Faire le pont entre l'observation et la preuve : Application au respect de la vie privée

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

Faire le pont entre l'observation et la preuve : Application au respect de la vie privée

Paul-Gauthier Noé[1] Andreas Nautsch[2] Driss Matrouf[1]
Pierre-Michel Bousquet[1] Jean-François Bonastre[1]
(1) Laboratoire Informatique d'Avignon, France
(2) Individu, Allemagne
paul-gauthier.noe@univ-avignon.fr

RÉSUMÉ

Le respect de la vie privée dans les technologies de la parole consiste généralement à dissimuler l'identité du·de la locuteur·rice. Dans cet article, nous traitons du respect de la vie privée dit *contrôlé* où un·e utilisateur·rice aurait la possibilité de dissimuler un ou plusieurs de ses attributs comme son sexe, son âge ou ses origines tout en préservant le reste de sa personnalité vocale. Lorsque l'attribut est binaire, le *secret parfait*, défini par Claude Shannon, nécessite d'avoir le log-ratio de vraisemblances (LRV) à zero. Cet article propose donc une approche qui permet, dans une représentation du·de la locuteur·rice, de manipuler le LRV relatif à l'attribut afin de l'annihiler. Nos expériences sur des x-vecteurs issus des bases de données VoxCeleb2 montrent que notre méthode est capable de dissimuler le sexe du·de la locuteur·rice tout en maintenant la possibilité de faire de la vérification automatique du·de la locuteur·rice.

ABSTRACT

A bridge between observation and evidence : An application to attribute-driven privacy.

Most of the time, privacy in speech technology aims in hiding the full speaker identity. In this paper, we deal with attribute-driven privacy which consists in hiding only a few speaker's personal attributes like its sex, age or ethnicity, while preserving the remaining voice richness. Defined by Claude Shannon, *perfect secrecy* requires that the posterior probabilities are the same as the prior ones resulting in no gain of information. When the attribute is binary, it means that the corresponding log-likelihood-ratio (LLR) should be zero. This paper presents an approach that allows to manipulate the LLR corresponding to the attribute in a speaker representation. For privacy, the LLR can be set to zero. Our experiments on VoxCeleb2 datasets' x-vector show that our method is able to conceal the speaker's sex while still being able to do automatic speaker verification.

MOTS-CLÉS : Secret parfait, vie privée, vérification du·de la locuteur·rice, *normalizing flow*.

KEYWORDS: Perfect secrecy, privacy, speaker verification, normalizing flow.

1 Introduction

Le respect de la vie privée dans les technologies vocales consiste généralement à dissimuler l'identité des utilisateur·rice·s. Ce concept est connu sous le nom d'anonymisation ou de pseudonymisation (Tomashenko *et al.*, 2020; Noé *et al.*, 2022a). Ici nous abordons un autre type de protection. Nous nous intéressons au respect de la vie privée que nous qualifions de *contrôlé* et qui consiste à dissimuler dans les données un ou plusieurs attributs que l'utilisateur·rice juge sensibles et désire cacher. On peut

penser entre autres au sexe, à l'âge ou à l'origine ethnique. Dans (Noé *et al.*, 2021) nous proposions une approche antagoniste de démêlage de l'attribut d'intérêt. Une fois la variable attribut isolée, il est possible de la fixer à une constante pour réduire la capacité de distinguer les modalités de l'attribut. Cependant, le choix de cette constante est arbitraire, cette approche n'est pas probabiliste et reste difficilement explicable. Ici nous proposons une nouvelle approche basée sur l'inférence Bayesienne et ayant pour but de d'approcher au mieux le *secret parfait*. Le secret parfait est atteint lorsque les cryptogrammes sont indépendants des messages (Shannon, 1949). Dans notre cas, ce sont les données qui doivent être indépendantes de l'attribut. Nous proposons donc de démêler, dans un vecteur d'observation, la *preuve* et le *résiduel*. Nous appelons ici preuve tout ce qui est utile à la tache de décision (Meester & Slooten, 2021) [1] et nous appelons résiduel tout ce qui ne l'est pas, autrement dit, tout ce qui, dans l'observation, est indépendant de l'attribut. Nous nous limitons ici aux attributs binaires. Dans ce cas, la preuve est représentée par le ratio de vraisemblances (RV) ou log-RV (LRV) entre deux propositions mutuellement exclusives (chacune correspondant à une modalité de l'attribut). Nous verrons que, pour atteindre le secret parfait, le LRV doit être mis à zéro pour toutes les observations. L'analyse discriminante linéaire (ADL) projette les données dans un espace où la dimension discriminante est le logit [2] de la probabilité a posteriori (Murphy, 2012). Cette dimension peut être ajustée de manière à ce que le LRV estimé soit mis à zéro et les données peuvent être reconstruites dans l'espace des caractéristiques observées à l'aide de la transformation inverse. Cependant, cette approche est linéaire et suppose que les variables observées, conditionnées sur les classes, suivent une loi normale multidimensionnelle de même matrice de covariance. Cette approche rencontre donc rapidement des limites lorsqu'il s'agit de l'appliquer à des données réelles. Cet article propose donc une analyse discriminante non linéaire [3] permettant de plonger les données dans un espace où la composante discriminante est le LRV et où les autres dimensions constituent le résiduel. Dans cette espace, le LRV peut être mis à zéro afin de rendre les observations indépendantes de l'attribut et ainsi s'assurer du secret parfait. Cette approche est basée sur le *normalizing flow* (Kobyzev *et al.*, 2021) qui permet d'apprendre une bijection entre l'espace des observations et un espace base. Si cette approche est généralement utilisée dans un cadre non supervisé, certains travaux proposent de l'utiliser de manière supervisée (Atanov *et al.*, 2019; Izmailov *et al.*, 2020). Dans notre cas, chaque variable conditionnée sur une modalité de l'attribut suit une loi normale multidimensionnelle dans l'espace base (et non dans l'espace des observations). Les paramètres de ces lois sont contraints de sorte que seule la première dimension est discriminante et donne une estimation du LRV calibré (Brümmer & du Preez, 2006; van Leeuwen & Brümmer, 2013). L'approche *normalizing flow* donnant une transformation inversible, il est donc possible de manipuler le LRV et ensuite d'appliquer la transformation inverse pour obtenir de manière non linéaire l'observation protégée.

La partie suivante rappelle la formule de Bayes, la notion de preuve et leur lien avec le secret parfait. La troisième section présente la solution que nous proposons pour annihiler le poids de la preuve dans les données, dans un souci de respect de la vie privée. La quatrième partie présente une application de notre approche où l'information relative au sexe du·de la locuteur·rice est dissimulée dans des représentations de type x-vecteur (Snyder *et al.*, 2018). Des résultats sur la base de données VoxCeleb2 (Chung *et al.*, 2018) y sont présentés afin d'évaluer la capacité de notre méthode à protéger les données tout en préservant l'efficacité des systèmes de vérification du·de la locuteur·rice. Cet article est une réécriture et mise à jour des travaux présentés ici (Noé *et al.*, 2022b).

1. Attention, en anglais, le terme *evidence* fait référence à ce que nous appelons ici *preuve* alors qu'en français, l'*évidence* fait référence à une manière d'exprimer les probabilités basée sur la fonction logit (voir Section 2.1).

2. La fonction logit, réciproque de la fonction sigmoid est définie comme $\mathrm{logit}(p) = \frac{p}{1-p}$ où $p \in]0;1[$.

3. L'analyse discriminante quadratique permet d'estimer de manière non linéaire le LRV. Cependant, elle se base aussi sur des hypothèses de Gaussianité des données et ne permet pas de transformation inverse (Hastie & Zhu, 2001).

2 Formule de Bayes et secret parfait

Dans cette section nous rappelons la formule de Bayes dans un cas binaire et comment cette formule peut être décomposée en un terme des probabilités a priori et un terme correspondant à la *preuve*. Nous montrons ainsi que pour atteindre le secret parfait, ce dernier terme doit être mis à zéro. Enfin, nous rappelons une propriété sur la distribution des LRV nécessaire au développement de notre approche.

2.1 Evidence, preuve et secret parfait

Soit un attribut binaire avec un ensemble de modalités ou classes $\mathcal{C} = \{c_0, c_1\}$. Considérons un·e attaqueur·euse qui cherche, à partir d'une observation x, à inférer la classe de l'attribut. La mise à jour de sa connaissance peut s'écrire avec la formule de Bayes :

$$\text{logit}\, P(c_i|x) = \log \frac{P(x|c_i)}{P(x|c_{\neg i})} + \text{logit}\, P(c_i), \text{ où } i \in \{0, 1\}, \tag{1}$$

où $P(c_i)$ et $P(c_i|x)$ sont respectivement les probabilités a priori et a posteriori. Cette manière d'écrire les probabilités a posteriori permet de les exprimer sous forme d'évidence. La formule de Bayes devient une somme entre un terme « subjectif » dépendant de la connaissance a priori de l'attaqueur·euse et un terme « objectif », que nous appellerons preuve, composé des vraisemblances : le log-ratio de vraisemblances [4]. Le *secret parfait* présenté par Claude Shannon (Shannon, 1949) est atteint lorsque les probabilités a posteriori restent égales aux probabilités a priori. Dans ce cas, l'avis de l'attaqueur·euse ne change pas après avoir observé les données, autrement dit, il n'a reçu aucune information utile et les observations sont indépendantes de l'attribut. Dans notre cas, le secret parfait est atteint lorsque le LRV est zéro. Ainsi, pour ôter d'une base de données l'information relative à un attribut binaire, l'idée est de mettre le LRV à zéro pour chaque échantillon. Nous proposons donc une méthode permettant de plonger un vecteur de caractéristiques dans un espace où la première dimension est le LRV et où les autres dimensions contiennent la variabilité résiduelle. Mais pour ce faire, nous devons d'abord comprendre quelle doit être la distribution des LRV dans un tel espace.

2.2 La distribution des log-ratio de vraisemblance

Le ratio de vraisemblance (RV) du RV est le RV (van Leeuwen & Brümmer, 2013; Meester & Slooten, 2021). Cette propriété connue sous le nom d'*idempotence* ajoute une contrainte sur les distributions des LRV en liant les paramètres de la distribution des LRV pour la classe c_0 à ceux de la distribution pour la classe c_1. En effet, si l'une des distributions conditionnelles est Gaussienne avec une moyenne μ alors l'autre est nécessairement Gaussienne avec une moyenne $-\mu$ et a la même variance $\sigma^2 = 2\mu$ (van Leeuwen & Brümmer, 2013; Good, 1979).

4. En pratique, ce terme a une part de subjectivité car il dépend des données et du modèle sous-jacent nécessaire à son calcul. Il est objectif dans le sens où il dépend uniquement des vraisemblances et ne dépend pas de la croyance a priori de l'attaqueur·euse.

3 De l'espace observable à la preuve et inversement

Soit $\mathcal{X}$ l'espace des vecteurs de caractéristiques de dimension n où chaque vecteur correspond à une observation. Supposons que chaque vecteur a une preuve représentée par un LRV $l \in \mathcal{L} \subset \mathbb{R}$ et un résiduel $r = (r_1, \ldots, r_{n-1})^T$ dans $\mathcal{R} \in \mathbb{R}^{n-1}$. On définit alors un espace base $\mathcal{Z} = \mathcal{L} \oplus \mathcal{R}$. Supposons qu'il existe une bijection f entre $\mathcal{X}$ et $\mathcal{Z}$. Une telle relation permettrait donc de passer de l'espace de l'observation à un espace où la preuve et le résiduel sont démêlés. Notre but est donc de s'approcher d'une telle relation.

3.1 Les densités conditionnelles dans l'espace base

Nous définissons un modèle génératif où les distributions conditionnelles dans l'espace base sont $z|c_0 \sim \mathcal{N}(\mu e_1, \Sigma)$ et $z|c_1 \sim \mathcal{N}(-\mu e_1, \Sigma)$, avec $\mu \in \mathbb{R}^+$, $e_1 = (1, 0, \ldots 0)^T$ et $\Sigma = \text{diag}(2\mu, 1, \ldots 1)$. Ainsi, la première dimension $z_0 = e_1^T z$ est le LRV et vérifie la propriété de l'idempotence présentée dans la Section 2.2. Les autres dimensions constituent le résiduel modélisé par une loi normale multidimensionnelle dont la moyenne est le vecteur nul et dont la matrice de covariance est l'identité.

3.2 Transformation inversible entre l'espace base et celui des observations

La méthode de *normalizing flow* permet, en maximisant la vraisemblance des données, d'apprendre une bijection g entre l'espace observé et l'espace base. Une telle transformation inversible permet donc de faire aussi bien de l'inférence que de la génération. Dans notre cas, chaque échantillon $x \in \mathcal{X}$ est associé à une classe $c \in \mathcal{C}$ et on note $\mathcal{D} = \{(x^{(0)}, c^{(0)}), \ldots (x^{(N-1)}, c^{(N-1)})\}$ l'ensemble de nos observations. La vraisemblance des données est :

$$\log p_{X|\theta_g,\mu}(\mathcal{D}) = \sum_{i=0}^{1} \left(\sum_{(x,c)\in\mathcal{D}|c=c_i} \log p_{X|c_i,\theta_g,\mu}(x) \right), \text{ où } \theta_g \text{ sont les paramètres de } g. \tag{2}$$

Parce que g est une bijection et qu'elle est appliquée sans distinction sur les échantillons des classes c_0 et c_1, la formule de changement de variable peut être utilisée :

$$\forall i \in \{0, 1\},\ p_{X|c_i,\theta_g,\mu}(x) = p_{Z|c_i,\mu}(z) \left| \det \left(\frac{\partial g(z)}{\partial z} \right) \right|^{-1}, \text{ où } x = g(z). \tag{3}$$

Dans notre cas, g est un réseau de neurones dont l'inversion est simple et dont le déterminant de la matrice Jacobienne (dans l'Equation 3) peut être calculé rapidement lors de l'optimisation du réseau (Dinh *et al.*, 2017).

3.3 Le paramètre des distributions dans l'espace base et son optimisation

Comme nous l'avons vu dans la Section 3.1, l'unique paramètre des distributions dans l'espace base est μ et se manifeste uniquement le long de la première dimension. Ainsi, on peut facilement obtenir un estimateur du maximum de vraisemblance pour ce paramètre :

$$\hat{\mu}_{\text{MLE}}(\mathcal{B}_Z) = -1 + \sqrt{1 + \frac{1}{|\mathcal{B}_Z|} \sum_{z\in\mathcal{B}_Z} (e_1^T z)^2}, \text{ où } \mathcal{B}_Z \text{ est un lot d'échantillons dans l'espace base.} \tag{4}$$

Nous proposons donc une optimisation itérative en deux étapes à l'image de l'algorithme espérance-maximisation. Dans notre cas, les paramètres θ_g de g sont optimisés avec μ fixé puis μ est optimisé avec θ_g fixé :

Choisir un paramètre d'adaptation $\alpha \in [0, 1]$,
Initialiser θ_g et μ,
pour *tout lot d'observation* $\mathcal{B}_X$ **faire**
  $\mathcal{B}_Z \leftarrow g^{-1}(\mathcal{B}_X)$
  $\theta_g \leftarrow \underset{\theta_g}{\operatorname{argmax}} \log p_{X|\theta_g,\mu}(\mathcal{B}_X)$
  $\mu \leftarrow \alpha\mu + (1-\alpha)\hat{\mu}_{\text{MLE}}(\mathcal{B}_Z)$.
fin

3.4 Manipuler la preuve dans l'espace base pour le respect de la vie privée

La relation entre l'espace observable et l'espace base, dont la première dimension est le LRV, étant inversible, il est possible de manipuler le poids de la preuve dans l'espace base et de reconstruire l'observation dont l'information relative à l'attribut aura été altérée. Plus précisément, étant donné que la transformation est inversible, le RV dans l'espace base est identique au RV dans l'espace des observations :

$$\frac{p_{Z|c_0}(z)}{p_{Z|c_1}(z)} = \frac{p_{X|c_0}(x)\left|\det\left(\frac{\partial g(z)}{\partial z}\right)\right|}{p_{X|c_1}(x)\left|\det\left(\frac{\partial g(z)}{\partial z}\right)\right|} = \frac{p_{X|c_0}(x)}{p_{X|c_1}(x)}. \tag{5}$$

Ainsi, plonger un vecteur d'observation dans l'espace base, mettre la première dimension à zéro et plonger en retour le vecteur transformé dans l'espace des observations permet d'enlever l'information relative à l'attribut. La prochaine section montre comment cette méthode peut être utilisée afin d'enlever dans une représentation locuteur·rice l'information relative au sexe du·de la locuteur·rice.

4 Dissimuler le sexe dans une représentation locuteur·rice

Cette section présente un exemple d'application de notre approche pour le respect de la vie privée contrôlé. Le but ici est de permettre à l'utilisateur·rice de profiter d'un système de vérification du·de la locuteur·rice tout en cachant son sexe. Pour cela nous proposons donc d'appliquer notre méthode sur des représentations de type x-vecteur (Snyder *et al.*, 2018) qui, une fois protégées, seront transmises au service d'authentification [5].

Dans notre expérience, les x-vecteurs utilisés sont de type TDNN extraits avec Kaldi [6]. Notre système de protection est respectivement entraîné et testé sur V2D (397032 utterances par classe) et V2T (9120 utterances pour la classe femme et 22559 pour la classe homme), respectivement un sous-ensemble de la partie de développement et de test de VoxCeleb2 (Chung *et al.*, 2018). Pour vérifier la capacité de notre système à cacher l'information relative au sexe, un classificateur est entraîné sur les données

5. Cela nécessite donc que l'extraction de la représentation et sa protection se fassent de manière locale. Cette configuration n'est pas forcément réalisable ou souhaitable pour certaines applications. Cependant, cette étude ayant pour but de proposer une modeste preuve de concept et n'ayant pas la prétention d'apporter une solution directement applicable, nous ne rentrerons pas dans ces considérations ici.

6. `https://kaldi-asr.org/models/m7`

transformées. Pour ce faire, V2T est donc divisée en deux parties : V2T-train (46 hommes et 25 femmes) et V2T-test (35 hommes et 14 femmes différent·e·s que pour V2T-train) respectivement pour l'entraînement et pour le test du classificateur. A partir des scores obtenus par le classificateur, deux métriques sont calculées : le $C_{\text{llr}}^{\text{min}}$ qui mesure le pouvoir discriminant des scores calibrés (proche de 0 : fort pouvoir discriminant, faible lorsque proche de 1) (Brümmer & du Preez, 2006) et le D_{ECE} qui mesure la quantité d'information, fournie par les scores à l'attaqueur·euse, moyennée sur l'ensemble de ses a priori possibles (Nautsch *et al.*, 2020). L'information mutuelle est aussi calculée entre chaque dimension du x-vecteur et la variable de classe (Ross, 2014; Pedregosa *et al.*, 2011).
Pour enlever l'information relative au sexe dans les x-vecteurs, nous proposons d'utiliser la méthode, que nous nommerons NFzLRV, présentée dans la Section 3. L'architecture de *normalizing flow* que nous utilisons est de type *Real NVP* (Dinh *et al.*, 2017) avec 6 couches de type *coupling* où les fonctions de *scaling* et de *translation* sont des perceptrons multicouche. Les deux sont composées de 3 couches linéaires avec deux activations LeakyReLU et une de sortie de type tangente hyperbolique pour la fonction *scaling* alors que la fonction de translation n'a pas d'activation de sortie. Pour l'apprentissage, l'algorithme Adam est utilisé avec un taux d'apprentissage de 10^{-4}, μ est initialisé à 10 et $\alpha = 0,99$. Nous comparons notre méthode à deux systèmes de référence. Le premier est basé sur l'ADL et correspond au blanchiment :

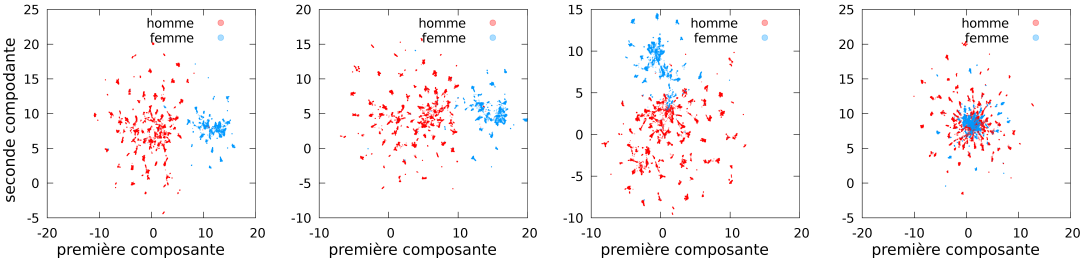
$$x \leftarrow x - \frac{ww^T}{||w||^2}x + \frac{1}{2||w||^2}(\mu_F^T\Sigma_W^{-1}\mu_F - \mu_M^T\Sigma_W^{-1}\mu_M)w, \tag{6}$$

où les distributions conditionnelles des données observées sont des lois normales de moyennes μ_F et μ_M, de matrice de covariance Σ_W, où la matrice de covariance interclasse est Σ_B et où $w = \Sigma_W^{-1}(\mu_F - \mu_M)$ est le vecteur propre associé à la valeur propre non nulle de $\Sigma_W^{-1}\Sigma_B$. Le second système de référence est l'approche antagoniste (adv-AE). Plus précisément, cette approche est basée sur une architecture de type auto-encodeur où l'encodeur essaie de tromper un classificateur qui cherche à prédire correctement le sexe du·de la locuteur·rice à partir de la représentation en sortie de l'encodeur. L'apprentissage concurrentiel entre l'encodeur et le classificateur tend à rendre la représentation intermédiaire indépendante du sexe. Pour permettre la reconstruction durant l'apprentissage du système, l'information relative au sexe est fournie sous forme d'un score au décodeur. Ici, pour diminuer l'information relative à l'attribut en sortie de l'auto-encoder, ce score est constant égal à 0,5 durant la phase de test. Pour plus de précisions, voir (Noé *et al.*, 2021).

Evaluation de la protection. Un perceptron à deux couches est entraîné, pour une tâche de classification du sexe du·de la locuteur·rice, sur V2T-train (et testé sur V2T-test) non protégé et protégé avec les trois différents systèmes. Les scores obtenus sont utilisés pour calculer le $C_{\text{llr}}^{\text{min}}$ et le D_{ECE}. Les résultats sont fournis dans le Tableau 1. Sur V2T-test, les valeurs proches respectivement de 1 et de 0 pour le $C_{\text{llr}}^{\text{min}}$ et le D_{ECE} pour le système NFzLRV montrent la difficulté du classificateur de généraliser. Cet écart de valeurs entre l'ensemble de train et de test suggère que le classificateur a sur-appris sur les données d'entraînement. De plus, la bonne protection assuré par NFzLRV est confirmée par des valeurs d'information mutuelle plus faible par rapport aux autres systèmes. Sur V2T, l'information mutuelle chute de 96,5% avec le système proposé alors qu'avec les méthodes ADL et adv-AE elle chute respectivement de 88,0% et 90,0%.
La Figure 1 montre une visualisation UMAP (McInnes *et al.*, 2018) des x-vecteurs de l'ensemble V2T avec et sans protection. On peut voir que cette méthode de visualisation non supervisée permet de distinguer un groupe homme et un groupe femme même avec les protections ADL et adv-AE alors qu'avec la méthode que nous proposons, les deux groupes se confondent.

TABLE 1 – Partie gauche : Performance de classification du sexe du·de la locuteur·rice sur des données non protégées et protégées. Pour une bonne protection, les performances de classification doivent être pauvres. Le $C_{\text{llr}}^{\text{min}}$ doit être proche de 1, et le D_{ECE} doit être le plus petit possible. Partie droite : Mesure d'information mutuelle entre les dimensions du x-vecteur et la variable binaire sexe y.

	$C_{\text{llr}}^{\text{min}} 10^{-2}$		D_{ECE}		IM [bit par dimension]	
	V2T-train	V2T-test	V2T-train	V2T-test	V2D	V2T
original data	8,39	2,12	0,658	0,703	18,7	19,0
ADL	12,20	57,75	0,628	0,295	1,43	2,28
adv-AE	30,43	74,21	0,493	0,179	1,0	1,90
NFzLRV	**48,45**	**95,75**	**0,362**	**0,029**	**0,14**	**0,67**



(a) Pas de protection, (b) protection avec ADL, (c) avec adv-AE, (d) avec NFzLRV.

FIGURE 1 – Visualisation UMAP des x-vecteurs de V2T avec et sans protection. Cette visualisation non supervisée permet de distinguer des groupes homme et femme lorsque la protection est effectuée avec les méthodes ADL et adv-AE contrairement à la protection avec la méthode NFzLRV.

TABLE 2 – Performance du système de vérification avec des représentations locuteur·rice protégées ou non. 2a donne les résultats lorsque les tests sont entre locuteur·rice·s de même sexe et de sexe différents alors que 2b et 2c les donnent pour des tests respectivement entre hommes et entre femmes.

(a) Comparaisons inter et intra hommes et femmes

	EER [%]	$C_{\text{llr}}^{\text{min}}$
original data	1,72	0,067
adv-AE	2,36	0,097
NFzLRV	**2,11**	**0,086**

(b) Intra hommes

EER [%]	$C_{\text{llr}}^{\text{min}}$
1,82	0,072
2,53	0,103
1,92	**0,079**

(c) Intra femmes

EER [%]	$C_{\text{llr}}^{\text{min}}$
2,76	0,109
4,14	0,151
3,03	**0,120**

Vérification automatique du·de la locuteur·rice. Bien que le système que nous proposons ait de meilleurs performances en terme de protection, on peut se demander s'il est toujours possible de faire de la vérification du·de la locuteur·rice avec des x-vecteurs ainsi protégés. Nous comparons dans cette section les performances en terme de vérification automatique du·de la locuteur·rice lorsque des x-vecteurs non protégés et protégés sont utilisés. Nous utilisons le même protocole que dans (Noé *et al.*, 2021). Une Analyse Discriminante Lineaire Probabiliste (PLDA) (Ioffe, 2006) permet de comparer deux énoncés. En effet, le service d'authentification compare un énoncé de référence avec

un énoncé de test envoyée par l'utilisateur·rice au moment de la tentative d'authentification. Dans un contexte de respect de la vie privée, les deux énoncés sont bien évidemment protégées pour éviter de diffuser l'information sensible au service d'authentification. La PLDA permet d'obtenir un score censé être le rapport de vraisemblance entre l'hypothèse *cible* (les deux énoncés proviennent d'un·e même locuteur·rice) et l'hypothèse *imposteur* (les deux énoncés proviennent de deux locuteur·rice·s différent·e·s). Le Tableau 2 reporte les mesures de Taux d'égale erreur (equal-error-rate EER) et de $C_{\text{llr}}^{\text{min}}$ calculés à partir de ces scores. Nous pouvons voir que protéger les représentations locuteur·rice réduit les performances du système de vérification. Cependant, le système que nous proposons ici semble plus précis que l'approche antagoniste. Ceci peut être du au fait que l'approche basée sur l'autoencodeur possède une erreur de reconstruction, contrairement à l'approche *normalising flow* dont le plongement est inversible. Les Tableaux 2b et 2c montrent un fossé entre les performances par sexe, qui peut s'expliquer par une sous-représentation des femmes dans les bases de données.

5 Conclusion

Cet article présente une analyse discriminante non linéaire permettant de plonger les données dans un espace où le log-ratio de vraisemblances (LRV), correspondant à un attribut binaire, et le résiduel, c'est à dire tout ce qui, dans l'observation, est indépendant de l'attribut, sont séparés. Parce que le plongement est basé sur une approche de type *normalizing flow*, il est inversible. Le LRV peut donc être manipulé et être mis à zéro réduisant ainsi le poids de la preuve dans un contexte de respect de la vie privée. Nos expériences sur les bases de données VoxCeleb2 montrent que notre approche permet de cacher le sexe du·de la locuteur·rice dans ses représentations utilisées par une technologie d'authentification vocale, tout en limitant la dégradation des performances de cette dernière.
Cette approche est basée sur le cadre théorique d'inférence Bayésienne d'une variable à deux classes où la preuve peut être représenté par le LRV. En réalité, beaucoup d'attribut personnels comme l'origine, l'âge,... ne sont pas binaires, voire ne sont pas discrets. Ainsi, nous aimerions, dans de futurs travaux, étendre les idées présentées ici à ces cas plus généraux.

Remerciements : Ces travaux ont été financés par le projet VoicePersonae ANR-18-JSTS-0001.

Références

ATANOV A., VOLOKHOVA A., ASHUKHA A., SOSNOVIK I. & VETROV D. (2019). Semi-conditional normalizing flows for semi-supervised learning. *arXiv preprint arXiv :1905.00505*.

BRÜMMER N. & DU PREEZ J. (2006). Application-independent evaluation of speaker detection. *Computer Speech & Language*, **20**(2), 230–275. Odyssey 2004 : The speaker and Language Recognition Workshop.

CHUNG J. S., NAGRANI A. & ZISSERMAN A. (2018). Voxceleb2 : Deep speaker recognition. In *Proc. Interspeech*, p. 1086–1090 : ISCA.

DINH L., SOHL-DICKSTEIN J. & BENGIO S. (2017). Density estimation using real NVP. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings* : OpenReview.net.

GOOD I. J. (1979). Studies in the history of probability and statistics. xxxvii a. m. turing's statistical work in world war ii. *Biometrika*, **66**(2), 393–396.

HASTIE T. & ZHU M. (2001). Dimension reduction and visualization in discriminant analysis - discussion. *Australian & New Zealand Journal of Statistics*, **43**, 179–185.

IOFFE S. (2006). Probabilistic linear discriminant analysis. In *Computer Vision – ECCV 2006*, p. 531–542 : Springer Berlin Heidelberg.

IZMAILOV P., KIRICHENKO P., FINZI M. & WILSON A. G. (2020). Semi-supervised learning with normalizing flows. In *International Conference on Machine Learning*, p. 4615–4630 : PMLR.

KOBYZEV I., PRINCE S. J. & BRUBAKER M. A. (2021). Normalizing flows : An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43**(11), 3964–3979.

MCINNES L., HEALY J. & MELVILLE J. (2018). Umap : Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv :1802.03426*.

MEESTER R. & SLOOTEN K. (2021). *Probability and Forensic Evidence : Theory, Philosophy, and Applications*. Cambridge University Press.

MURPHY K. P. (2012). *Machine Learning : A Probabilistic Perspective*.

NAUTSCH A., PATINO J., TOMASHENKO N., YAMAGISHI J., NOÉ P.-G., BONASTRE J.-F., TODISCO M. & EVANS N. (2020). The Privacy ZEBRA : Zero Evidence Biometric Recognition Assessment. In *Proc. Interspeech*, p. 1698–1702 : ISCA.

NOÉ P.-G., MOHAMMADAMINI M., MATROUF D., PARCOLLET T., NAUTSCH A. & BONASTRE J.-F. (2021). Adversarial Disentanglement of Speaker Representation for Attribute-Driven Privacy Preservation. In *Proc. Interspeech 2021*, p. 1902–1906.

NOÉ P.-G., NAUTSCH A., EVANS N., PATINO J., BONASTRE J.-F., TOMASHENKO N. & MATROUF D. (2022a). Towards a unified assessment framework of speech pseudonymisation. *Computer Speech & Language*, **72**, 101299.

NOÉ P.-G., NAUTSCH A., MATROUF D., BOUSQUET P.-M. & BONASTRE J.-F. (2022b). A bridge between features and evidence for binary attribute-driven perfect privacy. à paraître dans proc. ICASSP 2022.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

ROSS B. C. (2014). Mutual information between discrete and continuous data sets. *PloS one*, **9**(2), 1–5.

SHANNON C. E. (1949). Communication theory of secrecy systems. *The Bell System Technical Journal*, **28**(4), 656–715.

SNYDER D., GARCIA-ROMERO D., SELL G., POVEY D. & KHUDANPUR S. (2018). X-vectors : Robust DNN embeddings for speaker recognition. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5329–5333 : IEEE.

TOMASHENKO N., SRIVASTAVA B. M. L., WANG X., VINCENT E., NAUTSCH A., YAMAGISHI J., EVANS N., PATINO J., BONASTRE J.-F., NOÉ P.-G. & TODISCO M. (2020). Introducing the VoicePrivacy Initiative. In *Proc. Interspeech*, p. 1693–1697 : ISCA.

VAN LEEUWEN D. A. & BRÜMMER N. (2013). The distribution of calibrated likelihood-ratios in speaker recognition. In *Proc. Interspeech 2013*, p. 1619–1623.