A bridge between features and evidence for binary attribute-driven perfect privacy

Paul-Gauthier Noé, Andreas Nautsch, Driss Matrouf, Pierre-Michel Bousquet, Jean-François Bonastre

A BRIDGE BETWEEN FEATURES AND EVIDENCE FOR BINARY ATTRIBUTE-DRIVEN PERFECT PRIVACY

Paul-Gauthier Noé[1], *Andreas Nautsch*[2], *Driss Matrouf*[1], *Pierre-Michel Bousquet*[1], *Jean-François Bonastre*[1]

[1]Laboratoire Informatique d'Avignon, Avignon Université, France
[2]vitas.ai, Germany

ABSTRACT

Attribute-driven privacy aims to conceal a single user's attribute, contrary to anonymisation that tries to hide the full identity of the user in some data. When the attribute to protect from malicious inferences is binary, perfect privacy requires the log-likelihood-ratio to be zero resulting in no *strength-of-evidence*. This work presents an approach based on normalizing flow that maps a feature vector into a latent space where the *evidence*, related to the binary attribute, and an independent residual are disentangled. It can be seen as a non-linear discriminant analysis where the mapping is invertible allowing generation by mapping the latent variable back to the original space. This framework allows to manipulate the log-likelihood-ratio of the data and therefore allows to set it to zero for privacy. We show the applicability of the approach on an attribute-driven privacy task where the sex information is removed from speaker embeddings. Results on VoxCeleb2 dataset show the efficiency of the method that outperforms in terms of privacy and utility our previous experiments based on adversarial disentanglement.

***Index Terms*—** perfect privacy, forensic science, disentanglement, normalizing flow, speaker verification

1. INTRODUCTION

Attribute-driven privacy aims to make the data independent of an attribute a user wants to keep secret [1, 2]. In [1] we proposed a method based on adversarial disentanglement where the data is fed into an autoencoder and where there is an attribute's class classifier on the encoded representation. With the adversarial training, the encoder tries to fool the classifier and the latter tries to predict the right attribute's class. This reduces the discriminant information related to the attribute in the encoded representation. However, the way this approach protects the data may remain obscure and explainability may be jeopardised. In this paper, we explore a new vision of binary attribute-driven privacy inspired by probabilistic forensic science. Indeed, understanding how the attribute's information is embedded in the data helps in designing a protection system. Even if the term *evidence* is used sometimes to denote the observed data, we use evidence here to point what is useful for the decision task in the observation [3]. Also, we call *residual* anything in the observation that is independent of the variable (attribute) of interest. Therefore, our aim is to disentangle the evidence and the residual from the observation. The evidence is represented by a likelihood-ratio [4] between two mutually exclusive propositions which correspond here to the two classes of the attribute. For perfect privacy [5], the log-likelihood-ratio (LLR) must be zero to alleviate the *strength-of-evidence* of the data: this is zero-evidence. As far as we know, there is no literature in the speech community that deals with privacy with the terms of *perfect privacy* [6, 7] except [5] but it presents a privacy evaluation rather than a protection method. Even so, Linear Discriminant Analysis (LDA) [8] allows to linearly map the data into a space where the discriminant component lives in a logit posterior line on which the LLR can be identified. However, the Gaussian and shared covariance assumption of the LDA and the resulting linear mapping may result in a poor estimation of the LLR and is not flexible enough for real data. Therefore, we propose a non-linear discriminant analysis that maps the data into a latent space where the class-conditional densities are Gaussian with same covariance matrix. The first component is the only discriminant and is a LLR while the other components are the residual variabilities irrelevant for the attribute inference. Moreover, this mapping needs to be invertible in order to allow the manipulation of the evidence inside the latent space and the reconstruction into the original feature space[1]. In order to do so, the proposed method is based on Normalising Flow (NF) [11, 12]. This is a generative model that learns a bijective mapping between the feature and the latent space. Most of the time, it is used in an unsupervised framework where the latent distribution is a multivariate Gaussian distribution. Some works proposed to use NF in a supervised case [13, 14]. In our case, the two class-conditional densities are mapped into two multivariate Gaussian distributions where their parameters are constrained such that only the first latent dimension is discriminant and is a *calibrated* LLR [15, 16]. In this way, the LLR can be handled and set to zero for privacy. While, the proposed method can be extended to any binary attributes, this paper presents how it can be applied to remove the speaker's sex information into VoxCeleb2 [17] speaker embeddings. Indeed, to avoid being a victim of sexism or for any personal reason, users may want to hide their sex. Next section presents how Bayesian decision theory gives clue for perfect privacy. The third section explains the proposed evidence disentanglement system and the attribute protection strategy. Then, results of its application on the concealment of the sex attribute in speaker embeddings are given in section four.

2. BAYESIAN DECISION & PERFECT PRIVACY

Considering a set of classes $\mathcal{C} = \{c_0, c_1\}$ and an attacker who wants to infer the class of an observation x, the posterior probabilities are given by:

$$\text{logit}\, P(c_i|x) = \log \frac{P(x|c_i)}{P(x|c_{\neg i})} + \text{logit}\, P(c_i), \qquad (1)$$

where $i \in \{0, 1\}$ and $P(c_i)$ is the attacker's prior. *Perfect privacy* [5], also known as Shannon's *perfect secrecy* [18], is reached when the attacker's posterior knowledge remains its prior one. From

[1]Quadratic discriminant analysis is an example of a non-linear data transformation onto a logit posterior space. However, it is still based on a Gaussian assumptions and it does not allow a return into the feature space [9, 10].

Equation 1, one can see that this is achieved when the LLR is zero for all classes and observations. Therefore, in order to remove the information related to a binary attribute in some data, one would like to set the LLR to zero for all samples. This would remove any strength-of-evidence in the data making it attribute-independent [18]. In order to do that, we propose to map the observation's feature vector into a vector space where the first dimension is the LLR (i.e. the evidence) and the other dimensions contain remaining variability, called residual, which is independent of the attribute. Because this mapping is invertible, the evidence and residual can be mapped back into the original data's space. Next section presents a first solution that tends toward such mapping.

3. DISENTANGLING THE EVIDENCE FROM THE OBSERVATION

In this section, we formalise the disentanglement problem and propose a first solution. Let $\mathcal{X}$ be a n-dimensional feature space of some observed data. Assuming that each sample has an evidence expressed by a LLR $l \in \mathcal{L} \subset \mathbb{R}$ and has some residual $r = (r_1, \dots r_{n-1})^T$ in $\mathcal{R} \subset \mathbb{R}^{n-1}$, let's define a latent space $\mathcal{Z} = \mathcal{L} \oplus \mathcal{R}$. Our aim is to find, assuming it exists, an invertible mapping f between $\mathcal{X}$ and $\mathcal{Z}$ such that $z = f(x)$ and $x = f^{-1}(z)$ where $x \in \mathcal{X}$ and $z = (l, r_1, \dots r_{n-1})^T \in \mathcal{Z}$. In other words, f disentangles the observation into its evidence, and its residual i.e. everything that is independent of $\mathcal{C}$.

3.1. Class-conditional densities in the latent space

Due to the *idempotence* property of likelihood-ratio [3], if one of the class-conditional density of the LLR is Gaussian with mean μ, the other is also Gaussian with an opposite mean $-\mu$ and the same variance $\sigma^2 = 2\mu$ [16, 19]. For the first component of z to be a LLR, the class-conditional densities in the latent space must respect this property. Therefore, let's define them as follow:

$$z|c_0 \sim \mathcal{N}(\mu e_1, \Sigma), \tag{2}$$
$$z|c_1 \sim \mathcal{N}(-\mu e_1, \Sigma), \tag{3}$$

where $\mu \in \mathbb{R}^+$, $e_1 = (1, 0, \dots 0)^T$ and $\Sigma = \mathrm{diag}(2\mu, 1, \dots 1)$. It can be shown that in this way, the LLR is given by the first component $z_0 = e_1^T z$ and the idempotence property is respected [16].

3.2. Invertible mapping between features space and latent space

Normalizing flow allows latent variable inference and data generation by using an invertible mapping g, between the data space and the latent space, learned by data likelihood maximisation.
Let $\mathcal{D} = \{(x^{(0)}, c^{(0)}), \dots (x^{(N-1)}, c^{(N-1)})\}$ be a set of observed samples $x \in \mathcal{X}$ with the labels $c \in \mathcal{C}$. The log-likelihood of the observed data is:

$$\log p_{X|\theta_g,\mu}(\mathcal{D}) = \sum_{i=0}^{1} \left(\sum_{(x,c)\in\mathcal{D}|c=c_i} \log p_{X|c_i,\theta_g,\mu}(x) \right), \tag{4}$$

where θ_g are the parameters of g and because the mapping is bijective and is applied equally on samples coming from both c_0 and c_1, the change of variable formula gives:

$$p_{X|c_i,\theta_g,\mu}(x) = p_{Z|c_i,\mu}(z) \left| \det \left(\frac{\partial g(z)}{\partial z} \right) \right|^{-1}, \ \forall i \in \{0, 1\}, \tag{5}$$

where $x = g(z)$. Usually, g is a neural network that can be easily inverted and where the Jacobian determinant (in the formula of variable substitution) can be quickly computed for optimisation [11, 12].

3.3. Latent distributions parameter and optimisation

The only parameter of the latent distribution is μ. This parameter appears only on the first component of the latent variable. A straightforward computation shows that, given a batch $\mathcal{B}_Z$ of samples in the latent space, a maximum likelihood estimator of μ is given by:

$$\hat{\mu}_{\mathrm{MLE}}(\mathcal{B}_Z) = -1 + \sqrt{1 + \frac{1}{|\mathcal{B}_Z|} \sum_{z\in\mathcal{B}_Z} (e_1^T z)^2}, \tag{6}$$

where $e_1^T z$ is the first component of the vector z. Like the expectation maximisation algorithm, we propose to do a two-stage iterative optimisation. We summarise this optimisation strategy in the following where α is the adaptation parameter for the update of μ:

Choose $\alpha \in [0, 1]$,
Initialise θ_g and μ,
for *all batches* $\mathcal{B}_X$ **do**
  $\mathcal{B}_Z \leftarrow g^{-1}(\mathcal{B}_X)$
  $\theta_g \leftarrow \underset{\theta_g}{\mathrm{argmax}} \log p_{X|\theta_g,\mu}(\mathcal{B}_X)$
  $\mu \leftarrow \alpha\mu + (1 - \alpha)\hat{\mu}_{\mathrm{MLE}}(\mathcal{B}_Z)$.
end

3.4. Strength-of-evidence manipulation for privacy

This disentanglement of the LLR in an observation from the irrelevant information allows the manipulation of the data *strength-of-evidence*. Changing the first component of z permits to change the strength-of-evidence of x. Indeed both LLR are equal because the same bijective mapping is applied on samples from both classes:

$$\frac{p_{Z|c_0}(z)}{p_{Z|c_1}(z)} = \frac{p_{X|c_0}(x) \left| \det \left(\frac{\partial g(z)}{\partial z} \right) \right|}{p_{X|c_1}(x) \left| \det \left(\frac{\partial g(z)}{\partial z} \right) \right|} = \frac{p_{X|c_0}(x)}{p_{X|c_1}(x)}. \tag{7}$$

This can be used in order to make zero-evidence for perfect privacy. Indeed, the LLR can be set to zero following these steps: (1) Map features vector x into the latent space using $z = g^{-1}(x)$, (2) Set the first dimension of z to zero, (3) Map back into the original features space using $x_{\mathrm{llr}_0} = g(z_{\mathrm{llr}_0})$ where $z_{\mathrm{llr}_0} = (0, r_1, \dots r_{n-1})^T$. By setting the first dimension of the latent variable to zero, this approach allows to alleviate the strength-of-evidence of the data. In the next section, we show a practical application of this approach on attribute-driven privacy.

4. APPLICATION: ATTRIBUTE-DRIVEN PRIVACY

From now on, we assume that after optimisation, g^{-1} get sufficiently close to the ideal mapping f. We will therefore no longer distinguish them. In this section, we present how the proposed approach can be applied to binary attribute-driven privacy. As a toy example, we want to remove the speaker's sex information from embeddings used in speaker verification. As in [1], the user may want to use authentication-by-voice while not disclosing its sex to the service provider. Therefore, before releasing to the remote service the extracted speaker embedding, a sex protection system can be applied on it. We use here kaldi's x-vector [20] as speaker embedding.

4.1. Experiment

In our experiment, we use the same training and testing sets as in [1]. V2D and V2T are respectively subsets of VoxCeleb2 [17] development and test part on which we respectively train and test our protection system. However, for the protection assessment in [1], the $C_{\text{llr}}^{\text{min}}$ [15] and the ZEBRA's metrics [5, 21] were computed from scores obtained by a state-of-the-art sex classifier trained on non-protected data. With hindsight, we found that this way of assessing the protection can be misleading. Indeed, it only informs if we can fool this classifier and not if the attribute information remains in the protected embeddings. Therefore, these metrics will be here computed on scores obtained by a classifier trained on protected data. To do so, the set V2T (that was not used to train the protection system), is split into a training V2T-train and testing V2T-test part. 46 male and 25 female speakers are randomly chosen to build V2T-train and the remaining speakers i.e. 35 males and 14 females form V2T-test. These two sets contain respectively 17735 and 13944 utterance x-vectors. The classifier will be thus trained and tested respectively on V2T-train and V2T-test. We also provide as in [1] a mutual information (MI) measure between the embedding's dimensions and the sex class variable [22, 23].

Proposed system:
The protection system[2] we propose is based on the disentanglement and the protection strategy presented in Section 3. Here, the NF architecture is the Real NVP [12] with 6 stacked coupling layers where the scale and translation functions are multilayer perceptrons. The scale function is made of 3 linear layers with 2 LeakyReLU activations and an output Tanh activation. The translation function is made of 3 linear layers with also 2 LeakyReLU activations but no output activation. Adam algorithm with a 10^{-4} learning rate is used for optimisation. All μ initialisations we tried converge to a value close to 9, except for the initialisations really close to zero. The results reported here are those for a μ initialised at 10 with $\alpha = 0.99$.

Baselines:
We compare the proposed approach to two baselines. The first one is based on LDA and can be expressed with the following whitening:

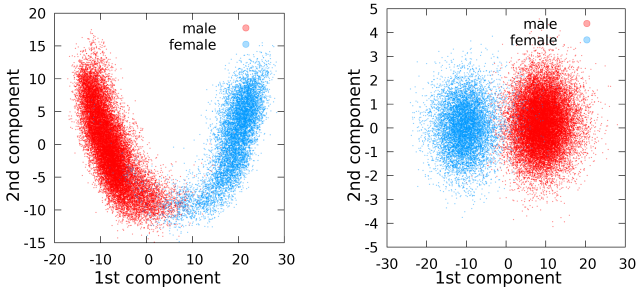
$$x \leftarrow x - \frac{ww^T}{||w||^2}x + \frac{1}{2||w||^2}(\mu_F^T\Sigma_W^{-1}\mu_F - \mu_M^T\Sigma_W^{-1}\mu_M)w, \quad (8)$$

where the class-conditional densities in $\mathcal{X}$ are assumed to be multivariate normal distributions with means μ_F and μ_M, shared covariance Σ_W, between covariance matrix Σ_B and $w = \Sigma_W^{-1}(\mu_F - \mu_M)$ is the non-zero eigenvalue eigenvector of $\Sigma_W^{-1}\Sigma_B$. This whitening is like setting the discriminant component such that the estimated LLR is set to zero. The second baseline called adv-AE were proposed in [1]. It is based on adversarial disentanglement autoencoding. An encoder-decoder and a sex-classifier on the encoded representation are trained in an adversarial manner. The encoder tries to fool the classifier while the latter tries to predict the right sex. This would reduce the sex information in the encoded representation. The decoder takes an additional input: a score related to the sex information. During testing, this score is set to a constant 0.5 in order to reduce the sex information.

[2]Code and models are available: https://github.com/LIAvignon/bridge-features-evidence

4.2. Results

Before discussing the protection and the utility results, let's visualise in Figure 1 how the data looks in the latent space. One can see that the mapping makes the two class-conditional densities Gaussian like and only the first component is discriminant as expected.



(a) Principal components of x (b) Two first components of z

Fig. 1: V2T's x-vectors visualisation in the original and latent space. In the latent space (b), class-conditional densities look Gaussian and only the first component is discriminant while this is not the case in the original feature space (a).

Protection ability assessment:
A 2-layers perceptron sex classifier is trained on four versions of V2T-train: without protection, with the LDA and adv-AE protection and with the proposed method that we call NFzLLR. Each classifier is tested on the corresponding V2T-test set. Results are shown in Table 1. Even if the classifier learned to separate relatively well the sex on the training data due to overfitting, with the proposed method, it hardly generalises to the test set in comparison to the baselines. The best baseline is the adv-AE approach because its $C_{\text{llr}}^{\text{min}}$ on V2T-test is larger than for LDA and the expected amount of sex information disclosed by scores (D_{ECE}) is lower. However, the proposed approach outperforms both baselines in terms of $C_{\text{llr}}^{\text{min}}$ and D_{ECE}.

Table 1: Sex classification performance on non-protected and protected V2T. The first line is for non-protected data, the second and third lines are for the two baselines and the last line is for the proposed system. For good privacy, $C_{\text{llr}}^{\text{min}}$ has to be close to 1 and D_{ECE} should be as low as possible on V2T-test. Based on that, the proposed system provides better protection compared to the baselines.

	V2T-train		V2T-test	
	$C_{\text{llr}}^{\text{min}}{}_{10^{-2}}$	D_{ECE}	$C_{\text{llr}}^{\text{min}}{}_{10^{-2}}$	D_{ECE}
original data	8.39	0.658	2.12	0.703
LDA	12.20	0.628	57.75	0.295
adv-AE	30.43	0.493	74.21	0.179
NFzLLR	**48.45**	**0.362**	**95.75**	**0.029**

Table 2: Absolute number of samples per tag on V2T-test (relative values obscure extrema). Bold numbers indicate the worst-case tag.

	A	B	C	D	E	F
original data	15	194	**13735**	0	0	0
LDA	11592	844	**1508**	0	0	0
adv-AE	12454	684	**806**	0	0	0
NFzLLR	13821	**123**	0	0	0	0

In [5], ZEBRA's metrics release also the score with the highest strength-of-evidence corresponding to the worst-case scenario. Along with it, a categorical tag is given: the latter is a letter from A to F, from better to worst. Table 2 gives the number of samples per tag for V2T-test. One can see the worst-case tag for the proposed system is B with a low amount of scores compared to the other lines with a worst-case tag C. Mutual information measures are given in Table 3. It confirms the better protection ability of the proposed system. On V2T the drop of MI is about 96.47% while for the LDA and the adversarial approach, the decrease is about 88.00% and 90.00%.

Table 3: Mutual information measures between the x-vectors and the sex class variable y on V2D and V2T.

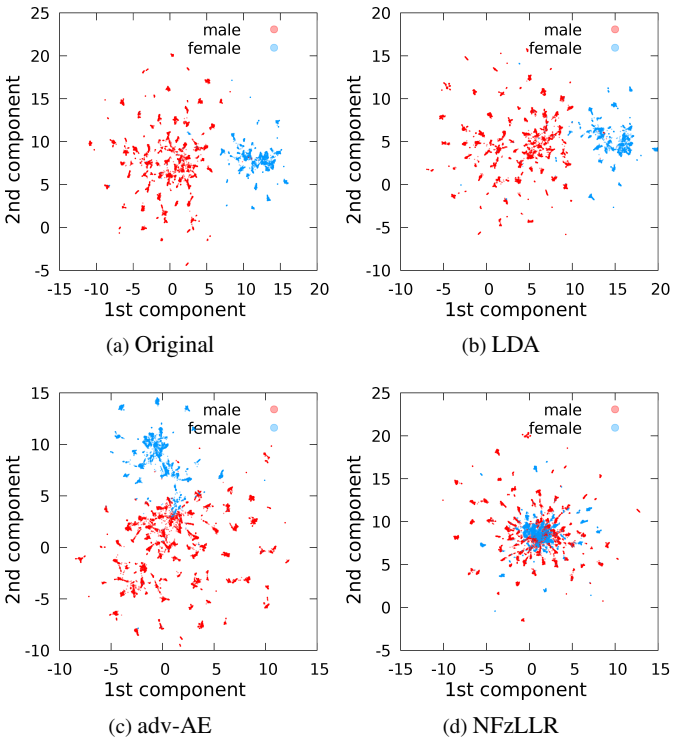
	MI $_{10^{-2}}$ [bit per dimension]	
	V2D	V2T
original data	18.7	19.0
LDA	1.43	2.28
adv-AE	1.0	1.90
NFzLLR	**0.14**	**0.67**

The better protection given by the proposed approach is also shown in Figure 2. It provides 2D visualisations of the embeddings using the UMAP dimensionality reduction and visualisation [24]. We can see that, even with the LDA and adv-AE methods, the unsupervised visualisation allows separation of the sex while with the new approach, male and female x-vectors are mixed.
In addition of making sure that the attribute has been hidden, we want to see if the utility is preserved i.e. if we can still do automatic speaker verification.

Table 4: Speaker verification performance. Even if the proposed system slightly alter the ASV performance, it does better than the adversarial-autoencoding system.

	EER [%]	$C_{\text{llr}}^{\text{min}}$
original data	1.72	0.067
adv-AE	2.36	0.097
NFzLLR	**2.11**	**0.086**

Automatic speaker verification results:
We use the same speaker verification protocol as in [1]. The common probabilistic linear discriminant analysis (PLDA) [25] is used to compute the scores from the comparisons between enrolment and probe segments. The PLDA is trained on non-protected data. Indeed, we suppose that the authentication side does not know that the user is protecting its data. Therefore, the authentication has a general ASV backend and the user applies protection on both reference and probe segments. The equal-error-rate (EER) and $C_{\text{llr}}^{\text{min}}$ for the original data, and protected with the adv-AE and the proposed approach are given in Table 4. Here, metrics are of course given for the classification between *target* and *impostor* trials while in Table 1, it was for the classification between *female* and *male* utterances. Because the protection ability of the LDA is not competitive, its results are not reported here. In addition of having a good protection ability, the method we proposed results in better speaker verification in comparison to the adversarial approach. This difference could be explained by the fact that the adversarial-autoencoder has some reconstruction error contrary to the normalising flow based approach. However, this adversarial-autoencoder's drawback can probably be reduced by investigating more flexible architecture and better tuning.



(a) Original (b) LDA

(c) adv-AE (d) NFzLLR

Fig. 2: UMAP visualisation (with euclidean metric and 30 neighbors) [24] of the original and protected V2T's x-vectors. Even with the LDA and adversarial based methods, unsupervised visualisation allows separation of the sex while with the proposed approach, male and female x-vectors are mixed.

5. CONCLUSION

This paper proposed a non-linear discriminant analysis that allows to map feature vector into a space where the LLR related to a binary attribute, and the residual i.e. everything that is independent of this attribute, are disentangled. Because the mapping is invertible, this approach can be used for attribute-driven privacy by manipulating the strength-of-evidence of the data. More precisely, binary information can be removed from some data by setting the estimated LLRs to zero. This would result in perfect privacy, also known as zero-evidence and perfect secrecy. Our experiments on the VoxCeleb2 speaker embeddings showed the ability of our approach to remove the sex information in the embeddings while preserving fairly well the speaker verification ability. The proposed approach is theoretically justifiable and outperforms, in terms of protection and utility, our previous work based on adversarial disentanglement.

The proposed approach is based on the LLR paradigm and zero-evidence which are well established for binary problems. However, most of the speaker attributes, like the nationality or the age, are not binary, even not discrete. Therefore, our future works will focus on the generalisation of our method and the zero-evidence concept to attributes that consist of more than two classes.

6. ACKNOWLEDGEMENTS

This work was supported by the VoicePersonae project ANR-18-JSTS-0001.

7. REFERENCES

[1] Paul-Gauthier Noé, Mohammad Mohammadamini, Driss Matrouf, Titouan Parcollet, Andreas Nautsch, and Jean-François Bonastre, "Adversarial Disentanglement of Speaker Representation for Attribute-Driven Privacy Preservation," in *Proc. Interspeech 2021*, 2021, pp. 1902–1906.

[2] Ranya Aloufi, Hamed Haddadi, and David Boyle, "Privacy-preserving voice analysis via disentangled representations," in *Proc. SIGSAC Conference on Cloud Computing Security Workshop*. ACM, 2020, pp. 1–14.

[3] Ronald Meester and Klaas Slooten, *Probability and Forensic Evidence: Theory, Philosophy, and Applications*, Cambridge University Press, 2021.

[4] Christophe Champod and Didier Meuwly, "The inference of identity in forensic speaker recognition," *Speech Communication*, vol. 31, no. 2, pp. 193–203, 2000.

[5] Andreas Nautsch, Jose Patino, N. Tomashenko, Junichi Yamagishi, Paul-Gauthier Noé, Jean-François Bonastre, Massimiliano Todisco, and Nicholas Evans, "The Privacy ZEBRA: Zero Evidence Biometric Recognition Assessment," in *Proc. Interspeech*. ISCA, 2020, pp. 1698–1702.

[6] Andreas Nautsch, Abelino Jiménez, Amos Treiber, Jascha Kolberg, Catherine Jasserand, Els Kindt, Héctor Delgado, Massimiliano Todisco, Mohamed Amine Hmani, Aymen Mtibaa, Mohammed Ahmed Abdelraheem, Alberto Abad, Francisco Teixeira, Driss Matrouf, Marta Gomez-Barrero, Dijana Petrovska-Delacrétaz, Gérard Chollet, Nicholas Evans, Thomas Schneider, Jean-François Bonastre, Bhiksha Raj, Isabel Trancoso, and Christoph Busch, "Preserving privacy in speaker and speech characterisation," *Computer Speech & Language*, vol. 58, pp. 441 – 480, 2019.

[7] N. Tomashenko, Brij Mohan Lal Srivastava, Xin Wang, Emmanuel Vincent, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Jose Patino, Jean-François Bonastre, Paul-Gauthier Noé, and Massimiliano Todisco, "Introducing the VoicePrivacy Initiative," in *Proc. Interspeech*. ISCA, 2020, pp. 1693–1697.

[8] Kevin P Murphy, *Machine learning: a probabilistic perspective*, 2012.

[9] Christopher M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag, Berlin, Heidelberg, 2006.

[10] Trevor Hastie and M Zhu, "Dimension reduction and visualization in discriminant analysis - discussion," *Australian & New Zealand Journal of Statistics*, vol. 43, pp. 179–185, 06 2001.

[11] I. Kobyzev, S. Prince, and M. Brubaker, "Normalizing flows: An introduction and review of current methods," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, , no. 01, pp. 1–1, may 5555.

[12] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio, "Density estimation using real NVP," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. 2017, OpenReview.net.

[13] Pavel Izmailov, Polina Kirichenko, Marc Finzi, and Andrew Gordon Wilson, "Semi-supervised learning with normalizing flows," in *International Conference on Machine Learning*. PMLR, 2020, pp. 4615–4630.

[14] Andrei Atanov, Alexandra Volokhova, Arsenii Ashukha, Ivan Sosnovik, and Dmitry Vetrov, "Semi-conditional normalizing flows for semi-supervised learning," *arXiv preprint arXiv:1905.00505*, 2019.

[15] Niko Brümmer and Johan 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.

[16] David A. van Leeuwen and Niko Brümmer, "The distribution of calibrated likelihood-ratios in speaker recognition," in *Proc. Interspeech 2013*, 2013, pp. 1619–1623.

[17] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech*. ISCA, 2018, pp. 1086–1090.

[18] C. E. Shannon, "Communication theory of secrecy systems," *The Bell System Technical Journal*, vol. 28, no. 4, pp. 656–715, 1949.

[19] I. J. Good, "Studies in the history of probability and statistics. xxxvii a. m. turing's statistical work in world war ii," *Biometrika*, vol. 66, no. 2, pp. 393–396, 1979.

[20] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[21] Paul-Gauthier Noé, Andreas Nautsch, Nicholas Evans, Jose Patino, Jean-François Bonastre, Natalia Tomashenko, and Driss Matrouf, "Towards a unified assessment framework of speech pseudonymisation," *Computer Speech & Language*, vol. 72, pp. 101299, 2022.

[22] Brian C Ross, "Mutual information between discrete and continuous data sets," *PloS one*, vol. 9, no. 2, pp. 1–5, 2014.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[24] Leland McInnes, John Healy, and James Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[25] Sergey Ioffe, "Probabilistic linear discriminant analysis," in *Computer Vision – ECCV 2006*. 2006, pp. 531–542, Springer Berlin Heidelberg.