Odyssey 2026
23-26 June 2026, Lisbon, Portugal

Evaluating voice anonymisation using similarity rank disclosure

Shilpa Chandra[1,3], *Matteo Pettenò*[1,4], *Michele Panariello*[1], *Nicholas Evans*[1], *Massimiliano Todisco*[1]
Tom Bäckström[2], *Dorothea Kolossa*[3], *Rainer Martin*[4], *Themos Stafylakis*[5,6,7], *Nicolas Gengembre*[8]

[1]EURECOM, France; [2]Aalto University, Finland; [3]Technische Universität Berlin, Germany
[4]Ruhr-Universität Bochum, Germany; [5]Omilia, Greece; [6]Athens University of Economics and Business, Greece; [7]Archimedes/Athena R.C, Greece; [8]Orange, France

{shilpa.chandra, matteo.petteno, michele.panariello, nicholas.evans, massimiliano.todisco}@eurecom.fr, tom.backstrom@aalto.fi, rainer.martin@rub.de
dorothea.kolossa@tu-berlin.de, tstafylakis@omilia.com, nicolas.gengembre@orange.com

Abstract

The evaluation of voice anonymisation remains challenging. Current practice relies on automatic speaker verification metrics such as the equal error rate (EER). Performance estimates dependent on the classifier and operating point provide an incomplete or even misleading characterisation of privacy risk. We investigate the use of similarity rank disclosure (SRD), an information-theoretic metric, which operates on feature representations rather than classifier decisions, providing a threshold-independent assessment of privacy and analysis of both average and worst-case disclosure. We report its application to speaker embeddings, fundamental frequency, and phone embeddings using 2024 VoicePrivacy Challenge systems. The SRD reveals privacy leaks and system-specific weaknesses missed by EER-based evaluation. Findings highlight the merit of representation-level metrics and demonstrate the potential of SRD as a flexible and interpretable tool for the evaluation of voice anonymisation.

Index Terms: voice anonymisation, privacy, evaluation

1. Introduction

Smart devices and cloud services are nowadays constantly capturing and processing speech data [1, 2]. Beyond voice identity, speech recordings can reveal sensitive attributes such as the speaker's age, gender, emotional state, etc. [2, 3]. The pervasive capture and inherent sensitivity of speech data, coupled with evolving privacy regulation [4] have stimulated growing research interest in privacy for smart speech technology.

The VoicePrivacy Challenge (VPC) [5], was founded in 2020 to foster the development of voice anonymisation solutions. Anonymisation is used to substitute the voice in a speech recording with that of another speaker or a pseudo-voice, preventing it from being linked to the original identity. Depending on the specific application, however, other utility-related attributes, such as the linguistic and paralinguistic content, should be preserved.

There is hence an inherent trade-off between privacy and utility. Perfect anonymisation can be achieved simply by removing the speech content entirely. This trivial solution, however, has no practical value; the result is functionally useless. While the evaluation of voice anonymisation systems hence requires careful assessment of the balance between privacy protection and utility preservation [6], our focus in this paper is exclusively upon evaluation of the former.

Privacy protection is usually evaluated using automatic speaker verification (ASV) and related metrics. For the VPC, the *Equal Error Rate (EER)* serves as the primary privacy metric. The EER and other related metrics, however, depend fundamentally on not just anonymisation performance, but also the ASV model. Estimates of privacy can then vary with the choice of ASV system, the features or embeddings employed, the training data with which it is trained, the distance measure, the operating point, etc. As a result, estimates of privacy can then be as much a function of the ASV system used for evaluation as of voice anonymisation performance.

We have investigated the use of a recently proposed privacy metric, the *Similarity Rank Disclosure (SRD)* [7], for the evaluation of voice anonymisation solutions. The SRD is not dependent upon the binary decisions of an ASV system or similar classifiers and can instead be applied directly to *any* features or embeddings likely to capture speaker-related information. The SRD can also provide estimates of average and worst-case privacy disclosure in information bits and could be applied readily to the evaluation of privacy leakage from the disclosure of other soft attributes beyond those immediately related to identity.

To demonstrate its flexibility, we demonstrate use of the SRD to estimate anonymisation performance and privacy disclosure from speaker embeddings, fundamental frequency and phone distributions [8–11]. Using 2024 VPC protocols and data, we furthermore show that the SRD can reveal anonymisation weaknesses that are missed by evaluations performed using ASV systems and EER estimates.

The remainder of the paper is structured as follows. In Section 2, we describe the existing metrics in voice privacy and current approaches to the evaluation of voice anonymisation systems. In Section 3, we describe our use of the SRD for evaluation. Our experimental setup is described in Section 4. We present results and discussion in Sections 5 and 6, followed by conclusions in Section 7.

2. Related work

The strength of any approach to privacy protection is usually estimated empirically according to a particular threat model [12] and simulated attacks launched to defeat the protection. The strength is then quantified according to some objective metric that indicates the attack success rate. In the case of voice anonymisation, the metric reflects the extent to which the original speaker of an anonymised utterance can be re-identified. In practice, re-identification is formulated as a binary ASV detection task.

A number of objective metrics have been proposed and applied to the evaluation of voice anonymisation systems [13]. The EER was adopted as the primary privacy metric since the first VPC held in 2020 [14]. EERs might not, however, reflect

*This work was performed while the second author was a PhD candidate at EURECOM and Ruhr-Universität Bochum.

10.21437/Odyssey.2026-17

the realistic cost model of a privacy attacker who might instead prioritise a lower rate of missed detections while tolerating a higher rate of false alarms. Moreover, it can be shown that, under calibrated scores, the EER corresponds to the upper bound of the total error rate [15, 16] and the worst possible decision policy of a privacy adversary [17]. Use of the EER can hence result in exaggerated estimates of privacy protection.

The *log-likelihood-ratio cost function* [18] offers one solution to overcome this limitation. Derived from the total log-likelihood-ratio cost (C_{llr}), it isolates discrimination capability via optimal calibration, often using the pool adjacent violators algorithm. $C_{\text{llr}}^{\text{min}}$ provides a scalar summary of the receiver operating characteristic (ROC) convex hull, offering a robust, threshold-independent measure that nonetheless still correlates well with the EER in the case of voice anonymisation.

The *Zero Evidence Biometric Recognition Assessment (ZEBRA)* framework [17] quantifies privacy disclosure based upon the empirical cross-entropy (ECE). The expected privacy disclosure D_{ECE} is an estimate of the average information disclosed to an adversary in bits, while the worst-case privacy disclosure quantifies the maximum evidence revealed for any individual.

While the C_{llr} and ZEBRA focus on information-theoretic aspects of privacy, the Linkability and Singling Out metrics introduced in [19, 20] provide a complementary, legally grounded perspective. Drawing on the GDPR and related privacy regulation, they capture residual re-identification risks that conventional measures may overlook. Linkability quantifies the probability that an anonymised sample can be correctly linked to the original speaker. Singling Out reflects the likelihood that an individual can be uniquely isolated from anonymised data, addressing risk even when identity remains unknown.

As we shall see, the SRD combines the merits of existing metrics into a single, flexible framework for the evaluation of voice anonymisation solutions and provides an interpretable understanding of how much personal identifying information (PII) remains in different representations of protected speech data.

3. Similarity Rank Disclosure

The SRD provides a framework to measure PII contained within speech utterances [7]. Information is quantified in bits, hence enabling comparisons between the disclosure of information contained in different speech characteristics. Operating directly upon speech features instead of classifier decisions, the SRD provides a classifier threshold-independent and entropy-grounded measure of privacy disclosure. With a full definition of the SRD available in the original work [7], we provide in the following only an overview of the SRD and necessary prerequisites relevant to its application as a metric for the evaluation of voice anonymisation solutions.

3.1. Similarity evaluation and empirical ranking

Computation of the SRD involves comparisons between feature representations extracted from input utterances x to a set of representations extracted from a database of N reference utterances y. The database contains one reference which matches the voice identity in each input x. We then proceed as follows.

1. **Ranking -** For every input x, we evaluate the similarity to each reference y. The set of resulting similarities is rank-ordered from the most similar (rank 1) to least similar (rank N). For each x, we then determine the rank k of the matching reference.

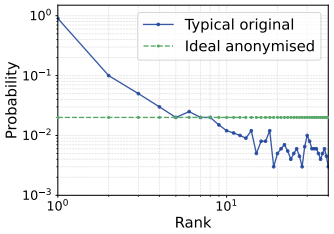

Figure 1: *Typical and ideal rank histogram distributions for original and anonymised data.*

2. **Distribution generation -** A histogram of matching ranks k is derived from a set of inputs x. The histogram is normalized as an *empirical probability distribution* $\tilde{p}_k$ of matching speaker rank.

Example empirical probability distributions are shown in Figure 1 for a database containing $N = 40$ references. The blue profile illustrates a typical distribution for original, unprotected data. The largely-negative slope indicates a greater probability of matching speakers being in lower than in higher rank positions. The voice in inputs x are readily identified and matched to the corresponding reference in the database. Lower probabilities for higher ranks imply little confusion between speakers in the input and other, non-matching references.

The green profile illustrates the ideal distribution expected for perfectly protected data. This distribution should be uniform ($\tilde{p}_k = 0.025$ for $N = 40$); the voices in the inputs x are as likely to correspond to the matching reference as to *any* other speaker in the database. In practice, protection is expected to be imperfect thus we expect distributions between those for unprotected and perfectly protected data.

3.2. Statistical modelling

With plentiful data, the normalised histogram $\tilde{p}_k$ can be used directly for further analysis. If, however, data is sparse, then further analysis can be performed using statistical, parametric distributions estimated from empirical distributions. The original work [7] suggests use of the beta-binomial distribution.

1. **Beta-binomial fit -** The fitting of a *beta-binomial distribution* to the empirical distribution $\tilde{p}_k$ supports the estimation of smoothed probabilities γ_j that input x corresponds to the j^{th} best match.
2. **Parameter estimation -** Due to a high penalty for distribution deviations at the crucial rank-1 position, for which disclosure is highest, model parameters α and β are optimised using the *constrained log-likelihood* loss function.

3.3. Disclosure quantification

The rank order disclosure ϵ_j quantifies the reduction in uncertainty (entropy) gained from observing the matching rank j.

1. **Rank order disclosure -** Before any observations, assuming all N identities are equally likely as in the case of a uniform distribution, then the prior information needed to encode identity is $\log_2 N$ bits. After observing the rank j of the matching reference, the posterior probability given by γ_j can be estimated using either $\tilde{p}_j$ or a beta-binomial approximation. The information required is then reduced to $-\log_2 \gamma_j$ bits. The *rank order disclosure* ϵ_j given rank j is then expressed in bits as:

$$\epsilon_j := -\log_2 \gamma_j - \log_2 N \tag{1}$$

The subtraction in (1), rather than addition, results in an expression of disclosure rather than of entropy or uncertainty.

2. **Statistical summary -** Further insights into privacy disclosure can be gained from rank order disclosure distribution statistics:
   - **Mean Disclosure (MeanD) -** The average disclosure across all observations.
   - **Maximum Disclosure (MaxD) -** The single, worst-case disclosure observed.
   - **Identification Rate (IdR) -** The probability that the matching reference is ranked first ($\tilde{p}_1$).
   - **Rank Spread -** The proportion of ranks whose probability exceeds the level of pure chance ($1/N$).

Use of the SRD provides a characterisation of privacy disclosure in bits. By focusing directly on feature representations, the SRD can also be used to compare privacy disclosure across any suitable feature representation, e.g., the fundamental frequency, phone embeddings, or speaker embeddings. By decoupling evaluation from classifier decisions, estimates of privacy disclosure better expose and quantify risks arising from information contained in the representation itself. We use the SRD to evaluate the privacy leakage in voice anonymisation.

4. Experimental Setup

While the original work [7] reports a study of privacy disclosure for original, unprotected speech data, we have applied the SRD to the study and comparison of privacy disclosure for speech data treated with different approaches to voice anonymisation [6]. In the following section, we describe the data used and a set of four feature representations.

4.1. Anonymisation systems, data and protocols

We evaluate both baseline and top-performing 2024 VPC systems [6] also used for the 2024 VoicePrivacy Attacker Challenge [21]. These include baselines B3, B4, and B5, and participant systems T8-5, T10-2, T12-5, and T25-1. We report results derived using the 2024 VPC evaluation set only.

VPC protocols define sets of enrolment and trial utterances, both of which may be anonymised. Privacy protection is then estimated using an ASV system and sets of target and non-target trials, with each involving a comparison between one enrolment utterance and one trial utterance. Using the SRD, instead of using ASV detection scores, we estimate privacy protection using similarity ranking. The 2024 VPC protocols provide for only 29 speakers that are common to both enrolment and trial sets. To provide for a larger number, we pool the sets of all enrolment and trial utterances (*libri_test_enrolls*, *libri_test_trials_f*, and *libri_test_trials_m*) and construct custom, utterance-disjoint input and reference sets (see Section 3.1) following the procedures in [7, Sec. IV]. This process yields 40 speakers common to both input and reference sets while also providing a greater number of observations per speaker.

4.2. Feature representations

We demonstrate the flexibility of the SRD with experiments using a selection of different feature representations known to capture speaker-dependent characteristics. The selection is inspired by the features used in [7], namely the fundamental frequency, phone, and speaker representations. For the latter, we use representations extracted using the same ECAPA-TDNN model [10] used in [7] as well as for VPC evaluations [6, 14, 22]. In addition, we make use of a WavLM model [23] trained to use non-timbral cues [11].

It is well known that more reliable interpretations of protection can be derived using stronger attacks and embeddings extracted using models trained with similarly anonymised data. This approach, referred to as the *semi-informed* attack model, is now standard practice for VPC evaluations [6, 14, 22]. We also use speaker embedding models trained this way.

We make no use of linguistic embeddings as in [7]. While also a source of speaker information, the VPC concerns strictly *voice* characteristics rather than spoken content as sources of PII. None of the baseline systems, nor participant systems, manipulate linguistic information.[1] Furthermore, the LibriSpeech dataset consists of audiobook recordings sourced from the LibriVox project [24]. It is hence a reasonable assumption that linguistic content reflects more the linguistic preferences of the book authors rather than those of the narrators/speakers. Consequently, linguistic information is less likely to be a source of privacy disclosure than the other features used in our work.

In the following, we describe the extraction of each feature representation used in our work and any differences to the implementations used in [7].

4.2.1. ECAPA-TDNN (ET)

We use the SpeechBrain [25] ECAPA-TDNN model [10] to extract speaker embeddings and train two models. The first, denoted ET_{orig} is trained using original speech. The second, denoted ET_{anon}, is trained under the *semi-informed* attack setting using anonymised speech, as described in Section 5. The empirical probability distribution is derived from the cosine distance between embeddings extracted from input and reference pairs.

4.2.2. Non-timbral embeddings (W-NT)

Recent work [11, 26] shows that non-timbral cues such as prosody, rhythm, speaking style, and accent can still support speaker re-identification after anonymisation. First, the authors use the self-supervised learning based WavLM model [23] for speech processing. Second, to better capture non-timbral cues, the model is fine-tuned using data processed using voice conversion (VC) to manipulate or obfuscate predominantly timbral characteristics. It is assumed that more non-timbral speaker-dependent cues such as rhythm and speaking rate are preserved and are then captured. Again, we use two models. The first, denoted as W-NT (WavLM, non-timbral), is trained using original data. The second, denoted as $W\text{-}NT_{\text{anon}}$, is trained using anonymised data under the same *semi-informed* attack setting. Both models were provided by the authors of [11]. The empirical probability distribution is derived in the same way as for ECAPA-TDNN embeddings.

4.2.3. Fundamental frequency

The fundamental frequency F0 captures the primary periodicity of voiced speech. F0 dynamics correlate with speaker-specific traits such as pitch range, prosodic style, and vocal characteristics. We estimate F0 using the pYIN algorithm [27], an improved version of the original YIN algorithm [8]. As in the original work [7], we limit the F0 range to between 65 and 450 Hz. pYIN estimates pitch lags as integers, which results in 107 distinct fundamental frequency values. We estimate F0

[1] It is acknowledged that this might not be the case for baseline B3 and other ASR-TTS-based anonymisation systems.

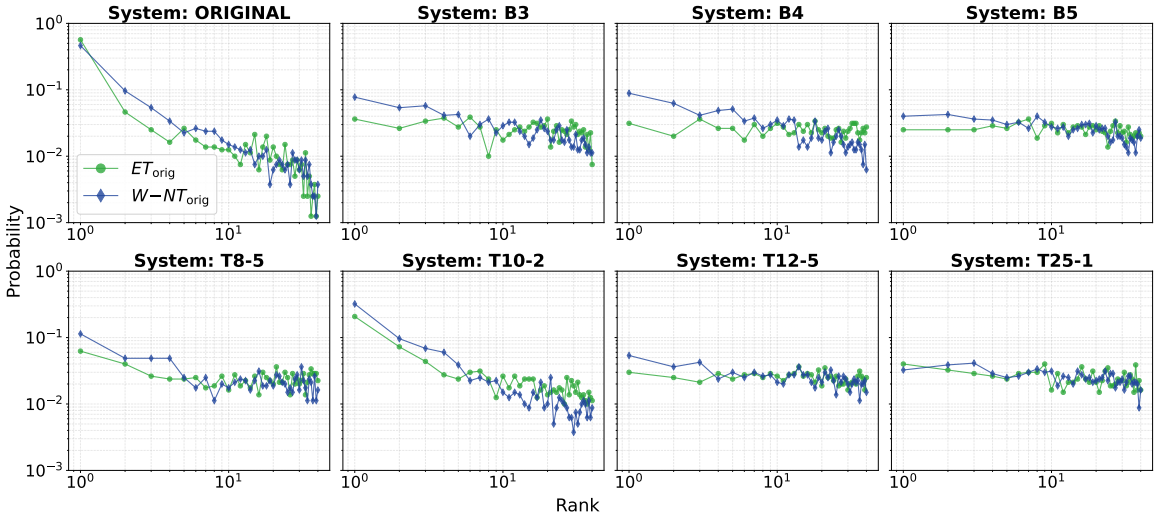

Figure 2: *Rank histograms for the ECAPA-TDNN embeddings ET_{orig} and non-timbral embeddings $W\text{-}NT_{orig}$, for original speech (top left plot only) and anonymised speech (all others).*

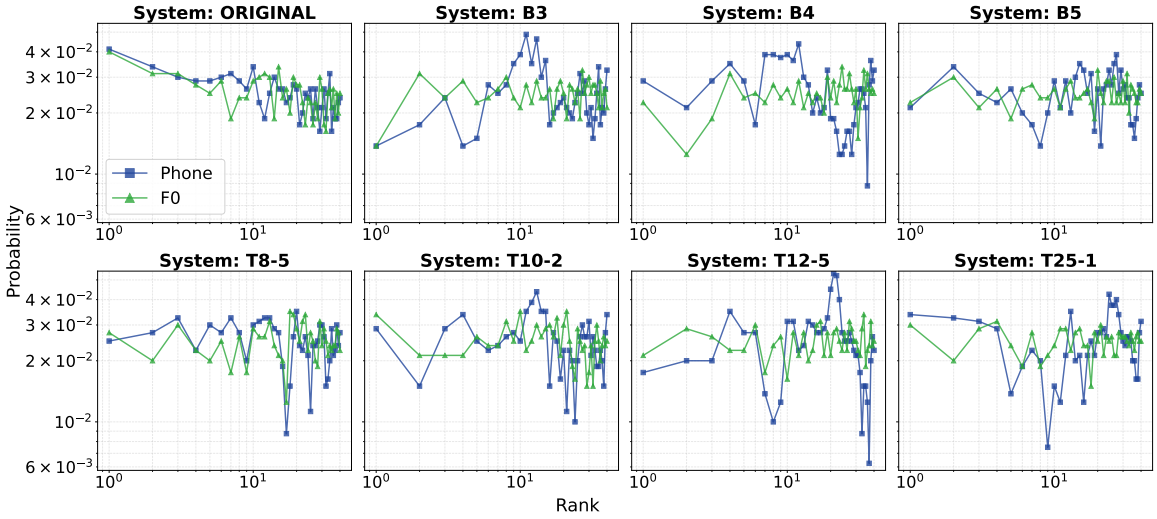

Figure 3: *As for Figure 2 except for fundamental frequency and phone embeddings.*

using frames of 30 s with a stride of 10 ms, thereby producing in the order of 3000 estimates over an interval of 30 s, though, after discarding non-voiced frames, this number is lower in practice. To obtain the empirical probability distribution we compute the euclidean distance between normalised F0 histograms for each input and reference.

4.2.4. Phone embeddings

Phone-related features capture speaker-specific articulation patterns and phonetic realisations that may persist after anonymization. We use a VQ-VAE-based voice conversion model to extract phone features [28]. The models learn discrete acoustic units (pseudo-phones) in an unsupervised manner, enabling phone-like segmentation without the need for transcriptions or text normalisation. Non-speech intervals are first removed using energy-based voice activity detection (VAD). Remaining speech is then partitioned into 20 uniform segments from which phone embeddings in the form of codebook histograms are extracted. To obtain the empirical probability distribution we compute the euclidean distance between phone embeddings for each input and reference.

5. Results

We present rank histograms for features described in Section 4.2 and differences in qualitative results derived using a stronger semi-informed attack model. We then present quantitative results derived using metrics described in Section 3.3, followed by a comparison to results from the use of statistical approximations described in Section 3.2.

5.1. ET and W-NT embeddings

Rank histograms for original speech and anonymised speech produced using each of the seven different anonymisation systems are shown in Figure 2 for ET_{orig} embeddings (green profiles) and $W\text{-}NT_{\text{orig}}$ embeddings (blue profiles).

Profiles for original speech shown to the top left show mostly similar trends. The comparatively high probabilities for the first histogram bins (rank-1) and the steep, negative slopes show that speakers can be readily identified using either embeddings; the true speaker identity is most often in the rank-1 position. The comparatively lower probabilities for other bins show that other reference speakers are rarely confused with the speaker in the input.

All other plots show rank histograms for anonymised speech. Rank-1 values are universally lower than for original speech for both ET_{orig} and $W\text{-}NT_{\text{orig}}$ embeddings, indicating that all systems provide some level of privacy. For systems B5, T12-5 and T25-1, the near-to-flat slopes indicate that other speakers are now more likely to be confused with the original speaker, indicating stronger privacy. Steeper slopes for system T10-2 and, to a lesser extent, also T8-5, indicate weaker privacy. Even if rank-1 probabilities are lower than those for original speech, the negative slopes for each profile indicate that, in most cases, the original speaker remains frequently identifiable.

Interestingly, while rank-1 values for original speech are slightly higher for ET_{orig} embeddings, for all anonymisation systems other than T25-1, they are higher for $W\text{-}NT_{\text{orig}}$ embeddings. This result is expected since most anonymisation systems obfuscate predominantly timbral rather than non-timbral cues. Consequently, after anonymisation, the latter rather than the former are often more informative. That we observe the opposite for the stronger T25-1 system is an indication that it is more successful at obfuscating both timbral and non-timbral cues.

5.2. Fundamental frequency and Phone embeddings

Rank histograms shown in Figure 3 show similar profiles for F0 (green profiles) and phone embeddings (blue profiles). Profiles for original speech shown in the top left plot show trends similar to those for speaker embeddings. While rank-1 probabilities are still highest, the negative slopes are less pronounced. These observations indicate that F0 and phone embeddings are less informative than speaker embeddings. Profiles for anonymised speech show that all anonymisation systems are successful in increasing the confusion between input and reference speakers and that, in most cases, the original speaker can no longer be identified.

5.3. Semi-informed attacks

All results illustrated in Figure 4 are computed using embeddings extracted using models trained with anonymised data according to the VPC semi-informed attack model. Rank histograms for embeddings extracted using models retrained with data anonymised by each respective system [11, 21] are shown in Figure 4. Profiles are shown for ET_{anon} embeddings (green profiles) and $W\text{-}NT_{\text{anon}}$ embeddings (blue profiles). Rank-1 probabilities for ET_{anon} and $W\text{-}NT_{\text{anon}}$ models (Fig. 4) are higher than those for ET_{orig} and $W\text{-}NT_{\text{orig}}$ models (Fig. 2). Profile slopes, while near-to-flat for extractors trained using original data, are mostly negative when retrained with anonymised data. For all but the T10-2 system, rank-1 probabilities are higher for the $W\text{-}NT_{\text{anon}}$ model than for ET_{anon}, confirming recent findings [11].

5.4. Privacy disclosure

While the more qualitative findings above provide an insightful picture, rank histograms alone do not offer a means to compare the strength of different systems. Quantitative comparisons are made using the SRD metrics described in Section 3.3. In keeping with VPC policy, and to facilitate performance comprisons, we present results derived using the ET_{orig} and ET_{anon} models only.[2] Table 1 shows maximum disclosure (MaxD), mean disclosure (MeanD), identification rate (IdR) and rank spread (RS) results for both original speech and speech anonymised using each of the seven systems (first column). Results for ET_{orig} are shown to the left and for ET_{anon} to the right. EER results, computed according to the standard VPC protocol under the semi-informed attack model are shown for comparison (last column).

Results again show the importance of using strong attack models for evaluation. While, for ET_{orig}, maximum and mean disclosures are lowest for systems B4, B5 and T12-5, they are lower for the T8-5 system in the case of ET_{anon} embeddings. The identification rate is lowest for system B5 in the case of ET_{orig} embeddings, but for system T8-5 in the case of ET_{anon} embeddings. While systems B4 and T12-5 have the highest rank spread initially, system T12-5 fares best for the strongest attack. Reassuringly, the better results for T8-5 in most cases correspond to the highest EER, albeit by a small margin.

5.5. Beta-binomial

Results for the ET_{anon} system derived from approximations to the empirical rank histograms using fitted beta binomial distributions (see Sec. 3.2) are shown in Table 2. While there are some numerical discrepancies, trends are identical. System T8-5 provides the lowest maximum and mean disclosures and the lowest identification rate, while the rank spread is again lowest for system T12-5, and now also system B3. It is evident that, despite the low number of speakers compared to the original work [7], differences between the performance of competing anonymisation systems are substantial enough so that identical trends are derived using either rank histograms directly or statistical, parametric approximations.

6. Discussion

The SRD provides revealing insights into the differences in privacy protection for competing anonymisation solutions. These go beyond a single snapshot like that provided from estimates of the EER in the form of the mean and worst case disclosure and the rank spread. By casting evaluation as an identification problem instead of verification, the SRD allows us to explore not just whether a voice is *sufficiently* dissimilar (to the original voice) after anonymisation, but *by how much* the anonymised voice is confusable with the voices of other speakers. Comparisons between mean and worst-case disclosure in bits also allows meaningful insights into privacy fairness. While we explored speaker embeddings, F0 and phone statistics in this paper, the SRD is readily adapted to representations of accent, of speaker sex, of prosodic features beyond F0, phone durations [29], etc., and supports the comparison of privacy disclosure coming from any such source of PII.

In one particular case, the SRD reveals stark differences to EER results: Table 1 shows that, in terms of EER, system T10-2 performs just as well as system T8-5, while SRD results uncover a different picture. Rank histograms for system T10-2 in Figure 2 show a high rank-1 probability, while Table 1 shows high maximum and mean disclosures. The identification rate is 70%. The cause of overestimated privacy in case of the EER is now well known [30] and stems from anonymisation being

[2] Results for other models and feature representations are available at `obscured_for_blind_review`.

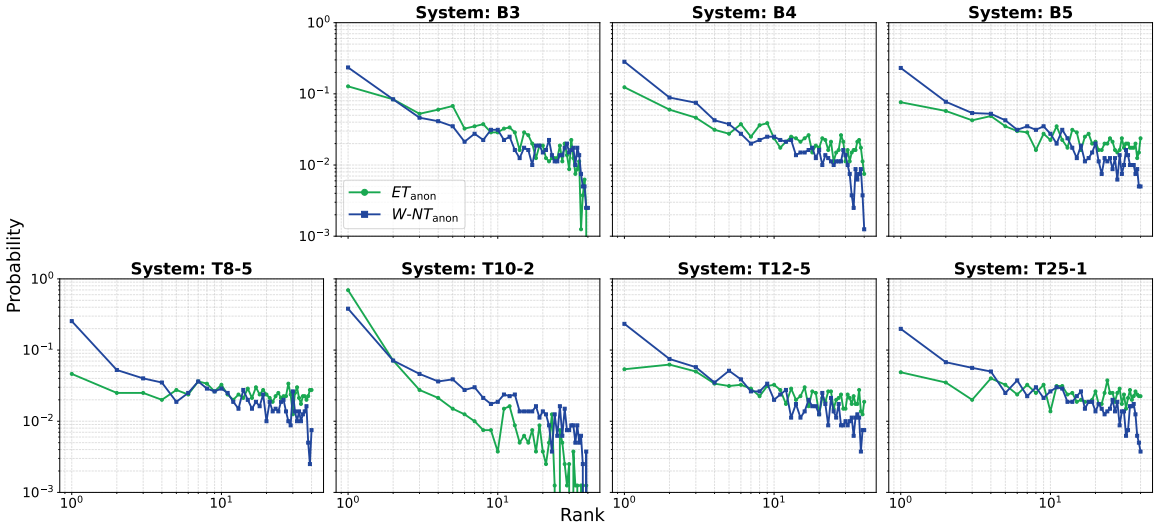

Figure 4: *As for Figure 2 except for ET_{anon} and $W\text{-}NT_{anon}$ embeddings and the semi-informed attack scenario for which embedding extractors are trained using similarly anonymised speech data.*

Table 1: *SRD results in terms of maximum (MaxD) and mean (MeanD) disclosure, identification rate (IdR) and rank spread (RS) for ET_{orig} and ET_{anon} embeddings. Last column shows corresponding EERs for the 2024 VPC.*

System	ET_{orig}				ET_{anon}				
	MaxD ↓	MeanD ↓	IdR (%) ↓	RS ↑	MaxD ↓	MeanD ↓	IdR (%) ↓	RS ↑	EER (%) ↑
Original	4.50	2.19	56.63	7.50	–	–	–	–	**4.6**
B3	0.63	0.06	3.63	45.0	2.35	0.52	12.75	37.5	27.3
B4	**0.53**	**0.02**	3.12	**47.5**	2.30	0.26	12.37	25.0	30.3
B5	**0.53**	**0.02**	**2.50**	42.5	1.60	0.14	7.63	30.0	34.3
T8-5	1.32	0.06	6.25	40.0	**0.88**	**0.03**	**4.62**	32.5	**40.9**
T10-2	3.05	0.52	20.75	20.0	4.79	3.12	69.37	7.50	40.8
T12-5	**0.53**	**0.02**	3.00	**47.5**	1.32	0.11	5.37	**40.0**	33.2
T25-1	0.67	0.05	4.00	40.0	0.96	0.05	4.87	32.5	39.8

Table 2: *As for Table 1 except for results derived using beta-binomial approximations to the empirical rank distribution. ET_{anon} only.*

System	MaxD ↓	MeanD ↓	IdR (%) ↓	RS ↑
Original	4.50	2.75	56.62	15.00
B3	2.35	0.59	12.74	**37.5**
B4	2.30	0.41	12.37	35.0
B5	1.60	0.12	7.62	32.5
T8-5	**0.88**	**0.02**	**4.62**	27.5
T10-2	4.79	3.39	69.37	10.0
T12-5	1.10	0.05	5.37	**37.5**
T25-1	0.96	**0.02**	4.87	30.00

performed at the speaker rather than utterance level, resulting in a weaker attack. Our own, further investigations reveal additional, new insights. They suspect that enrolment data for the T10-2 system are not anonymised. Our assumption is that enrolment utterances are simply re-synthesised with the voice of the original speaker meaning they are effectively unprotected. A high EER is then the result of mismatched enrolment and trial data rather than strong anonymisation. Since we use pooled data (see Sec. 4.1), rather than relying upon distinct enrolment and trial data as for evaluation using ASV and EER estimation, we avoid this potential evaluation pitfall.

Finally, despite the appeal of the SRD and new insights it provides, we stress the importance of using well-trained, strong attack models. Without them, just like for the EER [31], there is always potential for privacy to be overestimated, no matter what the metric, even the SRD. Use of the strongest available attacks hence remains fundamental to reliable evaluation.

7. Conclusions

We investigated use of the similarity rank disclosure (SRD) for evaluating voice anonymisation, providing an information-theoretic assessment of privacy. Compared to the automatic speaker verification equal error rate (EER), the SRD offers a more interpretable and fine-grained characterisation of residual privacy risk. Results for 2024 VoicePrivacy Challenge systems show that the SRD uncovers privacy leakage and system-specific weaknesses that were not immediately apparent from EER-based evaluation. In particular, systems with similar EER performance can exhibit markedly different levels of privacy disclosure, highlighting limitations of current evaluation practices and the need for more comprehensive metrics. By operating directly on feature representations, the SRD is independent of classifier decisions and can be applied to a wide range of representations that encode personally identifying information. The SRD hence offers a flexible and comprehensive tool for the evaluation of voice anonymisation.

8. Acknowledgements

This work was funded by the European Union's Horizon Europe research and innovation programme grant No 101168193. We would also like to thank Rayane Bakari, Nicolas Gengembre, and Olivier Le Blouch (Orange innovation, France) for providing the pre-trained models for W-NT.

9. References

[1] P. Wu, P. P. Liang, J. Shi, R. Salakhutdinov, S. Watanabe, and L.-P. Morency, "Understanding the tradeoffs in client-side privacy for downstream speech tasks," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2021, pp. 841–848.

[2] T. Bäckström, "Privacy in speech technology," *Proceedings of the IEEE*, vol. 113, no. 7, pp. 668–692, 2025.

[3] A. Nautsch, A. Jiménez, 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, B. Raj, I. Trancoso, and C. Busch, "Preserving privacy in speaker and speech characterisation," *Computer Speech & Language*, vol. 58, pp. 441–480, Nov. 2019.

[4] A. Nautsch, C. Jasserand, E. Kindt, M. Todisco, I. Trancoso, and N. Evans, "The GDPR & Speech Data: Reflections of Legal and Technology Communities, First Steps Towards a Common Understanding," in *Interspeech 2019*. ISCA, Sep. 2019, pp. 3695–3699.

[5] 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*. ISCA, Oct. 2020, pp. 1693–1697.

[6] N. Tomashenko, X. Miao, P. Champion, S. Meyer, X. Wang, E. Vincent, M. Panariello, N. Evans, J. Yamagishi, and M. Todisco. The VoicePrivacy 2024 Challenge Evaluation Plan. [Online]. Available: http://arxiv.org/abs/2404.02677

[7] T. Bäckström, M. H. Vali, M. Nguyen, and S. Rech, "Privacy disclosure of similarity rank in speech and language processing," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 34, pp. 196–205, 2026.

[8] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 04 2002. [Online]. Available: https://doi.org/10.1121/1.1458024

[9] B. van Niekerk, L. Nortje, and H. Kamper, "Vector-Quantized Neural Networks for Acoustic Unit Discovery in the ZeroSpeech 2020 Challenge," in *Interspeech 2020*, 2020, pp. 4836–4840.

[10] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Interspeech 2020*, 2020, pp. 3830–3834.

[11] R. Bakari, O. L. Blouch, N. Evans, N. Gengembre, M. Panariello, and M. Todisco, "The influence of non-timbral cues in voice anonymisation and evaluation," in *5th Symposium on Security and Privacy in Speech Communication*, 2025, pp. 35–42.

[12] M. U. Rahman, M. Larson, L. ten Bosch, and C. Tejedor-García, "Scenario of Use Scheme: Threat Modelling for Speaker Privacy Protection in the Medical Domain," in *4th Symposium on Security and Privacy in Speech Communication*, 2024, pp. 21–25.

[13] 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*, 2020.

[14] N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P.-G. Noé, A. Nautsch, N. Evans, J. Yamagishi, B. O'Brien *et al.*, "The VoicePrivacy 2020 challenge: results and findings," *Computer Speech & Language*, vol. 74, p. 101362, 2022.

[15] N. Brümmer, L. Ferrer, and A. Swart, "Out of a Hundred Trials, How Many Errors Does Your Speaker Verifier Make?" in *Interspeech 2021*, 2021, pp. 1059–1063.

[16] T. H. Kinnunen, K. A. Lee, H. Tak, N. Evans, and A. Nautsch, "t-eer: Parameter-free tandem evaluation of countermeasures and biometric comparators," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 2622–2637, 2024.

[17] A. Nautsch, J. Patino, N. Tomashenko, J. Yamagishi, P.-G. Noé, J.-F. Bonastre, M. Todisco, and N. Evans, "The Privacy ZEBRA: Zero Evidence Biometric Recognition Assessment," in *Interspeech 2020*, 2020, pp. 1698–1702.

[18] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2, pp. 230–275, 2006, odyssey 2004: The speaker and Language Recognition Workshop. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230805000483

[19] 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.

[20] N. Vauquier, B. M. L. Srivastava, S. A. Hosseini, and E. Vincent, "Legally validated evaluation framework for voice anonymization," in *Interspeech 2025*, 2025, pp. 3229–3233.

[21] N. Tomashenko, X. Miao, E. Vincent, and J. Yamagishi. The First VoicePrivacy Attacker Challenge Evaluation Plan. [Online]. Available: http://arxiv.org/abs/2410.07428

[22] N. Tomashenko, X. Wang, X. Miao, H. Nourtel, P. Champion, M. Todisco, E. Vincent, N. Evans, J. Yamagishi, and J.-F. Bonastre. The VoicePrivacy 2022 Challenge Evaluation Plan. [Online]. Available: http://arxiv.org/abs/2203.12468

[23] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[24] 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.

[25] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio. SpeechBrain: A General-Purpose Speech Toolkit. [Online]. Available: http://arxiv.org/abs/2106.04624

[26] N. Gengembre, O. Le Blouch, and C. Gendrot, "Disentangling prosody and timbre embeddings via voice conversion," in *Interspeech 2024*, 2024, pp. 2765–2769.

[27] M. Mauch and S. Dixon, "PYIN a fundamental frequency estimator using probabilistic threshold distributions," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 659–663.

[28] M. H. Vali and T. Bäckström, "Interpretable Latent Space Using Space-Filling Curves for Phonetic Analysis in Voice Conversion," in *Interspeech 2023*, 2023, pp. 306–310.

[29] N. Tomashenko, E. Vincent, and M. Tommasi, "Exploiting Context-dependent Duration Features for Voice Anonymization Attack Systems," in *Interspeech 2025*, 2025, pp. 5128–5132.

[30] N. Tomashenko, X. Miao, P. Champion, S. Meyer, M. Panariello, X. Wang, N. Evans, E. Vincent, J. Yamagishi, and M. Todisco, "The third VoicePrivacy challenge: preserving emotional expressiveness and linguistic content in voice anonymization," *arXiv preprint arXiv:2601.11846*, 2026.

[31] M. Panariello, S. Meyer, P. Champion, X. Miao, M. Todisco, N. T. Vu, and N. Evans, "The Risks and Detection of Overestimated Privacy Protection in Voice Anonymisation," in *5th Symposium on Security and Privacy in Speech Communication*, 2025, pp. 8–12.