Studying Voice Privacy Risks with Side Information through Partially Synthetic Data

Eulalie Thiombiano[1,2]*, Martha Larson*[1]*, Vincent Colotte*[2]*, Emmanuel Vincent*[2]

[1]iCIS, Radboud University, Nijmegen, Netherlands
[2]Université de Lorraine, CNRS, Inria, LORIA, Nancy, France

(eulalie.thiombiano, martha.larson)@ru.nl, vincent.colotte@loria.fr, emmanuel.vincent@inria.fr

1. Voice Data with Side Information

Privacy evaluations of speech data are typically conducted in a setting where only audio signals and a limited set of metadata are available. In practice, public speech corpora such as Common Voice, LibriSpeech, and VoxCeleb provide audio recordings, transcripts, and speaker identifiers, sometimes complemented by limited attributes such as gender, age, or accent. These datasets have become standard benchmarks in the speech community and are primarily used for tasks such as ASR [1], TTS, or speaker recognition [2], rather than privacy risk assessment. This results in a controlled or simplified setting, where demographic, geographic, social, or contextual information about speakers is often omitted to protect participants.

However, this simplified setting does not capture real-world conditions. In practice, an attacker may have access to additional side information about individuals obtained from external sources, such as occupation, place of residence, or contextual clues contained in spoken content. Such information can be combined with the voice signal to facilitate identification. Similar side-information attacks have been documented in other data domains, where external background knowledge can enable re-identification in anonymized demographic or behavioral datasets [3, 4]. Recent work on large-scale online deanonymization with LLMs shows that auxiliary information extracted from textual and online sources can substantially reduce anonymity [5]. This is consistent with GDPR principles, which state that identifiability should be assessed by taking into account all means reasonably likely to be used [6]. As a result, privacy evaluation using existing speech datasets remains incomplete when such side information is not considered. This limitation cannot be addressed simply by collecting richer datasets: speech data combined with detailed side information is highly sensitive, making it difficult to collect and even harder to share publicly in compliance with privacy regulations. Therefore, realistic privacy risk assessment requires alternative approaches to incorporate such information without relying on its direct availability.

2. Data Synthesis to Study Privacy

In the absence of such information, synthetic data has emerged as a potential solution. Synthetic data can be used in two ways. The first, which is the most widely studied, generates datasets that reproduce the statistical properties of real data to enable data sharing [7] or model training without exposing original records. The second, which is the focus of this work, constructs controlled scenarios that cannot be obtained from available real data. The literature distinguishes fully synthetic data, where all variables are generated, and partially synthetic data, where only a subset is replaced while the rest remains real [7]. Fully synthetic data would require generating both speech signals and attributes, which may alter data properties and confound privacy evaluation. In contrast, partially synthetic data preserves real speech while introducing controlled synthetic attributes, allowing the impact of side information on privacy risk to be studied. A concrete implementation would assign synthetic side information to real speech samples through statistical matching with external population-level data, such as census data, conditioned on available corpus metadata. These attributes would represent controlled attacker-side knowledge used to filter or re-rank candidate identities.

Most existing work follows the first approach, where synthetic data is used as a privacy-preserving mechanism [8], often relying on generative models [9] and, in some cases, differential privacy [10]. This line of work examines whether synthetic data can provide privacy guarantees. However, preserving statistical utility may also retain information that can be exploited to infer sensitive attributes or perform linkage attacks [11]. Moreover, generative models may reproduce structural properties of the training distribution, creating overlap between synthetic and real data [12].

Beyond this use, synthetic data can serve as an experimental tool, although this perspective remains under-explored. In this setting, synthetic attributes are introduced to study privacy risks or evaluate protection mechanisms. For example, generated data can support privacy auditing of machine learning models by simulating non-member examples in membership inference attacks [13, 14]. In this work, we adopt this perspective and focus on partially synthetic data, where real speech recordings are combined with synthetic attributes representing side information available to an attacker. Rather than aiming at privacy protection, we use synthetic data to construct evaluation scenarios that incorporate realistic side information and better reflect real-world attacks. However, the use of partially synthetic attributes raises methodological challenges. If synthetic attributes do not reflect realistic distributions or correlations with speaker identity, the measured privacy risks may be biased.

3. Outlook

This work opens several directions. Our study relies on partially synthetic data to model side information, raising methodological challenges regarding how synthetic attributes are designed and used. Incorrect assumptions about their distribution or relationship with speaker identity may bias privacy risk evaluation and lead to misleading conclusions. Future work will focus on principled approaches for constructing and validating partially synthetic data, to ensure privacy assessment in realistic scenarios involving audio signals, speech content, and metadata.

4. References

[1] 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 *Twelfth Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.

[2] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Interspeech*, 2017, pp. 2616–2620.

[3] L. Sweeney, "Simple demographics often identify people uniquely," *Health (San Francisco)*, vol. 671, no. 2000, pp. 1–34, 2000.

[4] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *2008 IEEE Symposium on Security and Privacy*, 2008, pp. 111–125.

[5] S. Lermen, D. Paleka, J. Swanson, M. Aerni, N. Carlini, and F. Tramèr, "Large-scale online deanonymization with llms," *arXiv preprint arXiv:2602.16800*, 2026.

[6] Article 29 Data Protection Working Party, "Opinion 05/2014 on anonymisation techniques," European Commission, Tech. Rep. WP 216, 2014. [Online]. Available: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

[7] J. Hu and C. M. Bowen, "Advancing microdata privacy protection: A review of synthetic data methods," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 16, no. 1, p. e1636, 2024.

[8] J. Jordon, L. Szpruch, F. Houssiau, M. Bottarelli, G. Cherubin, C. Maple, S. N. Cohen, and A. Weller, "Synthetic data–what, why and how?" *arXiv preprint arXiv:2205.03257*, 2022.

[9] P. Eigenschink, T. Reutterer, S. Vamosi, R. Vamosi, C. Sun, and K. Kalcher, "Deep generative models for synthetic data: A survey," *IEEE access*, vol. 11, pp. 47 304–47 320, 2023.

[10] D. Chen, R. Kerkouche, and M. Fritz, "A unified view of differentially private deep generative modeling," *arXiv preprint arXiv:2309.15696*, 2023.

[11] T. Stadler, B. Oprisanu, and C. Troncoso, "Synthetic data–anonymisation groundhog day," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 1451–1468.

[12] S. Mustaquim, A. Kotal, and H. Y. Paul, "When privacy isn't synthetic: Hidden data leakage in generative AI models," in *2025 IEEE International Conference on Big Data (BigData)*, 2025, pp. 4305–4314.

[13] M. Kazmi, H. Lautraite, A. Akbari, Q. Tang, M. Soroco, T. Wang, S. Gambs, and M. Lécuyer, "Panoramia: Privacy auditing of machine learning models without retraining," *Advances in Neural Information Processing Systems*, vol. 37, pp. 57 262–57 300, 2024.

[14] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017, pp. 3–18.