# Linkage-Based Adversarial Framework for Voice Privacy Evaluation

*Dāvis Šterns*[1,2], *Tom Bäckstrom*[1], *Catuscia Palamidessi*[2], *Natasha Fernandes*[3], *Konstantinos Drosos*[4]

[1]Aalto University, Finland
[2]Inria, France
[3]Macquarie University, Australia
[4]Nokia, Finland

## Extended Abstract

While some speech anonymization methods aim to provide built-in mathematical privacy guarantees (such as those based on the Information Bottleneck principle [1]), most techniques rely on empirical evaluation to assess their privacy leakage [2]. Conducting this empirical assessment requires simulating an adversarial attack. To summarize the results of these simulated attacks, the speech community has historically relied on the Equal Error Rate (EER), driven heavily by benchmarks like the VoicePrivacy Challenge [3]. However, as highlighted by Champion [4], EER is fundamentally an authentication metric; by relying on hard decision thresholds to evaluate average binary accept/reject errors, it fails to measure an attacker's actual probabilistic confidence in compromising a specific identity. Beyond the specific limitations of EER, empirical evaluations suffer from a broader conceptual issue. No empirically derived score represents a static privacy property of the anonymization tool itself. Instead, any such score merely describes the outcome of a specific event: the interaction between a defense mechanism, a dataset, and an attacker model. Consequently, blindly comparing these empirical scores across different studies can be misleading. Even under the same threat model assumptions, variations in datasets or specific attacker implementations can make their final results incomparable.

To address these methodological inconsistencies, this extended abstract introduces the Linkage-Based Speech Privacy Evaluation Framework, which systematizes the evaluation process by dividing it into two distinct phases. The first phase, Score Generation, formalizes the simulated attack strictly as a function of the defender, the dataset, and the attacker. While this phase does not eliminate inherent dataset mismatches, it forces researchers to explicitly state their evaluation context, thereby preventing false cross-study comparisons. This ensures that any subsequent privacy claims are transparent and can be directly mapped to legal requirements, such as the General Data Protection Regulation (GDPR) definitions of linkability and "singling out" [5]. The second phase, Score Interpretation, focuses entirely on the statistical analysis of the resulting similarity scores. By decoupling the metric from the attack simulation, this phase allows researchers to evaluate privacy using multiple metrics simultaneously, rather than relying solely on the EER.

Operationally, Phase 1 (Score Generation) executes the attack simulation by treating the attacker as a generic linkage function. This function takes an enrollment set and an anonymized trial utterance as inputs to compute a similarity score. The configuration of this attacker defines the empirical threat model (e.g., prior knowledge and capabilities), while the nature of its inputs dictates the alignment with GDPR definitions. For instance, comparing clear enrollment speech to anonymized trial speech evaluates cross-dataset linkability (e.g., linking an anonymized voice to a public social media account). Conversely, comparing solely anonymized speech measures internal linkability and helps to estimate the risk of "singling out" all utterances by the same person within an anonymized dataset. Crucially, the outcome of this simulation is uniquely defined by the specific interaction between the defender, the dataset, and the attacker. By exhaustively comparing the utterances, this phase always concludes by generating two distinct sets of raw data: *mated scores* (similarity scores between utterances belonging to the same speaker identity) and *non-mated scores* (scores between different speaker identities).

Phase 2 (Score Interpretation) utilizes the raw mated and non-mated score sets generated in Phase 1. At this stage, empirical privacy evaluation reduces entirely to analyzing the statistical separability of these two sets: if the anonymization succeeds, the mated and non-mated scores will largely overlap; if it fails, they will separate, allowing the attacker to confidently re-identify users. To the best of our knowledge, all existing linkage-based privacy metrics operate exclusively on this two score set principle. This reduction makes the interpretation phase completely independent of the specific security assumptions or algorithms used in the attack simulation. Consequently, researchers can seamlessly apply traditional biometrics metrics like EER, ROC, or DET alongside more rigorous, threshold-independent privacy metrics (such as the Log-Likelihood-Ratio Cost ($C_{\text{llr}}$) [6], ZEBRA [7], Linkability measure ($D_{\leftrightarrow}$) [8], Similarity Rank Disclosure [9], or legally validated privacy metrics [10]) to achieve the multi-faceted privacy evaluation previously advocated by comparative studies [11]. Furthermore, this strict decoupling ensures that newer, more robust privacy measures can be computed retroactively on the published score sets of older anonymization systems, without requiring access to the original audio data or attacker models.

By systematizing the knowledge surrounding voice anonymization privacy evaluation, this framework acts as a scaffolding that allows researchers to interchange attackers, defenders, datasets, and metrics while maintaining a common reporting language. An empirical privacy score should not be treated as a universal, static property of the anonymization system; rather, it is a contextual snapshot that merely indicates how a specific method performed against a defined attacker on a given dataset. High error rates are often mistakenly attributed to strong anonymization when they actually stem from a dataset mismatch or an under-trained attacker model [12]. By adopting this explicit, two-phase framework, the speech processing community can transition from publishing arbitrary, metric-bound numbers to making verifiable, standardized, and contextually sound privacy claims.

**Index Terms**: speech anonymization, empirical privacy evaluation, privacy metrics, GDPR compliance

## Acknowledgments

This work was funded by the European Union's Horizon Europe research and innovation programme grant No 101168193.

## 1. References

[1] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *Proc. IEEE Inf. Theory Workshop (ITW)*, 2015, pp. 1–5.

[2] T. Bäckström, "Privacy in Speech Technology," *Proc. IEEE*, vol. 113, no. 7, pp. 668–692, 2025.

[3] 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 *Proc. INTERSPEECH*, 2020, pp. 1693–1697.

[4] P. Champion, "Anonymizing Speech: Evaluating and Designing Speaker Anonymization Techniques," Ph.D. dissertation, Université de Lorraine, 2023.

[5] Article 29 Data Protection Working Party, "Opinion 05/2014 on anonymisation techniques," European Commission, Tech. Rep. 0829/14/EN WP216, 2014.

[6] N. Brümmer and J. Du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

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

[8] M. Gomez-Barrero, J. Galbally, C. Rathgeb, and C. Busch, "General Framework to Evaluate Unlinkability in Biometric Template Protection Systems," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 13, no. 6, pp. 1406–1420, 2018.

[9] 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.

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

[11] M. Maouche, B. M. L. Srivastava, N. Vauquier, A. Bellet, M. Tommasi, and E. Vincent, "A Comparative Study of Speech Anonymization Metrics," in *Proc. INTERSPEECH*, 2020, pp. 1708–1712.

[12] 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 *Proc. Symp. Secur. Privacy Speech Commun. (SPSC)*, 2025, pp. 8–12.