

[1] This metric differs from the linkability metric defined in [24, 25].

[2] https://github.com/brijmohan/voice-anonymization-legal-eval

vidual speaker can be isolated from the test dataset using residual speaker information. This metric is crucial because, even if the identity of that individual remains unknown, isolation makes him/her vulnerable. Linkability, by contrast, assesses the ability to match an utterance from the test dataset with another utterance from the same speaker among all speakers in a disjoint enrollment dataset. While no single utterance may suffice to fully reveal the speaker's identity, higher linkability increases the risk that an attacker aggregates multiple utterances to reconstruct a high-fidelity profile of the speaker.

To implement these metrics, we classically [18–20] assume that the linguistic content does not involve any direct or indirect identifiers, so the re-identification risk arises from the paralinguistic and extralinguistic attributes of anonymized speech only. Consequently, we also disregard the inference criterion, which refers to deducing sensitive attributes from anonymized data. This is because, contrary to tabular data with explicit attributes (e.g., age, gender, or ethnic origin), speech analysis methods are limited in the number and the accuracy of the attributes they can extract. Building an inference attack would require the attacker to possess a table containing the value and the uncertainty over these attributes for all enrollment speakers and to know the extent to which combinations of given attributes identify individuals, which appears highly unrealistic today. Metrics implementing the three legal criteria based on linguistic content are therefore left for future study.

2.1. Singling Out

The PSO framework [23] quantifies the risk that an attacker isolates an individual from an anonymized dataset with N individuals. To do so, it considers subsets $X = \{x_1^{\text{test}}, \dots, x_N^{\text{test}}\}$ containing a single entry x_i^{test} per individual. The attacker constructs binary predicates $p(\cdot)$ that aim to isolate one entry, i.e., $p(x_i^{\text{test}}) = 1$ for exactly one i and $p(x_j^{\text{test}}) = 0$ for all $j \neq i$. An anonymization system is considered *PSO-secure* if the probability of isolation across all subsets and predicates does not significantly exceed the baseline probability $\exp(-1) \approx 37\%$ achievable by a random predicate with expectation $\mathbb{E}_x\{p(x)\} = 1/N$.

In the speech context, the dataset entries are speaker embeddings (e.g., x-vectors) averaged over one or more anonymized utterances and speaker similarity is assessed via a similarity function $s(\cdot, \cdot)$ (e.g., cosine similarity). We assume that an enrollment dataset containing a subset $S \subset \{1, \dots, N\}$ of the test speakers is available and we define a predicate as

$$p(x^{\text{test}}) = \mathbb{1}\{s(x^{\text{test}}, x^{\text{enroll}}) > s^{\text{thresh}}\} \tag{1}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function, x^{enroll} is an enrollment speaker embedding and s^{thresh} is the optimal threshold ensuring that the predicate has an average value $\mathbb{E}_{x^{\text{calib}}}\{p(x^{\text{calib}})\} = 1/N$ over a dataset of $M \times N$ calibration embeddings x^{calib} with M embeddings for each of the N test speakers. The utterances in the test, enrollment, and calibration datasets are disjoint. Isolation succeeds if $p(x_i^{\text{test}}) = 1$ for exactly one entry x_i^{test}, *whether it is from the same speaker as* x^{enroll} *or not*. The Singling Out metric is defined as the probability of isolation over all subsets X and enrollment speakers:

$$\pi^{\text{sing}} = \Pr_{X, x^{\text{enroll}}}\{\exists i \text{ s.t. } p(x_i^{\text{test}}) = 1 \text{ and } p(x_j^{\text{test}}) = 0 \, \forall j \neq i\}. \tag{2}$$

2.2. Linkability

Linkability measures the risk that an attacker matches a speaker embedding x_i^{test} computed from one or more anonymized test utterances with the enrollment embedding x_i^{enroll} corresponding to the same speaker i among a set of N' enrollment speakers. Linkage succeeds if

$$s(x_i^{\text{test}}, x_i^{\text{enroll}}) > \max_{j \neq i} s(x_i^{\text{test}}, x_j^{\text{enroll}}). \tag{3}$$

The Linkability metric is defined as the probability of linkage over all test data:

$$\pi^{\text{link}} = \Pr_{x_i^{\text{test}}}\{s(x_i^{\text{test}}, x_i^{\text{enroll}}) > \max_{j \neq i} s(x_i^{\text{test}}, x_j^{\text{enroll}})\}. \tag{4}$$

A low π^{link} indicates that it is difficult for an attacker to correctly link anonymized recordings to the correct speaker, thus supporting the claim of effective anonymization.

3. Experiments

3.1. Anonymization system

In the following, we report the values taken by the Singling Out and Linkability metrics on anonymized data obtained from one anonymization system. We chose the B1 anonymization baseline of the Voice Privacy 2024 Challenge [20] as an example. It anonymizes speech by replacing the original speaker's x-vector with an anonymized x-vector computed from the average of a randomly selected subset of candidate x-vectors, followed by speech synthesis using a HiFi-GAN neural vocoder.

3.2. Attackers

We derive our attacker models from the framework presented in [26] and adopted in all Voice Privacy Challenges, which defines adversaries with varying levels of knowledge about the anonymization process. In this framework, an attacker's success in re-identifying speakers depends critically on their awareness of the underlying anonymization mechanism. Accordingly, we define three attacker profiles:[3]

- *Ignorant* attacker: Unaware that the data are anonymized, this attacker trains their speaker embedding model on original, untreated data.
- *Semi-Informed* attacker: This attacker is aware that the data have been anonymized but does not possess detailed information about the specific technique used. Consequently, they train their speaker embedding model on data processed by a similar, though not identical (e.g., slightly outdated), anonymization system. In our experiments, we use Baseline B1.a from the Voice Privacy 2022 Challenge [19], which is identical to the 2024 B1 baseline, except that it employs an earlier, two-step speech synthesis method consisting of an acoustic model and a neural waveform model.
- *Informed* attacker: This attacker has full knowledge of the anonymization process and access to the same anonymization system used to process the test data. They train their speaker embedding model on data anonymized with that system, representing a worst-case scenario.

We also report results in the *Original* setting where an attacker possesses the original test data before anonymization. By measuring the re-identification risk against these different attackers, we evaluate whether anonymization keeps it below an acceptable threshold for the most realistic attackers, which depend on the use case scenario.

[3] In the Voice Privacy Challenges, our Informed attacker is called Semi-Informed and our Semi-Informed attacker has no equivalent.

Table 1: *Dataset statistics: number of speakers (female, male, total), total duration (hours), and number of utterances.*

Dataset	**Female**	**Male**	**Total**	**Duration (h)**	**No. of Utterances**
LibriSpeech *train-clean-360*	439	482	921	364	104,014
Common Voice 11.0 A	5,601	16,423	22,024	323	234,945
Common Voice 11.0 B	1,354	3,595	4,949	1,409	996,971

3.3. Data

Our experiments rely on the Librispeech [27] and Common Voice 11.0 [28] datasets. Table 1 summarizes the number of speakers, the total duration, and the number of utterances for each dataset. The Librispeech *train-clean-360* subset is used to train the speaker recognition model, while two subsets from Common Voice 11.0 form the basis for our re-identification experiments. Subset A includes speakers with at least 2 min of speech, and subset B includes speakers with at least 3 min of speech; the two subsets are constructed so that each speaker in B appears in A with disjoint utterances. To compute the Singling Out metric B is used as the enrollment set and A as the test set, while for the Linkability and classical EER metric A is used as the enrollment set and B as the test set. Using Common Voice makes it possible to consider a much larger number of test (N) or enrollment (N') speakers than in the Voice Privacy Challenges, which better matches real use case scenarios.

3.4. Evaluation framework

To assess the re-identification risk according to the proposed legally validated criteria, we first anonymize all utterances and extract an x-vector speaker embedding from each anonymized utterance using the Sidekit toolkit [29]. The x-vector extractor is based on an ECAPA-TDNN architecture [30] and is trained on the anonymized LibriSpeech *train-clean-360* dataset.

A key component of our evaluation framework which also differs from the Voice Privacy Challenges is to consider the conversation length (L), that is the number of utterances used to compute the x-vector for each speaker. For $L = 1$, a single utterance is used. For longer conversation lengths $L \in \{3, 30\}$, the x-vectors from L utterances are averaged to obtain a more robust speaker embedding. This allows us to quantify how the amount of speech data per speaker influences the measured re-identification risk under the three metrics.

For the Singling Out metric, experiments span a range of test speaker counts N varying from 20 to 22,024. We randomly select 495 speakers from set B as enrollment speakers (S) and compute the corresponding averaged x-vectors x^{enroll} from 30 utterances per speaker. For each enrollment speaker, we then randomly select $10 \times L$ utterances from each of N test speakers in set A, including the enrollment speaker, which we split into $9 \times L \times N$ utterances for calibration (resulting in $M = 9$ x-vectors per speaker) and $L \times N$ utterances for test (resulting in 1 x-vector per speaker). The optimal threshold s^{thresh} for a given predicate is equal to the average of the 9th and 10th similarity scores between the enrollment x-vector and the x-vectors in the calibration set sorted by decreasing value.[4] The splitting of the $10 \times L$ utterances of each test speaker into calibration and test subsets is repeated 10 times in a cross-validation fashion, and the process is repeated over 5 random draws of the test speakers to obtain robust isolation probabilities.

Similarly, the Linkability metric spans a range of enrollment speaker counts N' from 20 to 22,024. We randomly select L utterances from each of the speakers in set B as a test set and compute the corresponding 4,949 averaged x-vectors. For each test speaker, we randomly select N' enrollment speakers from set A, including the test speaker, and compute the averaged x-vectors over all utterances of each speaker. This process is repeated over 5 random draws. We accelerate this process by pre-computing a score matrix of size $22,024 \times 4,949$ and then comparing the scores for several runs in parallel.

Finally, we compute the ROCCH-EER [31] (simply noted EER hereafter) using the same conversation lengths and speaker counts as above, and plot the quantity $1-$EER for simpler comparison to our metrics.

The attacker models in Section 3.2 are used to train three distinct attackers. The *Original* setting represents the baseline performance. We also compute the chance-level performance achieved by a random (a.k.a., *trivial*) attacker defined as 37% for Singling Out, $1/N'$ for Linkability, and 50% for $1-$EER.

4. Results

Figure 1 summarizes our results across the three metrics displayed in three rows, with columns corresponding to conversation lengths $L =$1, 3, or 30.

For example, at $L = 1$ the original data exhibit a Linkability of 82% with 20 enrollment speakers, decreasing to 35% with 10,000 speakers. After anonymization, the *Informed* attacker's Linkability decreases to 77% for 20 enrollment speakers and 21% for 10,000 speakers, while the *Semi-Informed* and *Ignorant* attackers yield even lower values approaching chance performance. Similarly, for Singling Out, the original data range from 57% down to 38%, and anonymization reduces this rate further, especially for less informed attackers. By contrast, $1-$EER remains remarkably stable at around 80–90% with very little variation.

As L increases, both the Linkability and Singling Out metrics indicate an increased re-identification risk. For $L = 3$ and $L = 30$, the original data reach very high values (up to 94–95%), and even after anonymization the *Informed* attacker achieves up to 96% (Linkability) and 99% (Singling Out), while the *Semi-Informed* and *Ignorant* attackers continue to perform above chance-level. This trend demonstrates that just a little more data per speaker, reflected by longer conversation lengths, leads to a much higher re-identification risk, a nuance that $1-$EER also mostly fails to capture due to its stability.

In summary, our results illustrate that the EER is not sensitive to the true re-identification risk. In contrast, besides ensuring legal protection, the proposed metrics reveal significant variations depending on the use case scenario, which provide a more sensitive and nuanced evaluation of voice anonymization.

[4] When a speaker has $2 \leq K < 10$ x-vectors only, $K - 1$ are used for calibration and 1 for test. The threshold is then equal to the average of the $(K-1)$-th and K-th similarity scores sorted by decreasing value. When $L = 30$, test speakers with less than $2 \times L$ utterances are excluded to ensure $K \geq 2$, hence the maximum N is below 22,024.

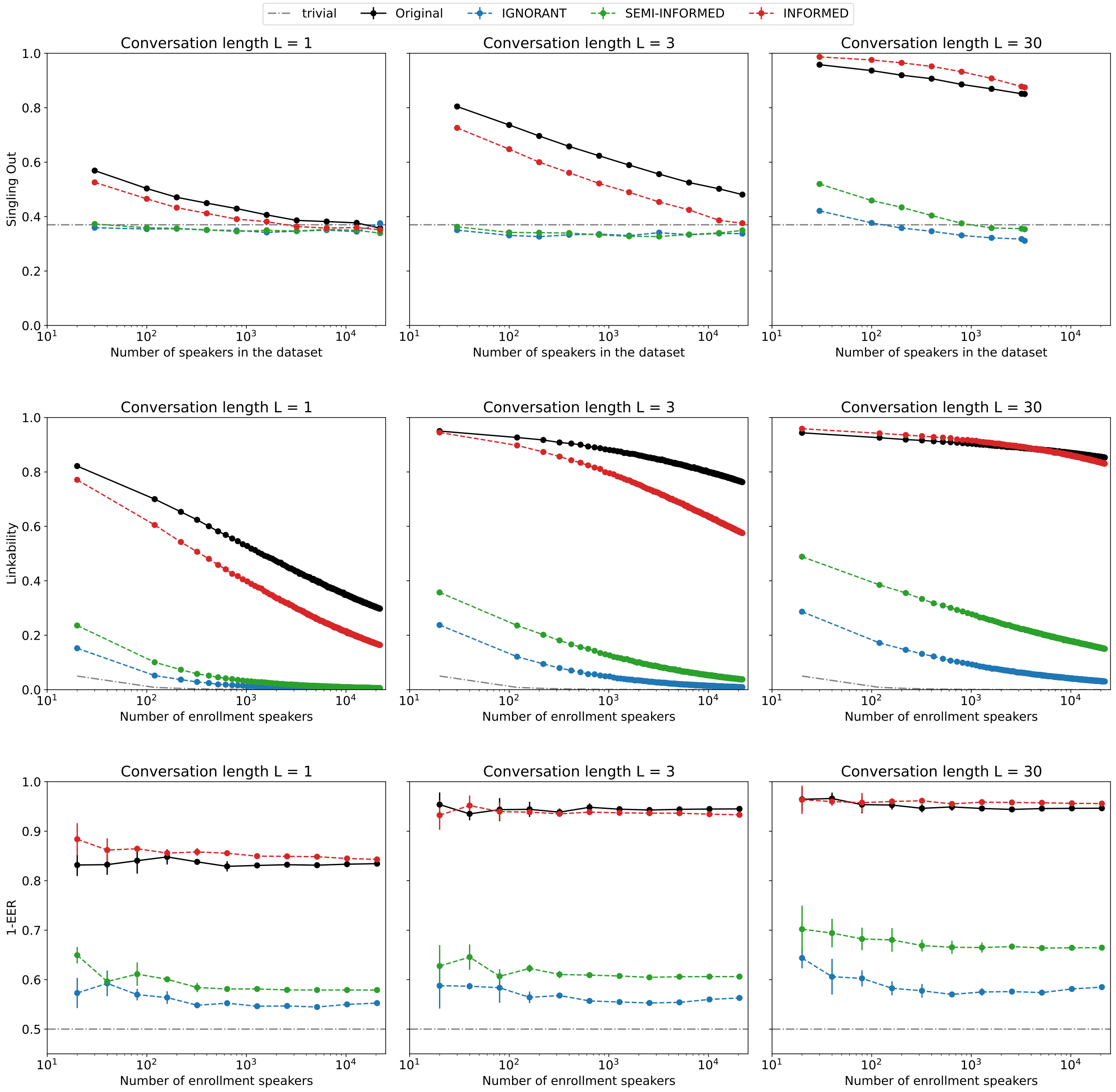

Figure 1: *Evaluation results in terms of Singling Out (row 1), Linkability (row 2) and* $1-$*EER (row 3). Columns indicate conversation length* $L = 1$, $L = 3$ *and* $L = 30$. *In each subplot, performance curves for original (non-anonymized) speech and anonymized speech under different attack models (*Informed, Semi-Informed, *and* Ignorant*) are shown alongside the trivial (chance-level) performance. In all curves, lower value indicates lower re-identification risk.*

5. Conclusion and future directions

In this paper, we addressed the critical gap between the conventional evaluation metrics and the legally required assessment of re-identification risk in voice anonymization. By introducing two legally grounded metrics, Linkability and Singling Out, we demonstrated that the conventional EER metric fails to capture the nuanced variations in residual privacy risk under varying attack conditions. Our experimental results showed that legal re-identification risk increases significantly with longer conversation lengths. By contrast, the EER remains nearly invariant, thus underestimating the actual risk.

Future work will explore extending our framework to capture other dimensions of privacy leakage based on the linguistic content and inference attacks, as well as evaluating the performance of emerging anonymization techniques across a broader range of datasets, languages, and worst-case scenarios. Additionally, refining the calibration of our legal metrics and integrating them into publicly available assessment tools will be a critical effort for advancing robust privacy protection in practical applications.

6. References

[1] P. Deepa and R. Khilar, "Speech technology in healthcare," *Measurement: Sensors*, vol. 24, p. 100565, 2022.

[2] J. Zhang, J. Wu, Y. Qiu, A. Song, W. Li, X. Li, and Y. Liu, "Intelligent speech technologies for transcription, disease diagnosis, and medical equipment interactive control in smart hospitals: A review," *Computers in Biology and Medicine*, vol. 153, p. 106517, 2023.

[3] A. Nautsch, A. Jimenez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, M. A. Hmani, A. Mtibaa, M. A. Abdelraheem, A. Abad, F. Teixeira, D. Matrouf, M. Gomez-Barrero, D. Petrovska-Delacrétaz, G. Chollet, N. Evans, T. Schneider, J.-F. Bonastre, and C. Busch, "Preserving privacy in speaker and speech characterisation," *Computer Speech and Language*, vol. 58, pp. 441–480, 2019.

[4] S. Tayebi Arasteh, T. Arias-Vergara, P. A. Pérez-Toro, T. Weise, K. Packhäuser, M. Schuster, E. Noeth, A. Maier, and S. H. Yang, "Addressing challenges in speaker anonymization to maintain utility while ensuring privacy of pathological speech," *Communications Medicine*, vol. 4, no. 1, p. 182, 2024.

[5] B. M. L. Srivastava, A. Bellet, M. Tommasi, and E. Vincent, "Privacy-preserving adversarial representation learning in ASR: Reality or illusion?" in *Interspeech*, 2019, pp. 3700–3704.

[6] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker anonymization using x-vector and neural waveform models," in *Speech Synthesis Workshop*, 2019, pp. 155–160.

[7] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, "Introducing the VoicePrivacy initiative," in *Interspeech*, 2020, pp. 1693–1697.

[8] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, "Speaker anonymisation using the McAdams coefficient," in *Interspeech*, 2021, pp. 1099–1103.

[9] B. M. L. Srivastava, M. Maouche, M. Sahidullah, E. Vincent, A. Bellet, M. Tommasi, N. Tomashenko, X. Wang, and J. Yamagishi, "Privacy and utility of x-vector based speaker anonymization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2383–2395, 2022.

[10] P. Champion, A. Larcher, and D. Jouvet, "Are disentangled representations all you need to build speaker anonymization systems?" in *Interspeech*, 2022, pp. 2793–2797.

[11] C. O. Mawalim, K. Galajit, J. Karnjana, S. Kidani, and M. Unoki, "Speaker anonymization by modifying fundamental frequency and x-vector singular value," *Computer Speech & Language*, vol. 73, p. 101326, 2022.

[12] S. Meyer, F. Lux, J. Koch, P. Denisov, P. Tilli, and N. T. Vu, "Prosody is not identity: A speaker anonymization approach using prosody cloning," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[13] M. Panariello, F. Nespoli, M. Todisco, and N. Evans, "Speaker anonymization using neural audio codec language models," in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 4725–4729.

[14] N. Kuzmin, H.-T. Luong, J. Yao, L. Xie, K. A. Lee, and E.-S. Chng, "NTU-NPU system for Voice Privacy 2024 Challenge," in *4th Symposium on Security and Privacy in Speech Communication*, 2024, pp. 72–79.

[15] Z. Cai, H. L. Xinyuan, A. Garg, L. P. García-Perera, K. Duh, S. Khudanpur, N. Andrews, and M. Wiesner, "Privacy versus emotion preservation trade-offs in emotion-preserving speaker anonymization," in *2024 IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 409–414.

[16] "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)," https://eur-lex.europa.eu/eli/reg/2016/679/oj, 2016.

[17] Article 29 Working Party, "Opinion 05/2014 on Anonymisation Techniques," https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf, 2014.

[18] N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P.-G. Noé, A. Nautsch, N. Evans, J. Yamagishi, B. O'Brien, A. Chanclu, J.-F. Bonastre, M. Todisco, and M. Maouche, "The VoicePrivacy 2020 Challenge: Results and findings," *Computer Speech & Language*, vol. 74, p. 101362, 2022.

[19] M. Panariello, N. Tomashenko, X. Wang, X. Miao, P. Champion, H. Nourtel, M. Todisco, N. Evans, E. Vincent, and J. Yamagishi, "The VoicePrivacy 2022 Challenge: Progress and perspectives in voice anonymisation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[20] N. Tomashenko, X. Miao, P. Champion, S. Meyer, X. Wang, E. Vincent, M. Panariello, N. Evans, J. Yamagishi, and M. Todisco, "The Voice Privacy 2024 Challenge evaluation plan," https://inria.hal.science/hal-04531444, 2024.

[21] N. Tomashenko, X. Miao, E. Vincent, and J. Yamagishi, "The First VoicePrivacy Attacker Challenge evaluation plan," https://hal.science/hal-04730990, 2024.

[22] S. Meyer, X. Miao, and N. T. Vu, "VoicePAT: An efficient open-source evaluation toolkit for voice privacy research," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 257–265, 2023.

[23] G. Cohen, M. Aloni, and K. Nissim, "Towards formalizing the GDPR's notion of singling out," *Proceedings of the National Academy of Sciences*, vol. 117, no. 15, pp. 8344–8352, 2020.

[24] M. Gomez-Barrero, J. Galbally, C. Rathgeb, and C. Busch, "General framework to evaluate unlinkability in biometric template protection systems," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 6, pp. 1406–1420, 2017.

[25] M. Maouche, B. M. L. Srivastava, N. Vauquier, A. Bellet, M. Tommasi, and E. Vincent, "A comparative study of speech anonymization metrics," in *Interspeech*, 2020, pp. 1708–1712.

[26] B. M. L. Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, and E. Vincent, "Evaluating voice conversion-based privacy protection against informed attackers," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 2802–2806.

[27] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[28] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *12th Language Resources and Evaluation Conference (LREC)*, 2020, pp. 4218–4222.

[29] A. Larcher, K. A. Lee, and S. Meignier, "An extensible speaker identification sidekit in python," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5095–5099.

[30] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech*, 2020, pp. 3830–3834.

[31] A. Nautsch, "Speaker recognition in unconstrained environments," Ph.D. dissertation, Technische Universität Darmstadt, 2019.