Voice Privacy from an Attribute-based Perspective

Mehtab Ur Rahman [1], *Martha Larson* [1,2], *Cristian Tejedor García* [1]

[1]Centre for Language Studies
[2]Institute for Computing and Information Sciences
Radboud University, Nijmegen, Netherlands

(mehtab.rahman, martha.larson, cristian.tejedorgarcia)@ru.nl

Abstract

Voice privacy approaches that preserve the anonymity of speakers modify speech in an attempt to break the link with the true identity of the speaker. Current benchmarks measure speaker protection based on signal-to-signal comparisons. In this paper, we introduce an attribute-based perspective, where we measure privacy protection in terms of comparisons between sets of speaker attributes. First, we analyze privacy impact by calculating speaker uniqueness for ground truth attributes, attributes inferred on the original speech, and attributes inferred on speech protected with standard anonymization. Next, we examine a threat scenario involving only a single utterance per speaker and calculate attack error rates. Overall, we observe that inferred attributes still present a risk despite attribute inference errors. Our research points to the importance of considering both attribute-related threats and protection mechanisms in future voice privacy research.

Index Terms: Voice privacy, Speaker attributes, Attribute inference, Speech data

1. Introduction

Speech is a rich signal that conveys far more than linguistic content. It carries a wide range of information about the speaker including identity, physical state, health, emotional state and demographic attributes. Such information is useful for many applications [1, 2, 3], but also constitutes sensitive personal information, leading to growing interest in voice privacy [4, 5]. Until now, voice privacy research has focused on a signal-based perspective. In other words, privacy risk is typically assessed by matching the speech signals of speakers. A key example is the Voice Privacy Challenge (VPC) [6, 7], the leading benchmark in voice privacy, other examples include [8, 9].

Such research overlooks the attribute-based perspective. Specifically, it does not consider matching between speakers represented by profiles of categorical attributes. Such profiles can be created by classifiers trained to infer the values of attributes from the speech signal. Attribute inference is possible on anonymized speech; for example, the anonymization approaches in the VPC 2024 [7], are designed to preserve speaker emotion. In general, any attribute not protected by anonymization, either intentionally or unintentionally, can be used to build a speaker attribute profile.

In this paper, we investigate the privacy of speaker attribute profiles and we point out that the attribute-based perspective on voice privacy should be considered alongside the conventional signal-oriented perspective. The importance of the attribute-perspective on speech privacy is motivated by prior work from the wider area of data protection that has shown that categorical data can still uniquely describe individuals, even in large datasets [10, 11, 12, 13]. In [14], the importance of speaker attributes is recognized for speaker verification as a complement to the speech signal. However, to our knowledge, we are the first to carry out a study of privacy protection on speaker attribute profiles and to provide a demonstration that such profiles still present a privacy risk despite attribute inference errors.

Our paper makes three main contributions: (1) We analyze the privacy risk of speaker attribute profiles at the speaker and at the utterance level. Our analysis studies speaker uniqueness for profiles inferred from original (unprotected) speech as well as from anonymized speech. (2) We carry out a re-identification attack in which the attacker matches the speaker attribute profile of the target speaker that has been inferred from a single utterance (original and anonymized) to speaker attribute profiles inferred from multiple utterances of a known speaker. (3) We release a set of annotations for four speaker attributes (gender, age, accent, and profession), aggregating and extending annotations previously released for the VoxCeleb2 data set and enabling experimentation with attribute-based speaker profiles.[1]

2. Related Work and Background

2.1. Uniqueness Analysis for Attribute-based Profiles

An attribute-based perspective of privacy has a long history in statistical disclosure control (SDC), which studies person-level data (microdata), with a specific focus on categorical data in table form [15, 16]. A central idea is that shared attribute values partition the dataset into equivalence classes, and privacy decreases when these classes are small. This intuition is formalized by k-anonymity, where the anonymity set size k is the number of records that share an attribute profile and uniqueness corresponds to $k = 1$ [11].

The importance of uniqueness analysis is supported by regulatory and evaluation perspectives. Guidance related to the GDPR identifies singling out as a practical privacy risk alongside linkability and inference [17, 18]. Although the GDPR does not formally operationalize singling out, uniqueness of attribute profiles provides a measurable interpretation of this concept. Recent work has proposed legally grounded metrics for singling out and linkability in voice anonymization evaluation and has shown that predicate singling out risk may not correlate with speaker verification metrics such as Equal Error Rate (EER) [19]. These considerations motivate us to use uniqueness and anonymity set sizes to assess privacy.

2.2. Attacker Specification for the Re-identification Attack

The Scenario of Use Scheme [20] provides a set of dimensions for the explicit specification of privacy threat situations. Based

[1]Annotations and code will be made publicly available.

on this scheme, the attacker scenario that forms the basis of our study of re-identification attack is specified in Table 1.

Table 1: *Attacker scenario for the re-identification attack*

Objective	The attacker aims to recover the real-world identity for a set of target (i.e., test) speakers.
Opportunity	The attacker has one spoken audio utterance for each target speaker. We study two cases: the audio is original (unprotected) and the audio is protected by an anonymization algorithm. The attacker has multiple spoken utterances for a group of reference speakers that are labeled with the speaker ID.
Additional Resources	The attacker uses the same attribute classifiers as in the uniqueness analysis, i.e., the attacker has access to original (unprotected) speech data drawn from the same distribution as the target data that has been labeled with the four attribute classes.

The specification of the attacker scenario is based on the most recent (i.e., 2024) Voice Privacy Challenge (VPC) [7], which studies privacy at the utterance level. However, we choose to study an attack scenario more challenging than what is studied in the VPC 2024. Specifically, for the cases involving anonymized test data, we assume that the attacker does not have access to the anonymization system and cannot anonymize the training data.

Like the VPC, our attack involves matching test speakers with reference speakers for whom the identity is known. A key difference is that the VPC studies signal-to-signal comparisons, which allow degrees of match, and our work studies comparisons between speaker attribute profiles in terms of exact matches or mismatches. Because we do not have degrees of match, we cannot adjust matching thresholds as needed for the Equal Error Rate used by the VPC. Instead, we calculate an Error Rate (cf. Section 3.3).

3. Experimental Setup

3.1. Dataset

Our research requires speech data that has been annotated with multiple speaker attributes, and we choose to build on VoxCeleb2 [21], which contains speech of celebrities collected from YouTube. Gender labels are directly available in the VoxCeleb2 metadata. We computed age labels from date of birth (from Wikipedia) and the recording year (from YouTube) following [22]. Accent is approximated using nationality labels scraped from Wikipedia following [14]. Further we generated profession labels by using information from Wikipedia and mapped to six categories adapted from [23]. Details on how we aggregated and extended existing attributes are in the supplementary material.

We experiment with two evaluation datasets: one with multiple utterances per speaker (*MultiEval*), with a total of 24,588 utterances, and one with a single utterance per speaker analysis (*SingleEval*), which we re-sample 10 times to control for sampling variance. These sets contains 72 of 118 VoxCeleb2 test speakers, which are the speakers for which all four attributes are available. Recall from the Section 1 that previous work on categorical data has shown that attribute profiles can isolate people in large scale data [10, 11, 12, 13]. For this reason, 72 speakers are enough for our purpose here and to our knowledge constitute currently the largest publicly available set of speakers fully annotated with at least four attributes. We train our attribute classifiers on the official VoxCeleb2 dev set and report classification results on the test set. In addition to the original speech, we evaluate anonymized speech on *SingleEval*, using the VPC 2024 baseline systems, McAdams (B2) [24], STTTS (B3) [25], NAC (B4) [26] and ASRBN (B5) [27]. Table 2 summarizes the statistics of the datasets used in our experiments.

Table 2: *Dataset statistics. Attributes used: Gender (2 levels), Age (3 levels), Accent (29 levels) and Profession (6 levels).*

Dataset	Description	#Speakers	#Utterances
Classifiers train/dev set	VoxCeleb2 Dev	5,994	1,092,009
Classifiers evaluation set	VoxCeleb2 Test	118	36,237
MultiEval	Speaker level analysis; Derived from VoxCeleb2 Test set; mean of 341.5 utterances/speaker; ground truth, original	72	24,588
SingleEval	Utterance level analysis; Derived from VoxCeleb2 Test set; 10 re-samplings for each ground truth, original, 4 anonymized versions	72	72×10

3.2. Method for Analyzing Speaker Uniqueness

For our speaker uniqueness analysis, we calculate the uniqueness of speakers in our speaker set on the basis of their attribute profiles consisting of the four attributes described above: gender, age, accent, and profession. We report uniqueness in terms of k, the number of speakers in the speaker set to which a given speaker is identical. We compute the percentage of speakers in the speaker set who are unique ($k = 1$) and the percentage of speakers that are below a target threshold ($k < 5$), which we consider in this work as an acceptable level of k. We also report the percentage of speakers at two other thresholds ($k < 3$ and $k < 10$) to give a more complete picture. Finally, we report the median k over all speakers in the speaker set.

The analysis is carried out on two types of speaker profiles: first the MultiEval profiles, consisting of attributes that were inferred over multiple utterances (speaker-level profiles) and SingleEval profiles in which the attributes were inferred over a single utterance (utterance-level profiles). The comparison reflects how privacy risk varies with the amount of speech data available. Recall that we test 10 re-sampled SingleEval sets so conclusions are not set dependent. Note that for the condition in which we use ground truth speaker attributes rather than inferred attributes, the speaker-level profiles and the utterance-level profiles are the same.

3.3. Method for Attribute-based Re-identification Attack

Next, we move our study of attribute-based profiles one step closer to a real-world threat scenario. Specifically, we measure the success of a re-identification attack with the SingleEval as the target (test) data following the threat model specification in Table 1. In the first case, the test audio is original speech (unprotected) and in the second case the test audio is protected with the four VPC 2024 [7] anonymization systems, specified in Section 3.1. The reference data is multiple utterances per speaker from the MultiEval set, each associated with the speaker ID. The attacker uses the attribute classifiers to infer speaker attribute profiles for all test data and also for all speakers in the reference data.

The attack is performed by matching the attribute profile of each test speaker to the attribute profiles of all reference speakers. For a given test speaker, if there is only a single reference speaker that matches, then the predicted identity of the test speaker is the identity of the matching reference speaker. If there are multiple reference speakers that match, then the attacker makes a random selection among all matching reference speakers and the predicted identity of the test speaker is the

identify of the selected reference speaker. We report the results of the attack in terms of the error rate, defined as the proportion of speakers that the attacker cannot identify.

We briefly discuss the similarities and differences between this error rate and the Equal Error Rate (EER) used in the VPC. As previously mentioned, the EER requires the ability to control the threshold of the matching decision. However, in the case of attribute-based profiles, we have a binary distinction between exact match and no match and there is no threshold to adjust. The error rate is influenced by the process of random draw, but this contribution remains constant across all conditions. We report the average error rate over the 10 re-samplings of SingleEval. Note that an attacker using partial matches would likely achieve higher attack success rates, and is relevant for future study of defenses against attribute-based attacks.

3.4. Setup for Attribute Inference

To train the four attribute classifiers (gender, age, accent, profession), we first extract speaker embeddings using the pretrained ECAPA TDNN model [28, 29], a well-established model for learning effective speaker representations. For each utterance, the encoder produces a 192 dimensional embedding that is used as input to lightweight attribute classifiers. Before classification, embeddings are normalized to unit length to reduce scale variation across utterances and improve stability. All attribute classifiers are multilayer perceptrons operating on the 192 dimensional normalized embedding.

The gender classifier is a single hidden layer MLP with ReLU activation and a single output logit for binary classification. Age, accent, and profession are classifiers with two hidden layer MLPs using LeakyReLU activations and a final linear layer producing logits for the target number of classes.

All attribute classifiers are trained on the VoxCeleb2 development set, with 10% of speakers per class held out for hyperparameter tuning. Hyperparameters are selected through grid search, and the configuration that achieves the best validation performance is retained. After selecting the best configuration, we retrain the final model on the full development set using the chosen hyperparameters. At inference time, attributes are predicted independently for each utterance by applying the trained classifiers, and selecting the class with the highest posterior probability. For speaker level inference, posterior probabilities from all utterances of a speaker are averaged, and the final attribute label is determined by the highest mean posterior. Note that we train only one set of classifiers on the original (unprotected) audio, meaning attack success will be more surprising and informative. However, future work on defenses should also study an attacker who has access to the anonymization system used to protect the speech data.

4. Results of Analysis and Attack

4.1. Analysis of Speaker-Level Attribute Inference

First, we consider the speaker-level case, in which multiple utterances are available for each speaker. Table 3 reports attribute inference. In column 'Speaker level' it can be seen that binary gender is inferred nearly perfectly. Performance is substantially worse for other attributes, although still remains above the 'Weighted baseline', which is class-weighted random classifier.

Table 4 summarizes the results of the evaluation at speaker level. The percentage of unique speakers ($k = 1$) is less when speakers are represented by inferred attributes (31.9%) rather than ground truth attributes (38.9%) and the median k rises by one (from 2 to 3), reflecting a modest improvement in privacy. However, considering $k = 5$, which we take to be our threshold of acceptable protection, the percentage of speakers with $k < 5$ rises, indicating that there are actually less speakers with an acceptably large group of identical speakers when speakers are represented by inferred attributes as opposed to by ground truth attributes. Considering $k < 10$ the drop is even higher. We interpret these results to demonstrate that the noise introduced by inference leads to less very small groups of identical speakers (including less unique speakers), but also less larger groups of speakers. In short, we do not see the errors introduced by attribute inference making a straightforward contribution to improving speaker privacy, as measured by uniqueness.

Table 3: *Attribute classifier performance on original data.*

	Baselines	Utterance level		Speaker level	
Attribute	Weighted	Acc	F1	Acc	F1
Gender	0.55	0.99	0.99	0.99	0.99
Age	0.71	0.76	0.79	0.83	0.85
Accent	0.39	0.64	0.66	0.75	0.73
Profession	0.42	0.53	0.54	0.60	0.57

Table 4: *Speaker uniqueness with respect to speaker-level (MultiEval) attribute profiles inferred on original audio data and with respect to ground truth attribute profiles.*

	Inferred	Ground truth
Unique speakers ($k = 1$) [%]	31.9	38.9
$k < 3$ [%]	43.1	55.6
$k < 5$ [%]	68.1	65.3
$k < 10$ [%]	80.6	72.2
Median k	3	2

4.2. Analysis of Utterance-Level Attribute Inference

Next, we consider the utterance-level case in which a single utterance is available for each speaker. Comparing the 'Speaker level' and the 'Utterance level' columns of Table 3, we observe that inference performance is indeed worse when only one utterances used for prediction across all attributes, except for gender, which is still predicted nearly perfectly.

Figure 1 reports the uniqueness of the speaker set, with the individual re-samplings of SingleEval represented as separate data points in the plot. The black Xs indicate uniqueness percentages with speaker profiles consisting of ground truth attributes and the leftmost (blue) stacks of data points are the uniqueness percentages with speaker profiles consisting of inferred ground truth attributes. Considering $k = 1$, the percentage of unique speakers in the case that speakers are represented with inferred attributes from the original data (Original) has a wide spread for the different resamplings. However, on average, its percentage is the same as with the ground truth. The same holds for $k < 3$. For $k < 5$, we again see that inference is increasing the percentage of speakers who fall under the acceptable threshold of $k = 5$ and the same happens with $k < 10$. In short, even with only a single utterance available for inference, classifier errors are not contributing to speaker privacy in a consistent manner.

To gain deeper understanding, we consider what changes at the level of individual speakers. Figure 2 shows that the change in uniqueness does not impact all speakers equally. The bars

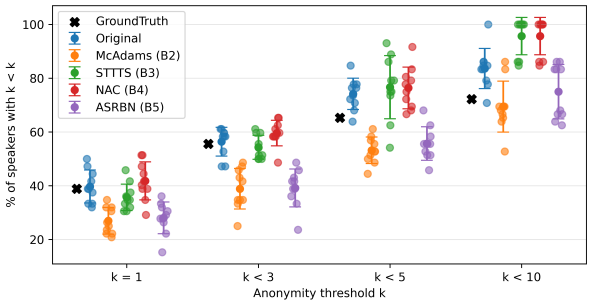

Figure 1: *Speaker singling out under original and anonymized single utterance per speaker setting. Each dot represents the percentage of speakers whose anonymity set size satisfies the given threshold in one independent run.*

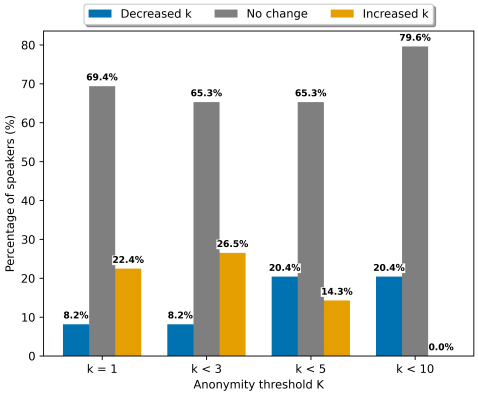

Figure 2: *Impact of attribute inference on the uniqueness across speakers. For different anonymity thresholds k, the figure shows the proportion of speakers whose anonymity set increased in size (reduced risk), decreased in size (higher risk), or remained unchanged relative to ground truth labels.*

in this graph show the percentage of speakers at each k that becomes lower (left), meaning their situation is worse with respect to their k-level, stays the same (middle), and becomes higher (right side), meaning their privacy situation has become better. At the $k < 10$ level, inference makes the situation worse for 20.4% of the speakers and improves the situation for no one.

4.3. Analysis of Inference on Anonymized Utterances

We now turn to the case in which a single utterance is available per speaker, and that utterance has been anonymized. It is important to note that, in general, anonymization techniques are developed to protect against de-anonymization, but not specifically designed to also offer full protection against attribute inference. The accuracy levels of inference on anonymized speech are quite low, as can been seen in Table 5. Recall that the attribute inference classifier is trained on unprotected speech.

Table 5: *Attribute classification accuracy (mean ± std) for different anonymization methods*

Attribute	McAdams	STTTS	NAC	ASRBN
Gender	0.80 ± 0.03	0.47 ± 0.05	0.67 ± 0.04	0.53 ± 0.03
Age	0.56 ± 0.05	0.31 ± 0.05	0.65 ± 0.05	0.36 ± 0.07
Accent	0.52 ± 0.05	0.38 ± 0.05	0.38 ± 0.05	0.49 ± 0.04
Profession	0.57 ± 0.03	0.26 ± 0.06	0.38 ± 0.05	0.39 ± 0.05

In Figure 1, we see that there is no large difference between the cases in which attributes are inferred from original speech and from speech anonymized with NAC or STTS. McAdams and ASRBN provide protection by lowering the number of speakers with $k < 5$, but none of the approaches are much different from the ground truth in the case $k < 10$. In short, analysis reveals that the error of attribute inference on anonymized speech is also not consistently contributing to improved privacy.

4.4. Re-Identification Attack

The re-identification attack compares target speaker profiles to reference profiles. When the ground truth versions are used the error rate is 0.47, compared to 0.53 when target and reference speakers are both represented with inferred profiles from multiple speaker (i.e., speaker level). Since for good protection the error rate should approach 1.0, these error rates are quite low.

The first line of Table 6 shows the performance on speaker attribute profiles that have been inferred from single original utterances. Surprisingly the error rate is lower when the reference profiles are inferred rather than ground truth. We attribute this difference to a correlation in the errors of the attribute classifiers: a wrong prediction does not protect privacy if it is consistently wrong. The remainder of Table 6 shows that inference on anonymized data sometimes leads to error rates in the same range as inference on original data. In general, in all cases, the error rate falls short of 1.0. Note that if the classifiers were to predict the majority class level for each attribute all speakers would be identical and the error rate would be $1/72 = 0.014$.

Table 6: *Attack error rate. Two types of speaker profiles used as reference. Inferred reference profiles are at speaker level.*

Target speaker profile inferred on	Reference: ground truth profile	Reference: profile inferred from original
Original (unprotected)	0.72 ± 0.04	0.67 ± 0.03
McAdams	0.76 ± 0.04	0.78 ± 0.04
STTTS	0.62 ± 0.05	0.82 ± 0.04
NAC	0.70 ± 0.06	0.71 ± 0.05
ASRBN	0.58 ± 0.05	0.82 ± 0.04

5. Conclusion and Outlook

We have taken a first look at voice privacy from an attribute-based perspective, studying speaker attribute profiles comprised of attributes inferred from spoken audio. We have shown that although the performance of attribute inference classifiers may be low, misclassifications do not always provide additional privacy protection. Specifically, misclassifications cause some speakers to become more unique (Figure 2) and can lower the number of speakers who enjoy a sufficiently large number of identical speakers (here taken to be $k \geq 5$) (Table 4 and Figure 1). Further, the very low attribute classification performance on anonymized data does not always contribute to lowering the uniqueness of speakers, as shown by our uniqueness analysis (Figure 1). Also, it does not always contribute to substantially raising the speaker re-identification error as shown by our re-identification attack (Table 6).

Future work should examine the pattern of classifier mistakes, since attribute classifiers that are consistent in their inaccurate predictions of the attributes for a given speaker can raise uniqueness and lower attack error. Looking forward, our work has laid the groundwork for future study of privacy from an attribute-based perspective as well as the development of defenses against attribute-based attacks.

6. Acknowledgments

This work was supported by the NWO research programme AiNed Fellowship Grants under the project Responsible AI for Voice Diagnostics (RAIVD) - NGF.1607.22.013.

7. Generative AI Use Disclosure

Generative AI tools were used for language editing and polishing, including grammar and phrasing. All scientific content, experimental design, analyses, results, and conclusions were developed, verified, and approved by the authors. The authors take full responsibility for the content of this paper, and no generative AI tool is listed as a co-author.

8. References

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

[2] G. Y. Peshkova, O. Zlobina *et al.*, "Digital transformation of banking with speech technologies," *European Proceedings of Social and Behavioural Sciences*, 2020.

[3] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International Journal of Speech Technology*, vol. 15, no. 2, pp. 99–117, 2012.

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

[5] A. Nautsch, A. Jiménez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, M. A. Hmani, A. Mtibaa *et al.*, "Preserving privacy in speaker and speech characterisation," *Computer Speech & Language*, vol. 58, pp. 441–480, 2019.

[6] 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.

[7] 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," *arXiv preprint arXiv:2404.02677*, 2024.

[8] X. Miao, R. Tao, C. Zeng, and X. Wang, "A benchmark for multi-speaker anonymization," *IEEE Transactions on Information Forensics and Security*, 2025.

[9] J. Yao, Q. Wang, P. Guo, Z. Ning, Y. Yang, Y. Pan, and L. Xie, "MUSA: Multi-lingual speaker anonymization via serial disentanglement," *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[10] Y.-A. De Montjoye, L. Radaelli, V. K. Singh, and A. S. Pentland, "Unique in the shopping mall: On the reidentifiability of credit card metadata," *Science*, vol. 347, no. 6221, pp. 536–539, 2015.

[11] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[12] P. Golle, "Revisiting the uniqueness of simple demographics in the us population," in *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society*, 2006, pp. 77–80.

[13] L. Rocher, J. M. Hendrickx, and Y.-A. De Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models," *Nature Communications*, vol. 10, no. 1, p. 3069, 2019.

[14] C. Luu, P. Bell, and S. Renals, "Leveraging speaker attribute information using multi task learning for speaker verification and diarization," in *Proceedings of Interspeech*, 2021, pp. 491–495.

[15] L. Willenborg and T. De Waal, *Elements of Statistical Disclosure Control*. Springer Science & Business Media, 2012, vol. 155.

[16] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, and P.-P. De Wolf, *Statistical Disclosure Control*. John Wiley & Sons, 2012.

[17] European Parliament, "Directive 95/46/EC General Data Protection Regulation," URL: http://data.europa.eu/eli/reg/2016/679.

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

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

[20] 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.

[21] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Interspeech 2018*, 2018, pp. 1086–1090.

[22] K. Hechmi, T. N. Trong, V. Hautamäki, and T. Kinnunen, "Voxceleb enrichment for age and gender recognition," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 687–693.

[23] Y. Wu, L. Chen, B. Elie, F. M. Suchanek, I. Vasilescu, and L. Lamel, "Who's speaking? predicting speaker profession from speech," in *International Congress of Phonetic Sciences 2023*. Guarant International, 2023, pp. 3086–3090.

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

[25] 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 *ICASSP - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

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

[27] P. Champion, "Anonymizing speech: Evaluating and designing speaker anonymization techniques," *arXiv preprint arXiv:2308.04455*, 2023.

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

[29] 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," 2021, arXiv:2106.04624.